HYBRID MACHINE LEARNING MODEL FOR COMPARATIVE OPINION MINING IN BRAND REPUTATION MONITORING

ONDARA, BERNARD OMOI (M.SC. COMPUTER SCIENCE)

J57/25718/2018

A THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF DOCTOR OF PHILOSOPHY (COMPUTER SCIENCE) IN THE SCHOOL OF PURE AND APPLIED SCIENCES OF KENYATTA UNIVERSITY.

NOVEMBER 2024

DECLARATION

I, Ondara Bernard Omoi, declare that this thesis is my original research work and has not been presented at any other university or institution for consideration for any certification or degree award. Where text, graphics, pictures, figures, or tables have been borrowed from other sources, these are accredited and references cited using the current APA system and in accordance with anti-plagiarism regulations.

Signature: …… Date: …….

Ondara, Bernard Omoi

J57/25718/2018

Department of Computing and Information Science

Supervisors

This thesis has been submitted for appraisal with our/my approval as University Supervisor(s)

Signature: …… Date: ……..……

Dr. Stephen T. Waithaka

Department of Computing & Information Science

Kenyatta University

Signature: …… Date: ……..……

Dr. John M. Kandiri

Department of Computing & Information Science

Kenyatta University

Signature: …… Date: ….……

Dr. Lawrence Muchemi

Department of Computing and Informatics

University of Nairobi

DEDICATION

This research is dedicated to God Almighty for His sustaining grace upon me, to my dear wife (Brigid), my dear children (Brighton, Brandon, Billy, and Benjamin), my dear brother (Julius), and my associates from whom I received awesome support.

ACKNOWLEDGEMENT

First, my greatest gratitude is to the Living God, for his grace and providence throughout the research period.

Second, my exceptional appreciations are to my three supervisors, Dr. Stephen Waithaka, Dr. John Kandiri, and Dr. Lawrence Muchemi, for their unwavering support and key inputs in the forms of guidance, motivation, insights, and other forms of support. Based on their vast amounts of knowledge and experience in both research and in the field of computing, it was possible to make significant progress, leading to the production of this document and the intended developments afterwards.

Third, I thank all my colleagues and friends for the moral support they offered me while I was working on this thesis. Special thanks go to Prof. Elizaphan Maina for his objective guidance on postgraduate studies, which directly influenced my studies and this research work.

I really want to thank Dr. Eric Araka, Dr. Lucy Gitau, and Dr. Peter Wamae for constantly encouraging me on my journey to completing this thesis.

Finally, I honestly express gratitude to my beloved wife (Brigid) as well as my dear children (Brighton, Brandon, Billy, and Benjamin) for their support to me during my studies. I also thank my dear brother (Julius) for his words of reinforcement that kept me going.

ABSTRACT

Social media platforms like X platform (formerly Twitter) and online review websites like Amazon Reviews allow people to express their opinions about a brand's products or services. To obtain competitive intelligence, brands can leverage this online user-generated content, through opinion mining, to extract useful insights to help them monitor their online reputation. Existing methods of brand reputation monitoring are mostly manual or automated to perform direct opinion mining with respect to a specific brand. In contrast, comparative opinions convey much more precise opinions about a specific brand relative to its competitors. Research in comparative opinion mining is rapidly gaining traction because of its extensive range of applications in areas such as brand reputation monitoring. Past studies utilizing machine-learning approaches have largely focused on applying single machine-learning models to perform direct opinion mining, targeting opinions about single entities. Results from the resultant tools are often misleading because they disregard opinions expressed towards other entities in comparative opinion data. Mentioning multiple entities in a comparative text potentially alters the polarity of opinions towards a target brand. Typically, existing models were built and tested using a limited number of comparative opinion labels and datasets, and were applied to a couple domains. Consequently, their reported performance may not be optimal in multi-label classification problems, comparative opinion mining, other application domains, and with larger datasets. Attempts at comparative opinion mining have largely focused on comparative sentence extraction using single machine learning models, thereby not leveraging the benefits of hybrid machine-learning models. In contrast, multi-label classification and exploitation of hybrid models consisting of machine learning models and/or deep learning models have shown performance improvements in model accuracy, transfer learning, data sparsity

handling, domain adaptation, robustness, and model generalization even on complex and huge datasets. Through systematic literature analysis, data analysis, empirical analysis, and statistical analysis methods, the researcher developed and validated a hybrid machine-learning model for comparative opinion mining using datasets from multiple domains. The model was applied to brand reputation monitoring for target brands as a proof of concept. The Multilayer Perceptron (MLP), which is a deep learning model, served as the base model because of its improved flexibility in feature extraction, minimization of prediction errors, and ease of integration with single models like Random Forest (RF) that served as the top-level model. The hybrid models outperformed the single models in accuracy and f1-score across multiple datasets, leveraging count vectors and trigram features. The lowest classification accuracy was 92.1%, while the highest was 93.0%. The MLP and RF hybrid model outperformed the other hybrid models and had a prediction efficiency of 0.1 milliseconds. The statistical tests show a significant difference between the performance (accuracy) of hybrid models and single models. Engaging three human experts in validating the hybrid model revealed that the hybrid models were generally more accurate and efficient than the single models. This is because hybrid models leverage the strengths while diminishing the weaknesses of single models. Therefore, hybrid models are more suitable for applications like brand reputation monitoring.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS & ACRONYMS

Abbreviation	Description
ANN	Artificial Neural Network
BOW	Bag of Words
CBOW	Continuous Bag of Words
CNN	Convoluted Neural Network
COM	Comparative Opinion Mining
CV	Count Vectorizer
DOM	Direct Opinion Mining
DT	Decision Tree
HML	Hybrid Machine Learning
KNN	K-Nearest Neighbor
LR	Logistic Regression
ML	Machine Learning
MLP	Multilayer Perceptron
NB	Naive Bayes
MNB	Multinomial Naïve Bayes
NLP	Natural Language Processing
OM	Opinion Mining
POS	Parts of Speech
RF	Random Forest
SA	Sentiment Analysis
SC	Sentiment Classification
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency

LIST OF APPENDICES

CHAPTER ONE

INTRODUCTION

1.1 Background Information

The quest for advanced intelligent business acumens derived from consumer opinions is snowballing predominantly attributable to globalization, the swift acceptance of mobile as well as web computing tools, and stiffening rivalry owing to intensified promotion competition between brands in the same business sector (Vidya et al., 2015). An opinion is a subjective appraisal, judgment, emotion, or view regarding an entity or subject matter, and it consists of a target and sentiment (Liu, 2012a). Opinion entities consist of products and/or services of various business brands, events, issues, and topics of interest. Opinions are conveyed directly (involving one entity) or comparatively (involving multiple compared entities). An instance of a direct opinion is “iPhone is good.” In this case, the opinion has a positive sentiment towards iPhone because of the positive opinion word “good.” In other cases, opinions can target multiple entities. An example of this is “iPhone is better than Nokia.” In this example, the reviewer has a more positive opinion towards iPhone, relative to Nokia. Regular or direct sentences have no comparisons between entities. Mining of direct opinions is relatively easy because the expressed opinion relates directly with the only mentioned entity in a text.

Comparative opinion sentences, on their part, express opinions by referring to multiple entities; hence, they present unequivocal arrangements amid various objects pertaining to the definite measurable properties (Varathan et al., 2017). An instance of a comparative opinion is, "iPhone is better than Nokia.” In this example, the opinion holder (reviewer) indicates his/her preference of iPhone (entity 1) over Nokia (entity 2) based on aspects of comparison not mentioned in the review. This example used the

comparative word “better.” Comparative words establish a directional relationship between two entities in a comparative opinion. Superlatives, conversely, are used to compare more than two entities, depicting the utmost grade in terms of the compared feature. One case of an opinion using a superlative is, “Mercedes is the best brand for cars.” In this example, the opinion holder has the positive view or sentiment that Mercedes brand is better than any existing brand on matters concerning cars (a product). This comparison does not explicitly mention the other entities in the opinion text.

Opinion mining refers to the application of natural language processing (NLP), computational linguistics, as well as text analysis in the identification and extraction of subjective information from textual material (Varathan et al., 2017). This involves using specialized computational resources to explore the opinions, sentiments, attitudes, and emotions of a person. It therefore presents immense discernments for business entities (Liu, 2015). It is one of the techniques brands used to monitor their online reputation (van Doorn et al., 2010; Gürsoy et al., 2017; Passon et al., 2019). Opinion mining offers a massive potential for commercial businesses, exclusively due to the amplified acceptance and substantial electronic customer data that businesses can be able to leverage from the data. Consequently, social media analysis has its core rooted in opinion mining (Liu, 2015). The huge volumes of data associated with customer opinions is cumbersome to evaluate, calling for automated methods for obtaining generalized opinion summaries (Ravi and Ravi, 2015). Opinion mining is also known by two alternative terms: sentiment analysis and opinion analysis (Ravi & Ravi, 2015; Varathan et al., 2017; Ligthart et al., 2021; and Wankhade et al., 2022).

Comparative opinion mining is a technique that involves comparing opinions about various brands in a particular market segment or industry. Brand reputation monitoring is important for any business that wants to gain key insights into how their customers as well as their competitors perceive the brand. This aids the businesses to make timely and informed decisions, enhances their customer relations, and aids the brand to remain competitive in their respective market (Li, 2012a). Businesses can apply comparative opinion mining to perform competitor analysis by comparing the opinions attributed to their brands versus those attributed to their competitor brands mentioned in the same texts (Baziotis et al., 2017). Brands can also leverage comparative opinion mining to determine customer preferences to help them develop business strategies in favor of customer preferences and better brand reputation (Kumar et al., 2010). These lead to market positioning based on strengths and weaknesses identified in customer opinions (Lee & Jeong, 2022). Lastly, comparative opinion mining is critical in brand reputation monitoring by providing real-time feedback through programmed monitoring of brand associated opinions on platforms available online.

To apply comparative opinion mining in brand reputation monitoring, several challenges must be overcome. First, the quality of comparative opinion texts needs to be improved for use by automated tools. Second, online comparative opinion texts may be dynamic and domain-specific (Pang & Lee, 2008), making it difficult for humans to handle in an effective and efficient manner. Third, handling comparative opinions that contain superlatives require that comparative opinion mining models be trained on adequate domain-centric datasets to learn the highest number of possible entities, features, and relations that could exist in comparative opinion texts that use comparative and superlative words (Ravi and Ravi, 2015). This challenge has proven difficult for

existing statistical and lexical approaches to handle. Lastly, humans cannot cope with the vast amounts of online opinions (Ravi & Ravi, 2015). For these reasons, brands require to leverage the benefits of automated comparative opinion mining models to automatically monitor their online reputation in a more effective and efficient manner.

Beyond the traditional approach of using direct opinion mining, a brand could monitor the performance of its reputation in relation to its rival brands. This is critical in helping the brand make informed decisions in support of its business operations as well as forecasting business performance (Gürsoy et al., 2017; Chen et al., 2017; and Younis, 2015). Traditional opinion mining methods demand that human agents manually classify or categorize and summarize opinions from text data. Consistent with this, business data analysts therefore utilize various feedback platforms such as online review sites and social media platforms to extract useful opinions, which help in monitoring the broad customers' feelings concerning the brand –an indicator of the brand's online reputation (Wang & Li, 2016; Wang et al., 2015). However, this method is sluggish and costly although useful in labeling opinion datasets (Vidya et al., 2015).

In keeping with Asghar et al. (2014), distinctive computerized scrutiny of opinion data for brand reputation monitoring can be achieved through opinion mining. Because of the benefits of using computing technologies, several research works have concentrated on how automated opinion mining may profit brands. To accomplish this, past studies have dissimilar automated models with varying performance results in comparative opinion mining. This is partly for the reason that diverse performance metrics as well as datasets are applied in diverse experimentations. Numerous studies have intensely researched on the exploitation of supervised ML methods, involving algorithms such as Random Forest, and Stochastic Gradient Descent (Vidya et al., 2015). Other studies

have involved DL models like the Multilayer Perceptron (Ondara et al., 2023), Convolutional Neural Networks, Long Short-Term Memory, and Recurrent Neural Networks, in line with Gulli and Pal (2017) and Ligthart et al. (2021).

Jindal and Liu (2006) conducted a pioneering study in comparative opinion mining, proposing the comparative sentence mining (CSM) and comparative sentence identification (CSI) tasks for identifying comparative sentences and extracting pre-defined comparative quintuples (subject, comparative aspect, object, comparison type, and relation word) from user-generated content. Their approach incorrectly assumed that all comparative opinion texts have single comparative relations (Liu, Xia, & Yu, 2021). Yang and Ko (2011) applied ML models including support vector machines (SVM) to extract comparisons. Pachenko et al. (2019) and Ma et al. (2020) used comparative preference classification (CPC) for the identification of explicit comparative preference (ECP) between two entities. However, CPC required pre-annotated entities, making its implementation unfeasible (Liu, Xia, & Yu, 2021). Besides, ECP relied on keywords like “better,” “none,” and “worse,” which sometimes lead to ambiguities. To overcome this challenge, Liu, Xia, and Yu (2021) extended the work by Jindal and Liu by adding more textual contexts such as adverbs and adjectives that follow relation words like “less” or “more.” This was besides handling negation in the comparisons. They recommended the use of deep learning models in future studies.

This chapter is organized as follows: - Section 1.2 presents the statement of the problem, followed by 1.3 on justification. Section 1.4 presents the research objectives, 1.5 covers research questions and 1.6 presents the research hypotheses. Section 1.7 covers the significance of the study, while section 1.8 outlines the scope and limitations. Assumptions are in Section 1.9. Lastly, 1.10 presents the language convention used.

1.2 Statement of the Problem

Brands with online presence are keen to benefit from the mentions of their brands (including products or services) alongside competitor brands, expressed in comparative opinion reviews. leveraging the mentions can help them improve customer satisfaction and manage their reputation (Weitzl, 2019). At present, direct opinion, mining models are largely in use. Their main problem, however, is that they disregard opinions about other mentioned brand entities. In contrast, comparative opinions offer precise, competitive information about an entity based a customer's experience with other entities (Gao et al., 2018). The presence of multiple entities in opinionated text potentially changes the opinion polarity towards a brand (Yueyang & Wang, 2019). Direct opinions present opinion judgment errors (Varathan et al., 2017). Comparative opinion mining attempts to address this. Hitherto, research on the extraction of comparative elements (sentences, entities, aspects/features, and relations) has been done. Still, lexical and statistical approaches have proven inadequate in handling language nuances like unusual comparison words and comparative words like "compare" and "comparison." These problems often require machine-learning and/or deep learning models to analyze such intricate language patterns and entity relations (Bengio et al., 2013). The hybrid model's performance could be improved through manual inputting of entity names, which would be more efficient than carrying out named entity recognition. Accuracy in feature selection could be improved through a deep learning algorithm such as a Multilayer Perceptron, which has the capacity to automatically learn and extract important features from the data. A machine learning algorithm such Random Forest, with known high accuracy in feature selection capabilities could be used too. In the hybrid model, features are selected using the base model such as Random Forest.

After examining numerous state-of-the-art machine-learning and deep-learning models for comparative opinion mining, the researcher observed that the current models do not have optimal performance. Araque et al. (2017) and Liu, Xia, and Yu (2021) recommend the use of deep learning in comparative opinion mining. Most studies have used single models yet each model has its strengths and weaknesses. Performance could be improved through hybrid classification models, which have proven to outperform single models in opinion classification accuracy by leveraging the strengths of the combined single models (Wankhade et al., 2022). Despite this, the researcher still found limited studies on the exploitation of hybrid machine-learning models in COM. According to Varathan et al. (2017), 88% of online customers trust online reviews in making transaction-related decisions. However, there are few studies on the application of comparative opinion mining models in brand reputation monitoring (Varathan et al., 2017). Therefore, this study aimed at developing a hybrid machine-learning model for comparative opinion mining in brand reputation monitoring.

1.3 Justification

Existing brand reputation monitoring tools are either manual, semi-automated, or automated but with the limitation of focusing on direction opinion mining (with potential opinion class judgment errors). Existing machine learning based tools for brand reputation monitoring primarily perform direct opinion mining, ignoring the fact that 10% of user-generated content has comparative opinions that communicated precise user opinions about specific brands and their competitors (Varathan et al., 2017; Liu et al., 2021). There is, therefore a need to develop a model that performs comparative opinion mining and apply it to brand reputation monitoring for improved competitive intelligence for brands. (Z. Liu et al., 2021)

1.4 Objectives

1.4.1 General Objective

To develop a hybrid machine-learning model for comparative opinion mining based on data analysis, theoretical analysis, and empirical results from the performance of objectively selected existing machine-learning models for comparative opinion mining.

1.4.2 Specific Objectives

1. To perform an empirical analysis of objectively selected, existing machine learning and deep learning models for comparative opinion mining.
2. To develop a hybrid machine-learning model for comparative opinion mining.
3. To evaluate the performance of the hybrid machine learning model using primary and secondary data.

1.5 Research Questions

RQ1: How would a hybrid machine learning model for performing comparative opinion mining on different datasets be designed using machine learning classifiers?

RQ2: What is the efficiency of the different machine learning techniques in performing comparative opinion mining on different datasets for online user reviews?

RQ3: Which one is the most efficient opinion mining classifier based on the comparison of different machine learning classifiers (hybrid classifiers versus single classifiers)?

1.6 Hypotheses

$H0_1$: There are no statistically significant performance differences between the hybrid ML models and the single ML models in COM.

HA_1: There are statistically significant performance differences between the hybrid ML models and the single ML models in COM.

$H0_2$: There are no statistically significant performance differences in COM between the different hybrid ML model.

HA_2: There are statistically significant performance differences in COM between the different hybrid ML models in COM.

$H0_3$: There are no statistically significant performance differences in COM between the Hybrid ML models based on the choice of dataset choice.

HA_3: There are statistically significant performance differences in COM between the Hybrid ML models based on the choice of dataset.

$H0_4$: There are no statistically significant performance differences in COM between the Hybrid ML models based on the choice of feature extraction technique.

HA_4: There are statistically significant performance differences in COM between the Hybrid ML models based on the choice of feature extraction technique.

1.7 Significance of the Study

Brand managers can use the hybrid ML model for COM developed in this study to monitor their online reputation by leveraging the model's capability to analyze comparative reviews to gauge the reputation of a brand relative to a competitor brand. This is achieved through the preferred brand indicator.

A corporate or individual brand with an online presence may leverage the hybrid ML model developed in this study to detect key words and aspects that represent what online customers / consumers are engaging on more frequently whenever a certain brand is mentioned. This could help brands identify specific brand, product, or service aspects that are performing poorly for the purposes of making improvements.

For researchers in artificial intelligence, machine learning, and natural language processing, this study could be important in selecting the right approach to implementing classification models for comparative opinion mining. This is because the results of this study show that hybrid machine-learning models outperform individual machine learning models on average.

Developers of machine learning and deep learning tools can use the findings of this study to select classification algorithms and feature extraction techniques with optimal results in COM tasks. This study's findings may be the default or preferred choice.

Business competitors may use the results obtained from the hybrid ML model in this study to monitor the performance of their competitors for purposes of competitive intelligence. This may translate to getting a competitive advantage in the industry.

Policymakers could use the results of this work to influence the making of online reviews and social media data freely accessible to enhance studies in COM. The findings of this study show that people trust only reviews. Therefore, public opinions on online platforms may need to be made freely or cheaply accessible.

1.8 Scope and Limitation

1. This research was limited to applying the model to two brands (entities). The computational complexity of handling more than two entities was a limiting factor.
2. Free comparative opinion datasets involve global brands. The researcher created datasets on local brands for benchmarking (Liu et al., 2021) but the sizes were small.
3. This study focused more on understanding opinion classes in relation to brands. Due to time and technical implementation constraints, aspect mining was minimal.
4. Pioneering studies used Chinese and Korean datasets. However, due to language constraints, this work used datasets in English langue, which are fewer smaller.
5. Opinion spam detection is a multifaceted issue. The developed hybrid model handled this at a basic level, by ignoring identical opinions from the same user.

1.9 Assumptions

This study assumed that the data used had balanced opinion classes. It assumed that the opinions in the data were trustworthy. The study also assumed that the computing environment used did not affect the effectiveness of the models in COM.

1.10 Language Convention

Published literature on opinion mining uses some technical words interchangeably in spite of the inherent technical differences between them. For uniformity, this study uses these phrases or words as were found in the original sources.

Phrase	**Often, Loosely Used Alternative Phrase**
Deep Learning Algorithm	Deep Learning Technique
Feature Extraction Technique	Feature Extraction Model
Hybrid Model	Hybrid Machine Learning Model
Machine Learning Algorithm	Machine Learning Technique
Opinion	Sentiment
Opinion Class	Opinion Label / Polarity / Sentiment Class
Opinion class	Opinion polarity / Sentiment Class (Polarity)
Opinion Mining	Sentiment Analysis

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

For at least ten years now, the Internet, supported by Information and Communications Technologies has enabled people to share their opinions vis-à-vis different entities including brands, brand products and services on online platforms (Alsaeedi & Khan, 2019). These online platforms include social media sites like X platform (formerly Twitter), microblogging sites, discussion forums, blogs, and product review websites, for instance, Amazon Reviews website (Younis, 2015; Saberi & Saad, 2017). In the year 2010, Amazon published 500,000 reviews about its electronic gadgets. After three years, this number rose to 2.5 million reviews (Marneffe & Manning, 2012). By the year 2023, this number grew to 18.3 million according to McAuley Lab (2023). Data Science experts forecasted a 4300% percentage increase in the next 10 years (Chen & Manning, 2014). Consequently, various enterprises exploit online reviews and social media in monitoring brand perceptions and seeking to fight off stiff business rivalry.

A survey carried out by the American Futures Group revealed that more than 90% of Forbes top-performing 500 global companies and 82% of large businesses adopt competitive intelligence (CI) for their risk management and strategic decision-making. To identify possible risks, such firms gather and scrutinize information on their rivals' plans and products. This in turn helps the firms learn the comparative strengths and weaknesses of their individual products and/or services for purposes of designing better campaigns and products or services that can countervail those of their rivals. Traditionally, CI information was obtained from trade journals and analyst reports, both of which were in the format of press releases. Recently, news sites and competitor sites

have proved useful sources of CI information. However, the information from these two sources is limited, and its objectivity is questionable. Fortunately, with the rapid proliferation of Web 2.0, there is a rising number of consumers with the prospect to define their opinions and sentiments about products and services of brands through diverse channels and platforms, such as social media network sites, online shopping sites, and blogs. Opinions gathered from these sources are a natural source for the CI needs of firms (Xu et al., 2011).

Here is an example of a comparative opinion: *Nokia C34 has a stronger signal than iPhone 6*. Many studies have largely concentrated on identifying opinion polarities that customers express towards products. Unfortunately, this does not address the main CI problem, which is to gather and analyze information about their competitors in a bid to identify potential risks and prepare appropriate strategies to compete better. Users frequently prefer comparing several competitive products that share some functions or features (Kessler & Kuhn, 2013). Consequently, customer reviews have become a rich source of comparative opinions used to identify the strong selling points and weaknesses of products, design new business strategies and products, and analyze threats and risks from a brand's competitors (Xu et al., 2011).

This chapter entails a comprehensive review and critique of published literature relating to opinion mining with an emphasis on comparative opinion mining. Section 2.2 presents the theoretical framework, highlighting theories behind the value of opinion mining as well as those for selecting ML algorithms and feature extraction techniques. Section 2.3 presents the fundamentals of brand reputation monitoring, while Section 2.4 introduces opinion mining. Section 2.5 presents the opinion mining process model. Section 2.6 covers comparative opinion mining. Section 2.7 handles comparative

opinion mining, while Section 2.8 addresses opinion-mining approaches. Sections presents ML algorithms used in opinion mining followed by Section 2.10 that covers DL algorithms used in opinion mining.

Section 2.11 documents existing hybrid machine-learning models. Section 2.12 presents the criteria for selecting ML algorithms for classification tasks while Section 2.13 addresses the Ensemble learning method for developing hybrid ML models. Feature extraction techniques are covered in Section 2.14, followed by model / algorithm performance evaluation metrics described in Section 2.15. Statistical tools and libraries for opinion mining are described in Section 2.16 followed by datasets in Section 2.17. The applications of comparative opinion mining are presented in Section 2.18 while the challenges of opinion mining are in Section 2.19. Section 2.20 presents the specific challenges in comparative opinion mining, followed by Section 2.21 covering the research gaps in comparative opinion mining. The conceptual model underpinning this study is in Section 2.22. Section 2.23 presents the conceptual process model. Finally, Section 2.24 summarizes this chapter.

2.2 Theoretical Framework

This section presents some of the relevant theories behind comparative opinion mining, brand reputation monitoring, the criteria for selecting machine-learning or deep-learning algorithms, and the criteria for feature selection techniques in comparative opinion mining.

2.2.1 The Affective Control Theory

The meaning that a person extracts from a given word is a fundamental aspect that is associated with the Symbolic Interactionism paradigm, which avers that the actions of people towards other individuals or things relies on the meanings they ascribe to the them (Money, 2023). The meaning associated with words excludes any emotive response that such words conjure in the minds of their users. The affective Control Theory asserts that people sharing a language and/or customs within a cultural background attribute explicit meanings to some words, which may differ across different cultures. Therefore, deprived of suitable contextualization, words could convey incorrect affective meaning. A word's transient meaning is based on the context of word in its usage, which is a key aspect in opinion mining particularly in determining opinion class. In return, opinion polarity could be used to assess brand reputation.

The semantic differential technique is used to detect objective meanings of different words used within compound dimensional semantic spaces (Ploder & Eder, 2015). Each dimension represents a scale such as good versus bad and cheap versus expensive. These dimensions were most substantial in defining the peculiarities amid the several meanings of words, which diminishes the semantic space to a three dimensional cube. An instance of a 3D dimension is Evaluation in the form of good versus bad. This evaluation is behind the creation of culture-precise dictionaries for specific countries like the United Kingdom and Japan.

2.2.2 Dissonance Theory

This theory postulates that consumers have some apparent differences as they interact with each other in relation to products or services offered by specific brands. According to the pioneer of this theory, dissonance depicts non-fitting connections amidst diverse cognitions. People experience some psychological disquiet when their opinions diverge from those of other people on the same subject matter (e.g. brand). A different study holds that customers engaging with a brand whose products or services are not so distinct from their rivals often also experience this dissonance (de Vries et al., 2023) and often end up sharing their opinions with other people about the specific entities they are currently engaged with.

Thus, out of dissonance, many customers engage in sharing their opinions about the products or services of specific brands in a bid to strengthen their belief or seek a better alternative from other reviewers with better experience in using rival products and services. Brands, using such reviews, end up launching comparative advertising to help bolster their competitive advantage (Soscia et al., 2017). This theory is useful in this study as it underpins the essence of opinion generation and reaction by users. The fans or customers of a specific brand are unhappy when their brand is mentioned negatively as this is a source of discomfort or dissonance. In brand reputation monitoring, brand managers are interested in understanding the positive or negative mentions about their brands, which helps them gauge the reputation of their brand over time.

2.2.3 Self-Categorization Theory and Social Identity Theory

These two theories help people to know that the individualities of persons combined with their group affiliations have a chief part to play in shaping an individual's attitudes as well as beliefs about different things within the same social or organizational setting that promotes teamwork and socialization (Reynolds et al., 2015). Therefore, the social identity of a person is a function of more than a few factors comprising social categorization, group membership worth, and group appraisal. As a result, a positive social identity leads to positive self-esteem among the people involved, while on the other hand, a negative social identify results in enduring rivalry and social movement activities to help produce a positive perception or reputation for the group one belongs. The incentive behind an individual member of a group contributing in social comparison is for such people to develop or promote positive distinctiveness of their group (Trepte & Loy, 2017). For this reason, the researcher of this work believed that whenever people publish online comparative reviews, they are motivated by the fact that the product or service of their brand may not be serving them better than a competitor's product or service.

2.2.4 Theories for Feature Selection Techniques

One of the critical factors in ML is feature selection, which hugely affects the prediction power and accuracy of a ML model (Hamdard & Lodin, 2023). It helps eliminate additional variables in a given dataset thereby reducing the overall model complexity. The utmost factor in opinion classification accuracy is the choice of relevant features (Gürsoy et al., 2017). Hence, much attention should be given to feature selection process. Accordingly, there are two crucial theories that could aid the efficient selection of features in ML projects.

i. **Information Theory** – this theory proposes the use of the Point-wise Mutual Information to characterize related information amid data features as well as classes (Medhat et al., 2014). PMI is used to measure the association involving comparing the probability of any two events happening together to the probability of such events happening independently. In this study, this theory was used to establish the relationship between variables (brand entities and their respective opinion polarities). For instance, the probability of an opinion towards a certain brand being positive when the brand is mentioned together with another brand as compared to the probability of the brand being positive when it is mentioned alone. This is critical when considering direct opinion mining in relation to comparative opinion mining with the understanding that comparative opinion mining provides much more precise opinion polarities about a certain brand unlike direct opinion mining. In a comparative opinion-mining model, determining the opinion classes for each entity in comparative opinion text may give different opinion results than when direct opinion mining is performing, assuming that only one entity was mentioned.

ii. **Rhetoric Structure Theory** – this theory is applicable in the chunking and tokenization tasks during data preprocessing before model training (Medhat et al., 2014). This may explicate the acceptance of n-grams in addition to word vector features. This has its application behind data pre-processing, involves breaking down text into tokens and vectors. For instance, in data preprocessing when carrying out comparative opinion mining, an opinion review may be broken down into words to form n-grams such as unigrams, bigrams, and trigrams, depending on the count of words combined to form a block. These blocks of words could be opinion words, brand entities, relation words, or features for use in COM.

2.2.5 Theories for Selecting Machine Learning Algorithms

A key theory in the selection of ML algorithms is the No Free Lunch (NFL) theorem. NFL theory advances that there does not exist an algorithm that has universally superior performance across different problem domains and tasks. Every ML algorithm exhibits complex relationship with the nature of the task or problem it was developed to solve. Therefore, there is no universal algorithm for various ML projects (Rushing, 2022).

2.3 Brand Reputation Monitoring

Considering the attractiveness of social media platforms like X and online customer review sites like Amazon Reviews, it is at this time beneficial for enterprises to leverage these platforms for business insight on their clients' perception of their brands. This is for the reason that online consumers certainly utilize these platforms in making product or service reviews, or simply share their feedback on their experience with specific businesses. Their feedback often carries a positive or negative with respect to the brands or entities of interest. Any recognized negative view (opinion) could help the reviewed brand to take corrective action to care for their reputation (Sebastiani et al., 2012).

2.3.1 Brand Reputation Fundamentals

The speedy propagation of electronic commerce across the globe has been driven by the snowballing implementation of mobile as well as web computing technologies. These two technologies have amplified the ease of access to online consumer reviews. This has made it possible for different brands to have an online presence. Subsequently, online consumers have an opportunity to interact with their favorite brands. For instance, the Amazon reviews website provides Amazon's customers with a chance to provide opinions on the goods they have purchased from the retailer, through the website. The feedback is typically, as a product reviews. Gürsoy et al. (2017) maintain

that such reviews benefit likely consumers in attaining a superior appreciation of a merchandise before concluding on buying it. Brands, alternatively, exploit these online reviews to screen the opinions of their customers concerning their products and brand (Gürsoy et al., 2017; Passon et al., 2019).

Accordingly, review websites are a prospective treasure house for brands that want to gauge the perceptions and engagement levels of their customers in relation to the brand (Neri et al., 2012). Brand innovation and online user interaction influence the reputation of a brand (Al-Dmour et al., 2023). However, because online users interact based on the products and services of a brand, capturing their opinions will also lead to capturing opinions relating to innovation factors that influence a brand's reputation. This is because; opinions targeting a brand are usually above a specific product or service of that brand. The Semantic Differential model finds its use in quantifying consumer responses concerning brands (Ploder & Eder, 2015).

The connection between a brand and its clients may be delicate. Before Web 2.0 and mobile computing technologies, the institution of online product review websites and social media platforms, customers' voices were secluded from the structure of managing a brand's reputation, as their voices were not able to extend to multitudes of other customers. Fortunately today, platforms like Amazon Reviews, X platform, and YouTube allow their clients to share their opinions about an entity they are interested in (Ahmad et al., 2018). Thus, it is stress-free for a grievance (negative opinion) from a discontented customer to the reach of many people instantly. These digital platforms also make it stress-free for prompt brand reputation disruptions for cogent or irrational purposes, with the potential of heaping immense pressure on the brand’s consumer rapport by adjusting how customers perceive a brand. This is often a cause for anxiety

among reputable brands. Businesses try to uphold a decent relationship with their clienteles to mitigate their brands against market impulses including customer churn.

To exemplify the impact of online consumer reviews, in the year 2011, Taco Bell restaurants were forced to swiftly react using both YouTube campaign and additional digital sources to address an alarm raised by a particular group. The group was calling for the United States Food and Drug Administration to measure the quantity of drugs in the company's foods. In a different instance, Domino Pizza were obligated to confrontation the problem of a YouTube video showing a member of staff inserting a pizza into his nose in advance of serving it to a customer. This unfortunate event took the company's president to reply similarly, through YouTube. The president confirmed that the assertions made against the business were derived from on a hoax video, possibly, a case of intentional reputation damage by a rival. In addition, the president made an apology to his loyal customers and highlighted the actions the establishment would take against the individual or individuals behind the fake YouTube video. These examples shows that actions taken by customers or competitors regarding a business brand can quickly ruin the company's reputation, lead to loss of loyal customers and hence revenue, and by extension, cause customer churn (Ahmad et al., 2022).

Therefore, business entities need to be a little more vigilant to safeguard their reputation in the wake of the progressively increasing number of enthusiastic online roles of consumers comprising opinion influencer in addition to consumer regulator (Shang et al., 2023). Accordingly, businesses need to find a tactic of relating with their shoppers, responding to their queries, solving bogging disputes, and supporting themselves. Thus, businesses should have vigorously involved Chief Executive Officers and other teams

across the various functions to address the reputational risks linked to negative business brand mentions online.

Setiawan and Sutarso (2023) found that persistent brand reputation monitoring, chiefly, in exceedingly interactive environments like online reviews, and social media platforms, must be addressed to manage the threats to a brand's online reputation. Nonetheless, the motive for brand reputation monitoring irrespective of the platform used or data source involved should be to shape and sustain customer trust via well-timed responses to voiced concerns along with enhancing consumer trust. Tripopsakul and Puriwat (2023) found that trust issues are common in brand reputation management. Online consumers have high levels of trust in blogs. This level is lower on opinions conversed via Facebook and X platform (Tripopsakul & Puriwat, 2023).

A research carried out by Invoke Solutions generated numerous acumens into the types of central features in the advancement of trust concerning consumers and their preferred brand. Foremost, they established that a website that allowed online participants to post positive as well as negative opinions was trusted more than the websites that restricted or prohibited the posting of negative reviews. Second, they observed that good websites had superior content, making reputation monitoring more statistically appropriate. Their observations demonstrate that the size of audience or the length of one's stay on a certain website or platform does not affect trust levels. Subsequently, a website or social media platform with several reviews generates better-quality results in a bid to define customer opinions towards the business the online customer reviewed. Online reviews in addition to social media are convenient; they have enormous amounts of data, which can leverage to gauge the reputation of a brand based on the opinions shared

by the online consumers. Due to intensified commercialization of goods and services in the worldwide market, customers are additionally attracted to genuine brands. Brand-linked notions for instance brand equity and brand loyalty have been delved into. Given the attractiveness of digital social media platforms for instance X, many companies endorse their merchandise online, through these platforms (Alsaeedi & Khan, 2019).

2.3.2 Techniques for Measuring Brand Reputation Online

To measure brand reputation online, one would need to track and analyze different metrics associated with a brand's presence and perception online (van Doorn et al., 2010). This section presents some of the methods commonly used to measure the reputation of a brand.

i. **Brand mentions:** This entails regularly monitoring the rate of recurrence and context in which a specific brand is mentioned online (Lusch & Vargo, 2014). Currently, tools such as Sprout Social, HootSuite, and Google Alerts help with tracking the mentions of a brand on news sites, social media platforms, and blogs (van Doorn et al., 2010).

ii. **Sentiment Analysis:** This technique entails the application of NLP together with machine learning algorithms to analyze the sentiment of online conversations about a brand. Examples of such tools are NetBase, Mention, and Brandwatch. These tools identify both negative and positive sentiments towards a brand.

iii. **Social Media Engagement**: This entails tracking brand metrics like comments, followers, shares, and likes on a brand's social media platform. A strong reputation and strong customer loyalty are depicted by high levels of social media engagement. Social media analytic tools like Buffer are applicable in measuring social media engagement levels pertaining to a brand (van Doorn et al., 2010; Smith et al., 2012).

iv. **Online Reviews and Ratings**: This entails tracking and consequently analyzing both online reviews and online ratings of a brand on sites like Amazon, Google My Business, and Yelp. The baseline is that both high ratings and positive reviews are deemed to depict a positive (good) brand reputation while low ratings and negative reviews translate to negative brand reputation (van Doorn et al., 2010).

v. **Web Traffic**: This entails measuring the volume and sources of web traffic flowing into a brand's website. Useful tools for this include Google Analytics, which tracks and identifies sources of referral traffic to a brand's website (van Doorn et al., 2010).

2.3.3 Online Reviews and X Platform as Sources of Opinion Reviews Data

There are four leading gains for brand monitoring via social media data and online reviews:

1) The X platform provides a rich source of comparative opinions for brand reputation monitoring (Fronzetti Colladon, 2019).
2) At times, users create X platform handles, targeting a specific brand to discuss a particular issue. This sometimes brings about viral content that can shape or kill a brand's reputation depending on the overall opinion polarity (Kaur, 2016).

3) Social media platforms including X and YouTube have vast data volumes. This provides sufficient data for reliable model testing and validation (Liu et al., 2017).
4) Finally, X platform data is available to developers via a Developer API. At the time of carrying out the pilot test, X platform data was freely available. However, later, when branding changed from Twitter to X platform, charges were introduced for accessing tweets. Nonetheless, opinion data on X platform is good for brand reputation monitoring using opinion mining tasks and processes (Das et al., 2018).

There are many reasons for choosing Online Reviews

1) Effectively, the whole lot of information on online reviews relate to a brand (a brand's products or services). This makes it easier for ML algorithms to detect entities like brand names, and service names to increase operational proficiency.
2) Products registered on Amazon.com, for instance, have accurate product / brand names. A classification algorithm applied to such data would demand less computational resources and still achieve high efficiency and effectiveness.
3) Online reviews such as Amazon reviews website often have a five-star rating system, which can be applied to cross-reference, confirm, or moderate outcomes from opinion mining. A 5-star rated review should have a positive opinion / view.
4) Online reviews are accessible via a Developer's Application Programming Interface (API).

2.4 Opinion Mining

As stated by Liu et al. (2017), a useful review involves some form of argument. Captivatingly, online reviews carry proof and motives behind one's opinion about a reviewed product / service besides just the sentiment. Ngo-Ye & Sinha (2014) applied a mishmash of text data and features like fiscal value to aid envisaging the efficacy of analyses acquired from online review websites like Amazon Reviews. This study benefitted from appreciating that online reviews are advantageous to online shoppers interested in making acquisitions at e-commerce platforms like Amazon and EBay.

2.4.1 Introduction to Opinion Mining

An opinion is a personal statement, assessment, view, feeling, or attitude expressed towards an entity from the perspective of an opinion holder (Liu, 2012a; Sebastiani et al., 2012). Opinion mining, which is otherwise also known as sentiment analysis is a common research area in research fields for instance NLP, Information Retrieval, Data Mining, Text Mining, and Web Mining for the extraction and analysis of diverse opinions frequently presented in the form of text (Saranya et al., 2016).

According to (Varathan et al., 2017), opinion mining is the application of NLP, Computational Linguistics, in addition to Text Analysis in the identification and extraction of subjective information found in textual content or materials. Opinion mining is therefore a key component in the advancement of better-quality products and services as well as aiding good business administration. Repeatedly, this leads to the analysis and classification of opinions about a certain entity (ideasov.com, 2019).

2.4.2 Applications of Opinion Mining

Recent times have witnessed a rapidly growing interest in opinion mining because it offers a plethora of tools for the scrutiny of public opinions on different topics (Varathan et al., 2017).

i. Business intelligence – the utilization of opinion mining in an enterprise to obtain the views of their customers towards different product and/or services for purposes of improving product quality or other aspects like pricing model (Diamantini et al., 2019).
ii. Would-be customers of a given merchandise or service benefit from the opinions of past customers in deciding whether or not to buy a product / service (Vidya et al., 2015)
iii. Governments could use opinion mining to understand the wants and needs of their citizens for the purposes of prompt action (Arunachalam & Sarkar, 2013).
iv. Politicians could benefit from opinion mining in their bid to understand the thoughts of their voters regarding them (Laver, Benoit, & Garry, 2003).

There are four common types of opinions:

i. *Direct opinions* – these are opinions about a certain entity without comparing that entity with another entity. These opinions are also known as regular opinions. For example:

   "The durability of a Dell Laptop is good."
ii. *Comparative opinions*, on the other hand, are opinions concerning an entity but with respect to comparable entity. In essence, this is achieved through comparative statements. Here is an illustration of a comparative opinion statement:

   "The durability of Dell laptops is better than that of HP laptops."

iii. *Explicit opinions* – opinions stated unequivocally in a statement/sentence. For instance:

"The resolution of Sony Cameras is amazing."

iv. *Implicit opinions* – opinions conveyed indirectly in a neutral sentence. For instance:

"The smartphone stopped charging yesterday."

2.5 Opinion Mining Process Model

This section describes the generic process of performing opinion mining. Notwithstanding there being considerable exploration in the area of opinion mining, the researcher found limited information that could help create a standardized model for comparative opinion mining. A case in point, Arora and Bansal (2020) used a six-step model. In their model, the initial step is data extraction. This is then followed by pre-processing, results and finally, making inferences. Then again, El Haddaoui et al. (2018) preferred an opinion mining model that is made up of four steps. Their model begins with data acquisition and finishes with results visualization.

However, from literature analysis, the entire model of opinion mining entails the subsequent major steps as presented: Data Collection, Data Preprocessing, Feature Extraction, Opinion Classification, Model Validation and Evaluation, as well as Integration (Kane et al., 2016). These steps may still be classified into four: Data Gathering, Data Preprocessing, Model Training, Opinion Classification, and Model Results (Kalaivani & Thenmozhi, 2019). Based on the model described earlier, for instance, data pre-processing includes feature extraction while the model excludes the integration phase as found in the model by Kane et al. (2016). Bjurstrom and Plachkinova (2015) developed the Desktop Methodology and endorsed it for the

development of opinion mining models. They used X data. The methodology involved three phases: Data collection, opinion mining, and visualization. Data pre-processing, model training and classification are grouped into a single step.

2.5.1 Data Collection and Feature Extraction

Opinion mining begins with the collection of opinion data. This data is used to develop an opinion-mining model, a process that involves training an algorithm on the collected data to learn the features that could help it predict opinion classes when presented with new unlabeled data. After data collection, the process involves feature extraction. Features may include opinion words, entities, natural language features like syntactic and semantic features, and machine features like n-grams. These features are obtained from textual data. The classification algorithm (also known as the classifier) then converts the mined features into a statistical representation. This numeric representation is usually in the form of a vector. Usually, each vector element denotes the rate of recurrence of a word as found in a lexicon (dictionary of polarized words. This is the vectorization process, also known as the feature extraction process. Characteristically, this consist of by means of bag-of-n-grams. A more topical improvement in feature extraction saw the creation of word embedding or word vectors to help capture synonyms of words with the sole aim of enhancing opinion classification accuracy.

2.5.2 Model Training

To carry out opinion mining, a classification model is first developed. This involves model training and validation using opinion data. During model training process, the selected algorithm learns from the presented text input, which is known as a training dataset. This dataset corresponds to precise outputs with respect to the test samples – provided through model training. The feature extractor transfers the input into a feature

vector. The sets of feature vectors with their associated outputs or tags are input the ML algorithm to create the ML model. During prediction, the feature extraction technique converts unfamiliar inputs to corresponding feature vectors. These are then input the ML model to produce predicted outputs. Usually, the outputs are in the form of opinion classes, which include positive and negative for binary classification or positive, negative, and neutral for trinary classification.

2.5.3 Classification of Opinions

This step consists of using ML models like Random Forest (RF) or DL algorithms such as Multilayer Perceptron (MLP) to classify opinions from the provided data in keeping with the corresponding opinion polarities.

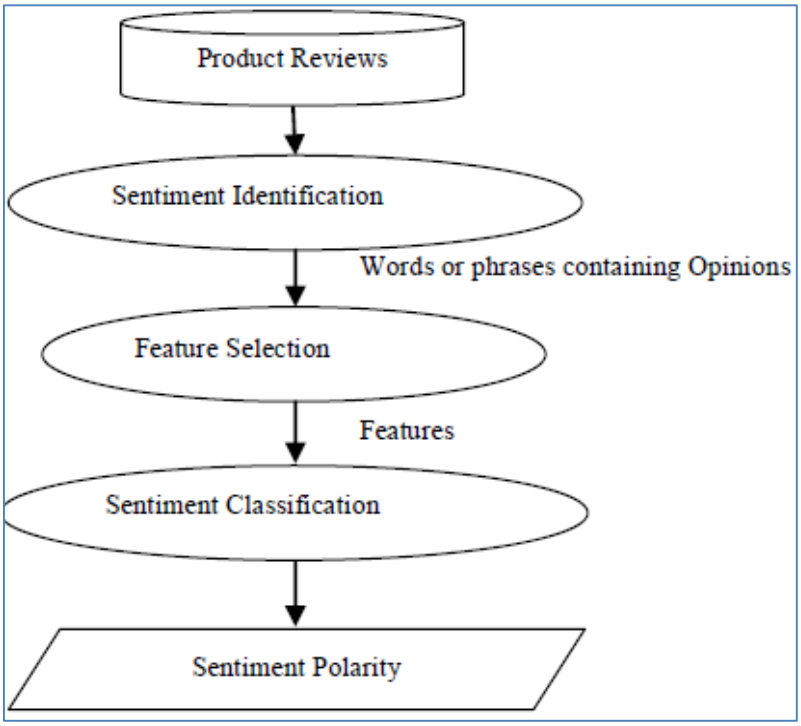

Figure 2. 1 Opinion Mining Process (Source: Rathor et al., 2018).

Upon effecting this stage and classifying opinions according to their polarities, the ensuing phase involves making a summary of the opinions gotten. Along with Arboleda et al. (2017), the bottommost level of opinion summarization yields raw summative

opinions specially for several reviews thus finalizing with the aggregate numeral of opinions that are positive versus opinions that are negative. This approach is common. One more method involves mining descriptive words or phrases or sentences that are most regularly be revealed in the dataset. In conclusion, summarization of opinions may concentrate on displaying how the extracted opinions concern a definite product or subject contained in the source of data.

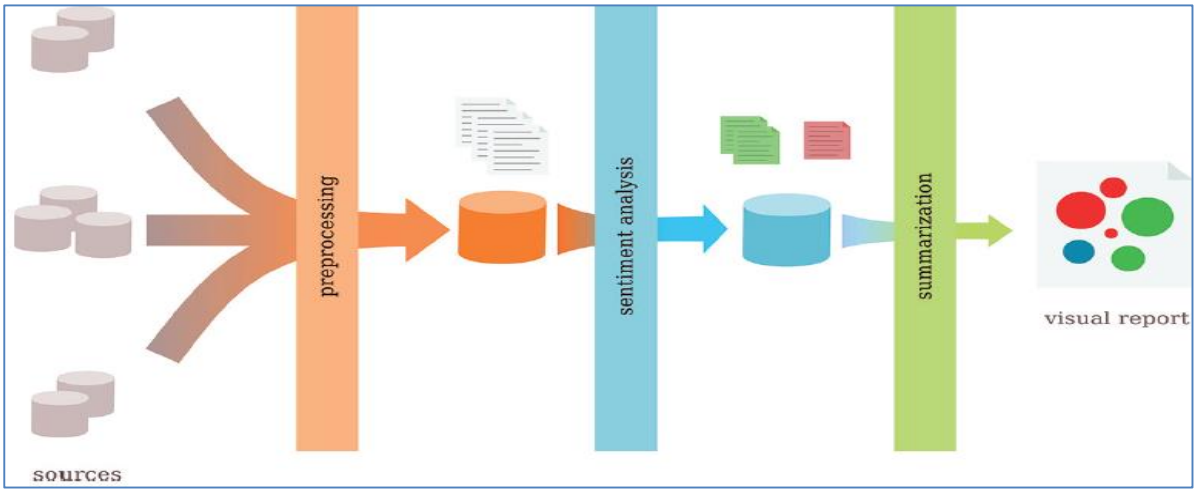

Figure 2. 2 The Process of Opinion Mining (Source: Rathor et al., 2018).

2.5.4 Lexicons or Dictionaries for Opinion Mining

There are many lexicons in use in opinion mining. These include but is not limited to Opinion Lexicon for Opinion Mining, SentiWordNet, Emoticon Sentiment Lexicon, and Sentiment Lexicons for 81 languages.

2.6 Comparative Opinion Mining

This section presents the fundamental of COM, how to mine comparative opinion elements (sentences, entities, features, and relations).

2.6.1 Fundamentals of Comparative Opinion Mining

Comparative opinion mining is a sub-field of opinion mining concerning the identification and extraction of information that is expressed in a comparative way (Varathan et al., 2017). They present a comparison of at least two entities such as brands from the perspective of an opinion holder. An example of a comparative sentence is "Nokia smartphones are more durable than Samsung smartphones." This sentence depicts a user's attitude towards the two brands based on the durability aspect of their smartphones. In the evaluation of various entities, comparative opinion mining is very significant given that it offers a point of reference for the comparison. Unlike direct opinion mining, which has witnessed significant research, COM has not been widely researched as it is an emerging field that concentrates on the expansion of approaches and tools for the automatic detection of opinionated information and subsequent determination of the opinion polarity towards a particular target (Varathan et al., 2017). Often, an opinion target is a named entity like a brand, topic, or event. This makes opinion mining a suitable area of interest for researchers interested in extracting valuable information regarding public views.

While comparative opinions are often expressed using comparative or superlative adjectives or adverbs, there exists comparative sentences that implicitly use comparative words or even those that use comparative words in a non-comparative sentence (Varathan et al., 2017). As such, mining comparative opinions is an exciting exploration area since comparisons are used in marketing and brand monitoring. Comparative opinions necessitate a dissimilar definition than direct opinions (Liu, 2015a; Xu et al., 2011). The following are five examples of comparative opinions, showing various language nuances in depicting comparisons between two entities.

Example 1: *Nokia N95 has a better camera than iPhone.*
> (Nokia N95, iPhone, camera, better)

Example 2: *Compared with Nokia N95, iPhone has a better camera.*
< (Nokia N95, iPhone, camera, better)

Example 3: *The Pearl and the Curve are both with high resolution camera.*
~ (Pearl, Curve, camera, high resolution)

Example 4: *The screen of iPhone is bigger than that of the curve, so I can read easily.*
> (iPhone, curve, screen, bigger)

Example 5: *The price of iPhone is much higher than that of the curve, so I can not afford it.*
< (iPhone, curve, price, higher)

For each instances above, the product features and the relationship between the two entities are placed in parentheses. For instance, Example 1 shows that there are two entities in the comparative text: Nokia N95, and iPhone. The feature upon which these two product entities are compared is the camera of the two phone products. The entity relation uses the comparative word "better" to separate Nokia N95 to the left and iPhone to the right. The flow of the opinion polarity indicates that the first entity (Nokia N95) and the second entity (iPhone). The ">" shows that the first entity is better than the second entity. The "<" shows that the second entity is better than the first entity. One could infer that the brand behind iPhone (i.e. Apple) needs to improve the camera quality of the compared iPhone to avoid revenue loss (Xu et al., 2011).

2.6.2 Mining Comparative Elements

In opinion research, comparative opinions are characterized by four elements, which require special techniques for handling, according to (Varathan et al., 2017). These four elements constitute four difficult tasks that are not easy to compare hence the need for different techniques to detect and handle them separately. These four elements constitute comparative opinion data. For this study, therefore, at least two entities were

required to ensure that there are relations to detect and analyze in order to establish directional opinions. In this study, the entities were represented by brand names. The features were the brand, product, or service aspects mentioned in the opinion while the relation is attributed to the opinion holder's attitude towards one brand relative to another by use of comparative words and phrases.

Table 2.1 Elements of Comparative Opinion Mining

Element	Description	Technique
Sentence	This is a comparative sentence.	Sentence detection
Entity	Comparative sentences have multiple comparable entities (e.g. brands) mentioned together.	Entity detection
Relation	This is the association/link between or among the entities mentioned in the comparative sentence.	Relation detection
Feature	This is an aspect forming the basis for the comparison of entities mentioned together.	Feature detection

2.6.3 Comparative Opinion Sentence Detection

A pioneering study on mining or extracting comparative sentences was carried out by Jindal et al. (2006). The intent in this study was to detect and extract sentences that had comparisons between entities. However, a study by Li et al. (2011) attempted extracting product comparisons from comparative texts. In general, research works spanning nine years (2006 to 2015) on COM concentrated on comparative opinion sentence detection using various approaches. The approaches studied included machine learning, lexicon approaches, and associative rule mining. The focus was largely on supervised learning methods like lexicon approach as well as association rule mining (Wang et al., 2015).

2.6.3.1 Sentence Detection Using Machine Learning Techniques

Recently, the primary techniques for detecting and extracting comparative opinion sentences are machine-learning based. This is evidenced in the upsurge in the number of studies exploring the application of machine learning techniques in comparative sentence detection or extraction. For example, SVM, which is a supervised machine learning technique to detect comparative sentences (Wang et al., 2015). However, this technique was integrated with a rule mining technique. A different study worked on identifying comparative sentences in Chinese texts using rule mining approaches (Sebastiani et al., 2012). In a number of these studies, rules were applied while in other studies; NLP approaches like syntactic and semantic rules were incorporated in identifying comparative sentences for final extraction using a ML technique.

2.6.3.2 NLP Techniques

A few studies have covered the application of syntactic approaches to comparative sentence detection. Gu and Yoo (2010) who used statistics and sentence structures to identify comparative sentences carried out one such study. In their study, any sentence that contained specific comparative words like “more” and “than” were extracted for eventual classification as being comparative sentences.

2.6.3.3 Association Rule Mining

Liu et al. (2013) studied the application of rule mining technique that was supervised learning based on the identification of comparative sentences. Again, this study involved Chinese texts, using syntactic structures. In later studies, sentence structures were explored. Dependency trees have also been studied to establish their effectiveness in mining comparative sentences. Unfortunately, for quite some time, past research focused on mining comparative sentences instead of comparative opinion sentences.

Sentences can have comparisons without any opinions stated in the sentence. Such sentences should not be mined or identified as comparative opinion sentences.

2.6.4 Entity Detection

This yields valuable information; which consumers need when making decisions on what product or service they should opt for while the same information helps business organizations to identify their present business rivals in a bid to handle their competition better. In the context of opinion mining, an entity has many representations such as a product, person, service, or other. However, for purposes of entity detection, the focus is in identifying the name of a brand or model. In comparative sentences, the identification of entities is a challenging task because in user-generated content, the text is often in non-standard form, the names are not always properly written or abbreviated, the spellings may be erroneous, and the grammar of the text may be improper. Thus, in comparative opinion mining, the researcher need to be more careful when performing entity detection to overcome the above nuances. Thus, entity detection in comparative opinions requires much more specific or specialized techniques than when dealing with handling formal texts since comparative opinions are expressed in informal language. There are four popular ways of carrying out entity detection for COM:

2.6.4.1 Entity Detection Using Machine Learning Techniques

Conditional Random Fields (CRF) technique has been applied to entity detection in comparative sentences. CRF is applied in entity identification by means of end word present in every element phrase in a specific domain knowledge. Liu et al. (2013) found that the CRF is better than the baseline technique in terms of recall and precision that

techniques that that rely on every word in a sentence to detect an entity. However, the CRF technique has only been used successfully to retrieve entities in Chinese texts text.

2.6.4.2 Entity Detection Using Lexicon Techniques

The lexicon approach is the most popular in entity identification. For example, Xu et al. (2011) used custom-generated lexicon dictionary consisting of names of mobile phones and specific attributes to make it possible to perform entity detection because entities can be identified by different names. For example, in online reviews and tweets, "iPhone 6 plus" is also identified as "iPhone6+" or "6+." In their dictionary (lexicon), they encompassed various variations of the product's name, brand name, and abbreviations of the product's name. A closest-first method was used to address the challenges of words referring to words like "it" or "the" referring back to a previously mentioned entity. To avoid duplication of entities, entities mentioned differently, such as "S2" referring to "Samsung 2" were addressed via anaphora through post processing.

Tkachenko and Laus (2014) realized that comparative entities ought to have been mentioned together before. Their focus was on sentences with strictly two entities using custom-made dictionary based on the matching approach especially because there were no named entity relations in the texts they were studying. Fieldman et al. (2017) identified entities on the basis that such entities must have been mentioned together in the same sentence. In their approach, a dictionary was created to capture all versions of the entity names used in the text. However, this approach has the limitation that it is not possible to capture all variations of an entity. To overcome this challenge, the focus shifted to using brand names as opposed to model names or product names. This is because product or model names often have variations while brand names often remain relatively the same hence yielding better results in entity detection. In fact, this is why

simple pattern matching applied to a dictionary or lexicon of brand names combined with POS tagging gave better results in entity detection as opposed to other techniques.

2.6.4.3 Entity Detection Using Rule-based Techniques

Rule-based techniques for named entity recognition depend on the use of dictionaries, predefined patterns, and grammatical rules in the identification of named entities (Sureka et al., 2009). This technique involves the use of nouns and pronouns to detect and extract entities in comparative sentences. This task requires manually selected comparative keywords, POS tags such as comparative adjectives and adverbs as well as superlative adjectives and adverbs. The rules specified based on these can be used to identified entities where relationships based on the above POS tags. The main limitation of this technique is that some texts or sentences considered comparative by this approach are unfortunately non-comparative but use comparative or superlative words. Such texts or sentences contain facts and only include comparative or superlative words, which does not make them opinionated comparative sentences. The technique is therefore misleading in this aspect (Jindal et al., 2006).

2.6.4.4 Entity Detection Using Text Mining Techniques

In this approach, every sentence is considered as having an object and a subject, which can be from the brands or models of a product. To differentiate between object and subject in a comparative sentence, the similarity value was used. If an entity has a high similarity value it is considered as the subject while the entity possessing a subordinate similarity value is treated as the object. Then, the Jaccard coefficient is applied to mine the entities found in comparative sentences (Varathan et al., 2017).

In conclusion, whereas correct entity identification or detection is key in comparative opinion since it sets the basis for handling relation detection and feature detection, the challenge is finding a reliable technique. As of 2017, CRF techniques were the only ways to detect entities in comparative opinionated texts when using machine-learning approaches. While the use of POS tag analysis, dictionaries, and domain knowledge may work, the process is tedious and in the world of big data, some entities may not be detected. Further to this, all past studies have only focused on comparative opinionated sentences with only two entities. For these reasons, further studies would be needed to explore other ML techniques like Random Forests in entity detection as well as the attempt to process sentences with more than two entities.

2.6.4.5 Entity Detection Using Deep Learning Techniques

DL models like BERT could leverage contextualization of information and the relationships in words to identify entities and their relationships with opinions (Devlin et al., 2019).

2.6.5 Relation Detection

The task of relation detection is paramount in the sense that it aids in establishing relationship between entities as well as how the entities are ranked in a comparative opinion text. A major benefit of relation detection is competitor analysis as it provided relevant information about how one entity relates with a competitor entity. On their part, customers will be able to compare products, services, or brands to determine which one is more suitable for their needs. From past studies on relation detection, machine learning techniques, lexicon techniques, graphical modeling techniques, and one-side association techniques have been attempted.

2.6.5.1 Machine learning techniques

A generated model was created and used to handle both entity-level and sentence-level comparisons based on both supervised and unsupervised ML techniques. This model managed to detect the preferred entity. At first, they attempted doing this by making use of the bag of words model, which proved inadequate. Then, they used the sigmoid ranking function to achieve better results in detecting the relation between entities in a comparative sentence (Tkachenko & Lauw, 2014). An entity "A" that is preferred to entity B does not automatically become the favorite entity in a set of comparative sentences. This is similar to how graphical models perform (Kurashima et al., 2008).

2.6.5.2 Lexicon techniques

Many researches have attempted using this technique in performing relation detection for use in comparative opinion mining. For example, for relation detection to be possible using lexicon approach then five elements must be present. This includes the relation word, entity 1, entity 2, features, and type (Jindal et al., 2006). In a comparative statement, entity 1 and entity two are linked with a relation word. The type element shows if the relation is equative, equal gradable, non-equal gradable, or superlative. The limitation of this method is that the rules for using this approach are inadequate.

2.6.5.3 Graphical model techniques

Kurashima et al. (2008) created a graphical model to represent entity relationship without considering features of such entities. Then, they ranked the entities against competing entities. Their model relied on query done by a user on one or many entities to show the general comparative relationship. One of the earliest studies on graphical model techniques proposed the application of SVM-based map to detect entity relations. In their proposed model, the map composed of entities, entity attributes, and

relations between entities based on the attributes. Their model was capable of identifying the existence of comparative relations using attributes linking entities together. While in their study, they were able to detect relations between three entities; their primary focus was on direct (simple) relations. Moreover, they used a small dataset of 217 comparative relations that they obtained from Amazon. Such a small dataset is not good for training machine-learning models because the performance of it on larger datasets could be significantly different (Xu et al., 2009).

Another graphical model was developed later to help with the interpretation of relations between entities. This model identified all entities in the dataset used. The benefit of this visualization model was that businesses could be able to use it tracking the strengths and weaknesses of their competitors versus their own strengths and weaknesses in relation to the products or services mentioned in comparative reviews. The challenge was that it could only handle direct relations such as “I like Nokia Smartphones.” This statement does not show relational direction hence difficult to establish entity preference. Later, they researched on detecting directional relations in comparative opinion reviews for business intelligence. Their new approach used a two-level CRF model that had no fixed interdependencies. This model proved more powerful in extracting useful relations but the performance was not very good (Xu et al., 2011).

Another attempt at graphical models involved developing a model to compare entity relations based on aspects like features, ease of use, and design. Every of the aspects was represented using an integrated graphical model that considered all comparative appraisals. This approach benefitted customers in identifying products with superior aspects based on feature superiority scores. Their models integrated K-means clustering with SVM as some of the techniques to attain better results. The limitation with this

approach was that some products features were irrelevant in helping customers make product choices. For instance, performance as an aspect or feature does not work for products like clothes and books. Another research that is similar to this one was done, relying on a search query to retrieve all competitor entities and features that had similarity with the search query results. With the entities and features ranked, users were able to have options from which they could choose their preferred entity or feature (Tkachenko & Lauw, 2014).

2.6.5.4 one-side association techniques

This approach was used to detect relations in comparative opinion texts and help identify the preferred entity. The focus of this study was to help consumers select the best options. The approach used to relate entities in a sentence based on opinion aspects in a sentence, which was derived from context dependent opinion that showed if a given comparative word bore a positive opinion or negative opinion. For instance, the existence of the comparative word "more" is not always indicative of a positive opinion without considering the context in which it is used and the entity it references. Therefore, in their model, a count of the frequency of co-occurrence of both features and comparative words to establish one-side associations. The model considered synonyms and antonyms of words. To detect preferred entities, adverbs, adjectives, and features were used. Negation was used to identify potential reversal of preference in a comparative sentence.

In conclusion, most of the studies on relation detection have used machine-learning approaches. Fewer studies used lexicon-based approaches while others used rule-based approaches. There is therefore room to study other approaches including unsupervised machine learning techniques, text-mining techniques, and deep learning techniques.

Identifying which entities co-occur together could be very helpful to businesses that want to know their rivals while consumers would be able to identify the best product or service quickly.

2.6.6 Feature Detection

This section explains the various techniques used in feature detection. This includes words, machine learning features, natural language features, statistical features and pattern matching.

2.6.6.1 Words as Features

With the wide-ranging application of opinion mining in the programmed classification of several online reviews, a majority of the present-day machine learning models use Random Forest and Stochastic Gradient Descent algorithms among others. The models assign classes to every user-generated opinion granting in a few cases, the attention of polarity is a sentence instead of the complete text. According to Carrillo-de-Albornoz et al. (2018), the bulk of these scholarly works have concentrated on choosing the most suitable and appropriate features since, herewith, the model's performance can be enriched. The bag-of-words (BOW) feature model is a predominant classification feature model in opinion mining. An improvement on the BOW model led to the creation of what is known as the Term Frequency — Inverse Document Frequency (TF-IDF). TF-IDF is premised on computing the value of each word in a piece of text (Pang et al., 2012). A different common feature is the sentiment-based dictionaries and subjective words or word-phrases (de Albornoz et al., 2010). Others consist of part of speech (POS) words and opinion words (Taboada et al., 2011).

Words are the central or core features in opinion mining (Kane et al., 2016; Kauffmann et al., 2019; Tripathy et al., 2017; Pantano et al. 2019). Conversely, since there are characteristically various kinds of words in comparative reviews, it aids to stipulate the specific words that are relevant to opinion mining. The key factor in classification model accuracy for ML algorithms is the choice of important features. Accordingly, the normal approach to feature selection is the utilization of a combination of vector representations of selected feature, which are classically referred to as n-grams (Almatarneh & Gamallo, 2018). In the n-gram model, if n = 1 then the features are one-word phrases (or unigrams). If $n = 2$ then there are two uninterrupted word phrases as features, that are called bi-grams. If the $n = 3$ then there are three consecutive word phrases (tri-grams). Hence, these three are the most frequently used features in opinion mining (Kalaivani & Thenmozhi, 2019). Features may take other forms and meanings. Polarity lexicons, bag of words, and word embedding among other features are vital in the detection of advanced opinion polarities like “least negative”, and “most positive” in addition to their application in the commonly-used positive, neutral, and negative opinion classes (Almatarneh & Gamallo, 2018).

Extracting and analyzing opinions, subjective information, and emotions from textual content is the foundation of opinion mining. Textual content consists of both natural language features and machine features that play diverse roles in seizing these distinctions within opinions. Two most prevalent features are opinion terms and POS tags (Varathan et al., 2017). Opinion terms and POS tags are the most popular features. Other features include keywords, statistical feature words, manual rules, words, entity types, sequence patterns, comparative words, and class sequential rules. For improved

nuanced and accurate opinion classification, it is important to combine natural language features and machine features.

2.6.6.2 Machine Features

Features that incorporate structural and quantitative attributes extracted from text content are referred to as machine features. They provide important patterns and contexts that help in opinion analysis and ultimately, opinion classification.

i. N-grams and Bag-of-Words: the extraction of contiguous sequences of n items (i.e. n-grams), together with representing these words using bag-of-words helps in capturing patterns of word frequency (Mikolov, Chen, Corrado, & Dean, 2013)

ii. Part-of-Speech Tagging (POS Tagging) – this is helpful in the identification of grammatical structures within a given sentence for purposes understanding of different opinion nuances (Hutto & Gilbert, 2014).

iii. Character and Word Counts – the frequency of characters and words can offer insight into the intensity and length of an opinion (Pang & Lee, 2008).

iv. Punctuation usage – punctuation marks in text can also indicate tone, emphasis or emotion (Kiritchenko et al., 2014).

v. Emoji usage and capitalization – emoji and uppercase letters can convey intensity and opinion class (Novak et al., 2015).

vi. Readability Metrics – Flesch-Kincaid readability score is one of the readability metrics that can give insights into the difficulty of opinions (Flesch, 1948).

2.6.6.3 Natural Language Features

i. Opinion Words – to reveal sentiment orientation (opinion class), the identification of explicit opinion expressions such as “I love” is useful (Wiebe et al., 2005).

ii. Opinion Words –opinion polarity is directly affected by the presence of positive and negative sentiment words (Liu, 2012b).

iii. Subjectivity Markers – in the identification of subjective opinions, it is important to identify subjectivity markers like “in my opinion” or “I think” (Wilson et al., 2005).

iv. Negation handling – it is important to handle negated phrases or any negation in text to achieve accurate opinion mining (Taboada et al., 2011).

v. Dependency Parsing – to understand how entities relate with opinions in the same text, it is important to analyze the grammatical dependencies existing between words (Manning & Schütze, 2000).

Since feature information is critical in comparative opinion mining, product or service features are often considered. Features are the foundation for decision making when comparing entities. For example, “Nokia phones are more durable than Samsung phones” provides durability as feature for comparison and this is useful information for those interested in buying durable phones. Features therefore add significance to entities in comparative statements.

2.6.6.4 Statistical techniques

Different attempts have been made to detect features for purposes of performing comparative opinion mining using statistical techniques. Sun et al. (2009) built a database of features based on product-related sentiments. In their model, products were compared using features stored in the database. To support customer requirements in comparing products based on features / aspects, tree models were built. Using past

reviews, a recommender system was proposed that accounted for all features in previous reviews. Using product features from Amazon, supported by an evolutionary tree to establish how a product evolves over time, they were able to create a system that informed of the evolution of a product based on its features. Unfortunately, this approach did not address the issue of synonyms of features found in the various reviews. Synonyms like “photo” and “image” were wrongly treated as different features.

2.6.6.5 Pattern matching techniques

A generative model encompassing information derived from the sequence of specific features was suggested by Tkachenko and Lauw (2014). The features are then categorized into negation features and syntactic features. In this approach, negation features were associated with targeted words while syntactic features were aimed at analyzing the position of words in a sentence. Frequently used product features were included. The findings revealed that word sequence is imperative in feature detection.

Until 2017, statistical techniques and pattern matching techniques were applied in feature detection for comparative opinion mining. Whereas statistical approaches yielded more feature aspects, pattern matching was instrumental in the identification of the preferred entity using pre-identified features. Therefore, potential approaches or techniques such as topic modeling could be used in feature identification for comparative opinion mining using comparative opinion reviews (Varathan et al., 2017).

2.7 Comparative Opinion Mining Approaches

Three approaches to COM have been explored in the past. They are:

1. Machine Learning Approaches
2. Rule Mining Approaches
3. Natural Language Processing Approaches

2.7.1 Machine Learning Approaches

Machine learning is a sub-field of computer science that involves the study of algorithms and methods that make predictions based on training data from which they learn. Therefore, machine-learning techniques are employed to perform different tasks in COM. For instance, ML algorithms have been applied in the identification of comparative sentences from a collection of both comparative and non-comparative sentences. Some of the most common ML algorithms in comparative opinion mining are Random Forest, Naive Bayes, and Support Vector Machines (SVM).

The ML approach generally utilizes supervised classification algorithms to detect opinions and classify according to opinion polarities (classes), which include positive and negative classes, or positive, negative, and neutral classes. This is known as opinion polarity classification (Yueyang & Wang, 2019). For this method, categorized data (pre-annotated data) is appropriate to train the algorithms. The context of words is considered. For example, an opinion review with 'not good' reveals a negative opinion while that with 'very beautiful' communicates a positive opinion or view.

2.7.1.1 Supervised Learning

A ML method that learns a function from a training dataset with the goal of predicting the class attribute for some unlabeled data is said to be a supervised learning method. Such machine learning classifiers undergo training on features extracted from a training dataset. Opinion terms like great and love, together with the POS tags are among the most popular features employed in annotating words with the correct syntactic behavior. The result of POS tagging is the identification of words that are verbs, as well as adjectives and adverbs. An example of a supervised machine-learning technique is SVM, which has been applied to classify sentences as being comparative or non-comparative or simply to extract comparative sentences from text (Wang et al., 2015). Xu et al. (2009) used a multiclass SVM model for use in comparative relations classification for more than two classes. In their model, entity types, POS tags, and opinion words were used as features.

Using customer reviews, Xu, Liao, Li, and Song (2011) used CRFs to extract comparative relations. A graphical model utilizing two-level CRF extracted and visualized comparative relations between products. The features used included opinion words, entities, and relations between entities. This model was also applied on blogs, SMS, emails, and epinions.com. Liu et al. (2023) also applied CRFs to extract comparative elements in comparative sentences. They used opinion words, domain-specific words, POS tags, and entity – comparative word distance as features. Naïve Bayes classifier was applied as the classification algorithm as it is more suited for datasets with many dimensions. This method uses the Bayesian theorem and computes the probability of a class following a bag-of-words approach. Another studies that have used the Naïve Bayes model includes Tkachenko and Law (2014).

2.7.1.2 Unsupervised Learning

This type of machine learning technique or algorithm learns hidden structures in unlabeled data. A popular unsupervised approach is clustering, which is used to group similar data into the same group. A popular algorithm for clustering is K-means clustering. It is used to group similar data into k clusters, ensuring that each data or observation is found in a cluster that has the nearest mean. Besides the generative model created by Tkachenko and Laus (2014) for performing comparative opinion mining, past studies have not delved much into the application of unsupervised learning methods to perform comparative opinion mining.

2.7.1.3 Reinforcement Learning

This is a sub-field of ML that involves how agents perform actions within an environment with the intent of maximizing cumulative reward. The agent interacts with the environment following specific steps: observing current state, choosing appropriate action determined by a policy, and getting rewards as feedback. Based on the feedback, the agent is guided to learn the specific actions that would lead to the most rewarding outcome (Sutton & Barto, 2018). Reinforcement learning has been applied to aspect-based opinion mining. In this case, the approach is used to detect certain features or aspects of an entity mentioned in a piece of text. For example, a smartphone camera could be a useful feature in a comparative opinion review. In other cases, the approach has been used to extract opinion targets and handle other natural language nuances particularly on social media platforms because new expressions and terms often emerge on these platforms (Wang, Lu, & Zhang, 2020)

2.7.1.4 Hybrid Learning

Hybrid learning approach involves the combination of various learning techniques for purposes of leveraging the strengths while overcoming the limitations of the individual learning techniques. The application of hybrid learning in opinion mining would involve a combination of two or all of these learning techniques: unsupervised, supervised, and reinforcement learning (Kharde & Sonawane, 2016). The primary goal of hybrid learning is to achieve advanced nuanced text data analysis that involves ambiguous and complex opinions. Popular hybrid learning methods include a combination of machine learning with lexicon-based models. In this case, the lexicon-based method provides a list of opinion words while the ML models like Random Forest perform text classification using the features from the lexicon model. This enhances opinion mining by overcoming the limitations of each approach. In this case, lexicon methods are limited in handling context-dependent opinions whereas ML methods demand huge labelled datasets (Sharma & Dey, 2012).

2.7.2 Rule Mining Approaches

A popular data mining method used in the discovery of interesting relations or associations between data is the association rule mining. In text mining, this approach has the objective of discovering strong patterns and rules between terms. Another method used in rule mining is called sequence pattern mining, which takes into consideration the order of terms in every sentence (Varathan et al., 2017). The definition of rule mining is expressed as X => Y, where variables X and Y are both subsets of items. Rules mining has also been applied to comparative sentences to identify the preferred entity (Ganapathibhotla et al., n.d.). In their study, compared entities, comparative opinion words, compared entity features, and type of comparison

were treated as features. One-side association measures were employed to generate the association between a feature and a comparative word. Class Sequential Rules (CSR), which is one of the types of sequential pattern mining was used to reveal common patterns in comparative sentences in a study by (Jindal et al., 2006).

Sequence patterns derived from comparative key phrases and POS tags were used as opinion features. Based on the manually selected key phrases, each pattern's opinion was determined. Their experiment proved that combining manual rules with CSR showed greater effectiveness in comparative sentence identification. This approach was used later to identify comparative sentences in a study by (Liu et al., 2013) except that in their study, syntactic patterns, adverbs, and comparative words were used as features. In addition, they combined machine-learning techniques with sequential rules mining in this CSR-based approach. The output was a classification of comparative sentences into three classes based on sentence polarity.

Liu et al. (2013) attained better classification results by incorporating CSR features in their model to identify comparative sentences in Chinese language. In a different study, rues that indicated comparisons were to detect comparative sentences by incorporating comparison words and sentence pattern rules (Gu & Yoo, 2009). Kurashima et al. (2008) employed language patterns in the detection of entities and the extraction of comparative relations. In their work, manually generated language patterns were used and the experiments were run on movie reviews with the results showing effectiveness in the extraction of comparative relations.

2.7.3 NLP Approaches

Natural language processing (NLP) is one of the sub-fields of computer science. Its role is in the development of methods for processing and analyzing language (spoken or written). NLP operates at both semantic analysis levels and syntactic analysis levels.

Syntactic Analysis

For the identification of syntactic relations existing between words, dependency parsing (a type of syntactic analysis) is used. A syntax tree is then output, which reflects a sentence's syntactic structure. Dependency parsing is useful in the generation of dependency grammar graph, which is then used in the identification and assignment of grammatical roles to words in a text. The identification of grammatical roles is significant in performing COM as this can help in recognizing the direction of a comparative relation and/or in the mining of feature opinion pairs. Both of these benefits are important in the development of recommender systems. Such systems benefit from the creation evolution trees that help in the visualization of how a product has evolved (Sun et al., 2009). A study by Liu et al. (2013), aimed at the identification of similar syntactic patterns in comparative sentences.

Semantic Analysis

One of the challenges in NLP is how to handle language ambiguities. Semantic analysis attempts to address this challenge by means of extracting the meaning of each given sentence. A popular semantic analysis method that is used in detecting semantic roles and how they related with predicates or verbs in a sentence is known as Semantic Role Labelling (SRL). In tasks like text summarization, assigning semantic roles is of great value. Hou and Li (2008) used SRL in the extraction of comparative relations from Chinese-based text. Kessler and Kuhn (2013) built a semantic-based model that

employed SRL for the detection of predicates and entities in comparative sentences. They identified arguments and predicates. They employed logistic regression classifiers to classify the arguments as positive or negative. Applying this to blog posts, the model outperformed the baseline model by Kessler et al. (2010).

Semantic network analysis approach uses a network to reflect how various concepts are related. This is a good way of depicting how different entities relate with each other based on their attributes. WordNet is probably the popular semantic network that contains a huge English-based lexical resource. In WordNet, there are links between sets of synonyms that represent groups of words. The linking is achieved by means of lexical and semantic relations. For this reason, WordNet has found applications in comparative opinion mining. For example, it has been applied in identifying different synonyms of opinion features (Liu et al., 2005); detection of synonyms and the enhancement of comparative opinions words list (Jindal et al., 2006); and identify the relations between opinion words and features (Anuradha et al., 2023). Another useful application of semantic networks was in aspect extraction, which is an important tasks in comparative opinion mining (Chaturvedi et al., 2018). Through a graph-based method, semantic network analysis was utilized to present entities as nodes and edges as relations between any two entities, thereby modeling how a prospective customer behaves when searching for a preferred entity (Kurashima et al., 2008).

Kim and Zhai (2009), Li et al. (2011), and Fujimoto (2012) also carried out studies on the application of semantic analysis in comparative opinion mining. Most studies machine learning or rule based approaches to perform various tasks in comparative opinion mining. While pattern-based approaches are appropriate in carrying out COM, better results could be obtained by combining the approach with methods from other

approaches. For instance, NLP approaches and unsupervised ML approaches could be explored for application in this area considering that they are currently underexploited in this regard (Varathan et al., 2017).

2.7.4 Hybrid Approaches

2.7.4.1 Basics of Hybrid Models

Previously, hybrid approaches have involved a mixture of ML and lexicon approaches. In this traditional usage, a hybrid model would use a sentiment lexicon for features while the ML algorithm would classify the opinions. In later studies, hybrid techniques arising from a combination of ML techniques have been used to carry out direct opinion mining. An example of this is study is research by Al Amrani et al., (2018) that used a hybrid ML model consisting of SVM and RF for direct opinion mining.

In their study, the single models (SVM and RF) achieved accuracies of 81% and 82% correspondingly while the resultant hybrid model consisting of both machine-learning algorithms had an accuracy of 84% on an amazon.com based product reviews dataset. This study shows that hybrid machine-learning models could be viable in achieving improved performance in opinion classification. Unfortunately, there are few studies that have delved in creating hybrid techniques involving a lexicon-based technique and a machine learning technique, making this an area of interest for further research (Wankhade et al., 2022).

Conversely, as far as the researcher's knowledge is concerned, hybrid machine-learning models have seen limited exploitation in comparative opinion mining and especially in brand reputation monitoring. Whereas the traditional application of the concept of hybrid models referred to models created from fusing lexicon-based techniques with a

machine learning techniques, recent advancements in the development of opinion mining tools and solutions have witnessed the development of hybrid techniques that involve other combinations. This includes Multiple Machine-Learning Techniques; Multiple Deep-Learning Techniques; Machine Learning and Deep Learning Techniques; Machine Learning and Lexicon-based Techniques; Deep Learning and Lexicon-based Techniques; Lexicon, Machine Learning; and Deep Learning Techniques; and Machine Learning; Deep Learning, and Lexicon based techniques.

2.7.4.2 Benefits of Hybrid ML Models (Machine Learning + Deep-Learning Models)

i. **Improved Feature Extraction** – deep-learning algorithms like MLP or CNN have the ability to extract complex features from text content, by capturing global and local patterns automatically. The features then are fed into the top-level model (in this case, a machine learning model like RF) to additionally learn the intricate relationships between the features thereby ultimately improving the accuracy of opinion analysis (Kim, 2014). Machine learning models on the other hand, frequently demand manual feature engineering, which is domain-specific and inefficient (Bengio et al., 2013).

ii. **Efficient Utilization of Resources** – deep learning models often require significant amounts of computation resources such as powerful TPUs and GPUs for training as well as inferencing. These resources might not be required in some machine-learning algorithms (Chollet, 2017). ML models instead are more resource friendly and efficient in tasks like classification while deep learning models are better in feature extraction. Therefore, combining them leverages the strengths of DL and ML hence improved computational resource efficiency (Kotsiantis, 2007).

iii. **Handling of Varied Data Types** – deep learning models perform well in processing of unstructured text like online reviews and social media data like tweets (Zhang et al., 2021). Machine learning on the other hand is more effective in handling tabular or structured data. A hybrid of these two therefore leverages the strengths of these approaches in processing diverse sources of data for much more comprehensive understanding and processing of opinions (Yang et al., 2016).

iv. **Explainability and Interpretability** – deep learning models are harder to interpret than machine learning models. Therefore, using a machine-learning model as the top-level model makes it is easier to interpret and explain the decision-making process behind the predictions made by the model. This improves understanding and trust in the model for use in applications like comparative opinion mining (Caruana et al., 2015).

v. **Improved Generalization** – when training deep learning on limited data, the problem of overfitting is likely to happen. Using a ML model as a top-level model would help regularize the predictions using learned representations. Thus, a hybrid of deep learning and machine learning aids prevent model overfitting thereby improving the model's overall opinion classification accuracy on unseen, new data (Zhang et al., 2021).

vi. **Domain Adaptation and Transfer Learning** – deep learning models are created by pre-training them on large datasets for purposes of capturing general language nuances and patterns (Yosinski, Clune, Bengio, & Lipson, 2014). When a DL model is used as base model in a hybrid architecture and fine-tuned using domain-specific

data, the resulting hybrid model can attain higher performance even when applied to a small amount of domain-specific labelled data (Howard & Ruder, 2018).

2.8 Opinion Mining Approaches

This section discusses the various approaches used to perform opinion mining. This includes ML approaches, DL approaches, lexicon approaches, and hybrid approaches.

2.8.1 Machine-Learning- Based Approach

A ML based approach to opinion mining necessitates using classification algorithms to extract relevant features that help improve the accurate determination of opinion polarity. Processes utilizing this approach require the availability of annotated (labelled) corpuses (Turney, 2002). Largely, ML approaches are categorized into supervised ML and unsupervised ML. A few studies have demarcated a third class of ML based approaches. This class is known as semi-supervised learning and it combines a few elements of supervised ML learning and unsupervised ML methods (Ligthart et al., 2021). Machine learning encompasses the optimization of algorithm performance using sample training data (Chao, 2011). Computers learn from presented training data to understand the patterns, knowledge, and experience used in making accurate predictions with test (real) data. The output of a ML process is a ML model. Using machine learning presents the following benefits or advantages:

1) It helps in processing very intricate tasks that regular computer programs cannot handle. Such tasks could be accomplished by humans but they are often beyond the capability of people to perform efficiently and effectively (Rathor et al., 2018).

2) ML enables adaptivity, which is unmanageable with orthodox programming. Since tasks vary over time and through dissimilar users, ML presents program actions that allow for adaptation to those changes (Shalev-Shwartz & Ben-David, 2013).

3) ML does not depend on the programming language used. A machine-learning model can be developed using any specific programming language and its relevant programming tools that support ML, simplifying the implementation of ML models using preferred languages such as Python, Scala, and R (Huda, 2017).

4) ML provisions for automation. Computers, with negligible human intervention, do ML model training, testing, and running. When the performance of a ML model has been successfully found to be satisfactory, the model can be left to perform a task automatically, deprived of human involvement. This leads to minimal amount of work on the part of humans, depending on the outputs of the ML models.

In machine learning, it is essential to use at least two different data sets. These are the training set and testing set. The ML classifier uses the training set to establish distinctions among text features. The testing set contains data for approximating the classifier's performance.

2.8.1.1 Supervised ML Algorithms

These are the kinds of ML algorithms developed by training them on labeled data and later exposing them to unlabeled data to classify it. In essence, this algorithm models the relationships and existing dependencies between input features and the target prediction output. Consequently, algorithms in this class are used often to predict or classify data inputs. There are various supervised ML techniques employed in classifying text into positive or negative classes based on sentiments. Supervised learning approach involves the following classifiers: Decision Tree, Linear Classifiers, Rule-based Classifiers, and Probabilistic Classifiers. Linear classifiers contain the following opinion mining techniques: Support Vector Machines, Neural Networks, Naïve Bayes, Bayesian Networks, and Maximum Entropy (Saranya et al., 2016). ML algorithm performances are affected by the dataset and features used.

2.8.1.2 Unsupervised ML Algorithms

This category of algorithms involves training the ML algorithm on unlabeled data to learn the patterns and be able to use such patterns to predict the patterns in unknown data (Saberi & Saad, 2017). Unsupervised learning algorithms include K-means clustering, Principal Component Analysis (PCA), Association Rule Mining, and t-Distributed Stochastic Neighbor embedding (Iqbal et al., 2019; Alsaeedi & Khan, 2019)

2.8.1.3 Semi-supervised Machine Learning Algorithms

The semi-supervised class of algorithms blends a subset of features from both supervised and unsupervised ML techniques (Madhoushi et al., 2015). They work with a subset of labelled training set, while the other subset is unlabeled. This is useful where labeled data is missing and/or the cost of labeling all the data is high. Semi-supervised algorithms include Generative Models, Graph-Based, and Self Training Algorithms.

2.8.1.4 Reinforcement Machine Learning Algorithms

In the case of this class of algorithms, a built algorithm uses its experiences and observations from interacting with its environment to perform appropriate actions that aim at minimizing the risk and/or maximizing the reward. The commonest algorithms in this class are Q-Learning, Asynchronous Actor-Critic Agents (A3C), Monte-Carlo Tree Search (MCTS), and Temporal Difference (Cumming, 2015).

2.8.2 Deep-Learning- Based Approach

Deep learning is a sub-field of machine learning, which typically utilizes deep neural networks. Recent times have seen rapid interest in and the adoption of DL in opinion mining. Some of the popular DL algorithms have been reviewed in a study by Ligthart et al. (2021) include Convolutional Neural Network (CNN), Multilayer Perceptron (MLP), Deep Neural Network (DNN), Recurrent Neural Network (RNN). There are varying implementations or variations of these algorithms, including the Long-Short Term Memory (LSTM).

2.8.3 Lexicon-based Approach

This approach is linguistically driven. The process involved entails analyzing text to identify valuable features that indicate semantic polarity (Arora et al., 2012). Lexicon-based approach depends on using a predefined lexicon (dictionary) of words, known as a sentiment lexicon. The lexicon contains phrases or words with the corresponding sentiment scores that show the positivity, neutrality or negativity of a word. This method has the main goal of determining the general opinion of text with the help of sentiment scores / values associated with each word in the text. However, the methodology has the limitation of not being capable of properly handling language nuances like negations, which depend on context (Cambria et al., 2013).

2.8.4 Hybrid Approach

A hybrid approach traditionally combines a few elements of ML and lexicon-based approaches, or, according to recent studies, any two of the four approaches described in the previous sections. This approach consists of two sup-approaches: Dictionary-based approach as well as Corpus-based approach. The corpus-based approach similarly has twofold opinion mining techniques: semantic and statistical techniques (Saranya et al., 2016). Studies have exploited hybrid models of regular opinion mining using hybrids made by combining ML algorithms, or ML with DL algorithms, or ML with lexical approach, or DL with lexical approach, or ML with both DL and lexical approach. Section 2.11 presents some of the hybrid models created this way.

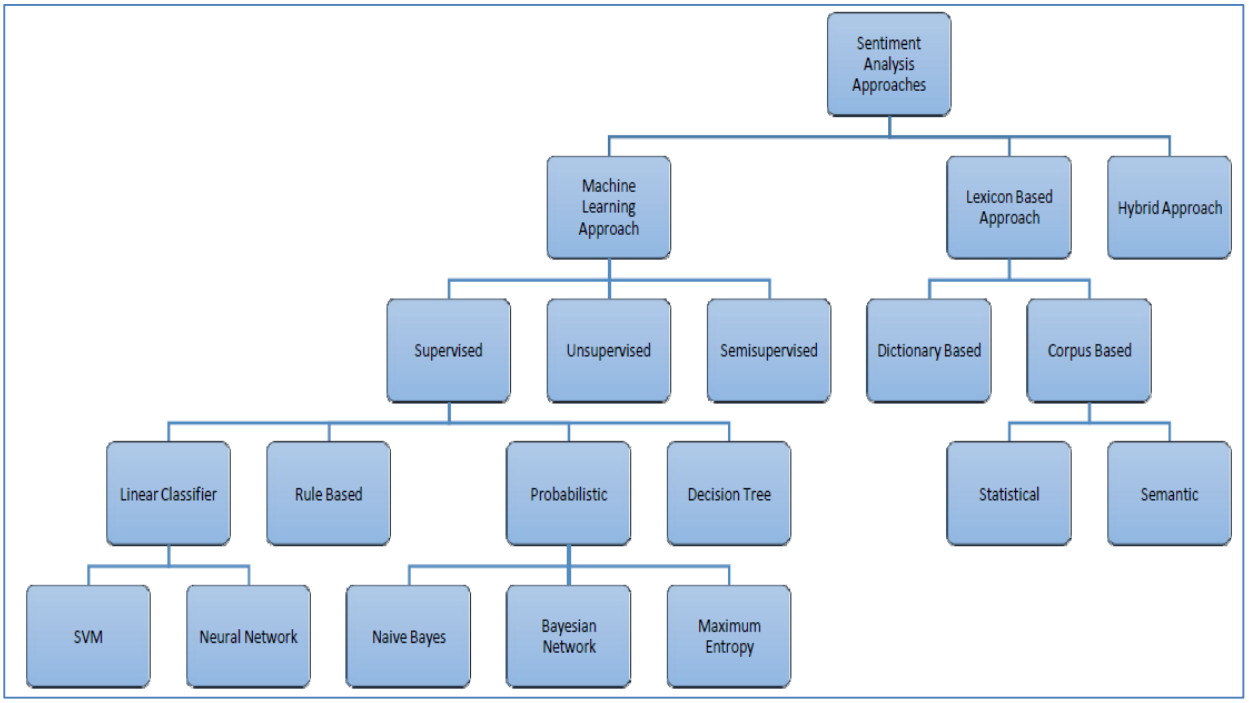

Figure 2. 3 Opinion Mining (Sentiment Analysis) Approaches & Techniques

(Source: Chauhan & Singh, 2017)

2.9 Machine Learning Algorithms

This section describes some of the state-of-the-art ML algorithms used in opinion mining.

2.9.1 Support Vector Machine (SVM)

This algorithm uses supervised ML to handle opinion classification and is widely used (Schrauwen, 2010). It follows a maximum margin classification approach that tries to gain an understanding of how all data points available are represented to permit unrelated opinion labels to be separated by a clear, adequately large gap (Doni Abdul Fatah et al., 2023). In quite a few classification and recognition problems, SVM is used as a baseline for comparing the performance of algorithms. Its primarily handles quadratic programming problems in which computational time is subject to the number of features are modeled. Thus, where huge data is present, feature transformation takes more time during model training. Generally, SVM has no intuitive parameterization, requires a small amount of memory, has an average tendency to model overfitting, takes more time to train, but is efficient in generating outputs (Chao, 2011).

SVM is prevalent in the opinion mining as a consequence of its greater performance (Ahmad et al., 2018). Several researchers consider it as the best algorithm in text classification tasks such as opinion mining (Yu & Kak, 2012). The algorithm works by examining information, unfolding optimal parameters, and performing calculations (depending on the components) in the input space. Significant information is obtainable using twofold vectors whereby each vector takes a size *m*. Each vector is a class. Subsequently, SVM isolates the borderline unraveling the classes. This limit has to be distant from each other in the training dataset samples. This boundary forms the text

classification edge. Widening of this particular edge causes abridges ambivalent choices. Consistent with Khairnar and Kinikar (2013), the SVM model is more effective in terms of performance than the Naïve Bayes model in different problems in opinion classification. Chauhan and Singh (2017) correspondingly specify that SVM algorithm has a premier accuracy in comparison with other text classifiers. This result is comparable to that by Borele and Borikar (2016). However, this finding is not consistent with some studies that have revealed that other ML models like Random Forest actually outperform SVM in text classification (Ondara et al., 2023).

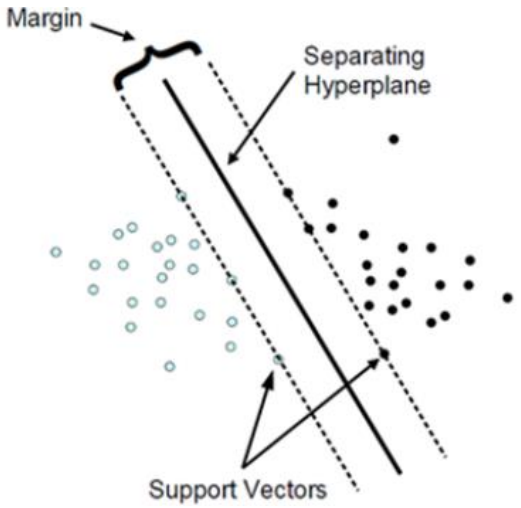

Figure 2. 4 Operation of the Support Vector Machine (

(Source: Dey, 2016)

2.9.2 Naïve Bayes (NB)

This algorithm is prevalently used in the text document classification as well as opinion mining. It is established on a probabilistic model utilizing the accommodating probabilities or likelihoods of particular terms and their related classes to estimate the prospect of a definite group being in possession of a text document as probable input. NB has no parameterization involved, uses small amounts of memory, has low tendencies to overfitting, takes little time to learn and its prediction latency is low, being a fast algorithm (Schrauwen, 2010). It is a very powerful algorithm. The NB presumes

that the possibility of every single word in a text document does not hinge on its context and position in the real document. Hence, every document is treated as word count vector. Thus, the likelihood of a text document's category or class is handled as a multinomial distribution. Assuming x is the word count, the multinomial distribution is as presented in equation X (Ankit & Saleena, 2018).

$$p(\boldsymbol{x}|c) = p(x_+|c)\, p(\boldsymbol{x}|c, x_+)$$
$$= p(x_+|c)\, \frac{x_+!}{\prod_d x_d!} \prod_d p(d|c, x_+)^{x_d}$$

2.9.3 K-Nearest Neighbors (KNN)

This algorithm utilizes a test tuple that is associated with several equivalent training tuples, where each tuple characterizes a certain point within n-dimensional space. The value of *n* represents the tuples' features. When KNN is given an unidentified text, it probes the space of patterns to discover the *k* training tuples adjacent to an unknown tuple. The *k* training tuples are, in this case, the k-nearest neighbors to the unknown tuple. To delineate the closeness of the tuples, the Euclidean distance (ED) is used to measure the distance (Mukwazvure & Supreethi, 2015). Subject to the distance of the neighbor from the query point, a weight value is apportioned. KNN algorithm is effective when used in classifying short, small sentences (Kaur & Kumari, 2016).

2.9.4 Logistic Regression (LR)

This algorithm performs opinion classification by depending on both training set and testing set to establish opinion polarity classes. It is a fast algorithm when implemented as a text classification model, achieving enhanced model generalization, which is essential in managing or overcoming overfitting. This makes models based on LR to perform well on unknown data. The algorithm is presented by the formula (Younis et

al., 2020). The algorithm has simple parameterization, uses small memory sizes, has low overfitting tendencies, is a weak learner, and takes a little longer to predict the outcomes (Harfoushi et al., 2018).

$$e\_0+\_1x1+\_2x2$$

2.9.5 Decision Tree (DT)

This is a common supervised ML algorithm in many AI problems such as text classification. A DT consists of internal nodes comprising of labeled features. The boundaries from the tree's nodes have labels or annotations on tests performed on the weights of features, in conjunction with the leaves labeled by classes or categories. DT algorithm has efficaciously been used in many NLP applications (Schrauwen, 2010). Their non-linear functions are suitable in tasks that necessitate inductive extrapolations. DTs are k-array trees consisting of decision nodes that embodies decisions based on unequivocal features found in the particular input data; and the leaf node, which is allied to exact feature values well defined at the decision node. In view of that, if a decision node has only positive instances, it characterizes a positive class. On the contrary, if it comprises only negative instances then it signifies a negative class.

This algorithm categorizes documents by means of a top-down method –root through branches down to the leaf nodes. Hereafter, each leaf node predicts a class of data (Chauhan & Singh, 2017). The algorithm begins with training tuples from a specified dataset, selecting what it regards as the top distinctive, which adds to a test node. From the test node, a top-down technique is followed in building a DT from available test nodes using test quality values (Suresh & Bharathi, 2016). Generally, decision tree algorithm has intuitive parameterization, requires large memory sizes to work, have

very high tendencies to overfitting, require little time to learning / training and takes little time to predict the outcomes.

2.9.6 Random Forest (RF)

The RF model is a versatile and powerful tool in opinion mining because of its many benefits. These benefits include its capability to handle intricate feature interactions in textual data; demonstrated efficiency in handling imbalanced datasets, which is a common occurrence in comparative opinion mining applications; not counting its ability to handle high dimensional feature spaces, which is critical in comparative opinion mining (Fernandez-Delgado, Cernadas, Barro, 2014). Random Forest is a self-determining algorithm that exploits several decision trees. It applies the principle of joining numerous DTs (for their extrapolative significance) and the popular vote technique to build a final class that is made up of the bulk of votes. This process uses bagging as well as bootstrapping as key concept, thereby compensating for overfitting to a particular training dataset and as a result, plummeting the problem of DT overfitting with their increasing depth (Khanvilkar & Vora, 2019). This grants a crucial value of the RF model (Reel et al., 2019).

The high accuracy of RF is predominantly ostensible on datasets with low dimensionality (Novalita et al., 2019). Regularly, this algorithm achieves pronounced results. With simple variations on its hyper-parameters and strict selection of features, RFs produce remarkable results (Khanvilkar & Vora, 2019). Generally, RFs have intuitive or simple parameterization (i.e., the number of decision trees), requires very large memory capacities to train and run, has average tendency to model overfitting, exhibits costly time for learning, and takes a little longer to predict the outcomes.

2.9.7 Stochastic Gradient Descent (SGD)

SGD is a powerful optimization method, which outperforms many machine-learning models in opinion mining as well as in others classification tasks. Its superior performance is attributed to several factors (Ruder, 2016). First, SGD is highly scalable and efficient, which makes it a good choice for working with huge datasets as is often the case in opinion mining. Due to its stochastic nature, SGD uses one data point every single time to update its model parameters, which leads to a decrease in computational complexity as well as an improved convergence. Second, SGD is capable of handling high-dimensional feature spaces. In comparative opinion mining, feature spaces can be vast due to the huge vocabulary of words.

SGD effectively traverses these high-dimensional spaces leading to a more rapid convergence to attaining an optimal solution. Third, through the customization of SGD's loss functions as determined by opinion (sentiment) classification objective to achieve optimized model performance in the performance of opinion mining. Fourth, opinion patterns and data distribution may exhibit nonlinearities. SGD uses its frequent updates of its model parameters to achieve a generally optimal solution. Finally, SGD has inherent randomness, which makes the model more generalizable and able to better address model overfitting in handling sentiment patterns and classification over different datasets (Mikolov, Chen, Corrado, & Dean, 2013).

2.9.8 Popular Existing Machine Learning Models for COM

Several machine learning techniques were identified through a systematic literature review. The review provided useful information on opinion mining including state-of-the-art machine learning techniques, feature extraction techniques, text vectorization techniques, dataset sources, and algorithm performance metrics (Ondara et al., 2022).

Table 4.2 shows the popularity ranking of the state-of-the-art machine learning algorithms used in opinion mining from the reviewed works. Besides being popular in most studies, these algorithms have high accuracy, robustness, and achieve high model generalization with easy- to-interpret results in opinion mining tasks (Yang et al., 2016). This may explain their wide application among researchers and software developers.

The first eight algorithms were ranked from the reviewed literature. The last two were identified in other literature but not evaluated in the systematic literature review. The Stochastic Gradient Descent (SGD) was included because it is an emerging, powerful technique in optimization of tasks like opinion mining where it has proven to outperform many machine learning techniques (Bottou et al., 2018). The multilayer perceptron (MLP) was included in this study because it has been tried in classification tasks as a deep learning model to leverage its power to model very complex relationships for obtained good features for different opinion scenarios thus improving opinion mining across different entities (LeCun et al., 2015). From the study, although there are attempts to use other DL techniques, the MLP is still the most popular.

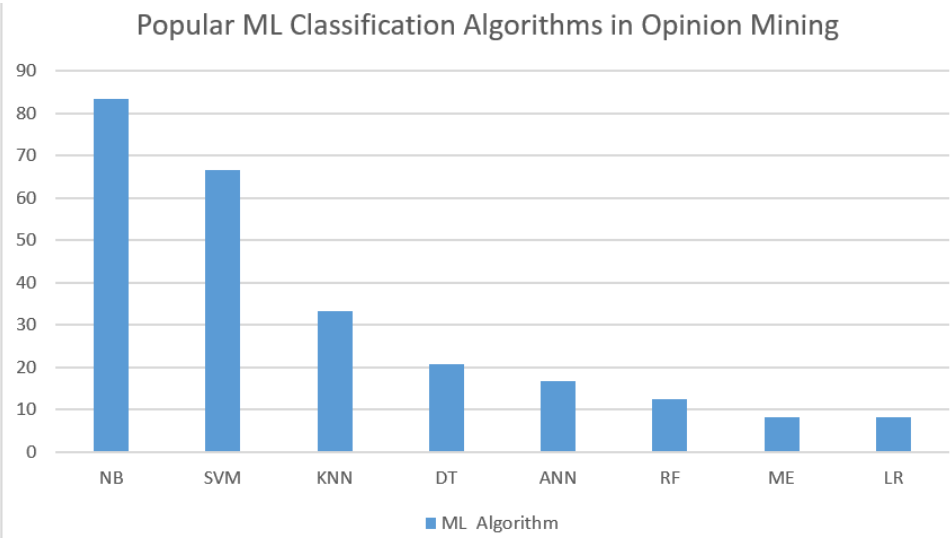

Figure 2. 5 Popular ML Classification Algorithms in Opinion Mining

(Key: The vertical axis represents percentage accuracy measure)

From *Figure 2.5*, it is evident that the Naïve Bayes algorithm is the most popular in opinion mining research. This is according to a systematic review conducted by Ondara et al. (Ondara et al., 2022) that revealed the popularity of the eight machine learning algorithms in opinion mining. The logistic regression and Maximum Entropy models are the least popular. The Artificial Neural Network (ANN) is the fifth most popular algorithm. This is also the only deep learning algorithm identified in the same study. The following is a list of single machine learning algorithms / models that have been applied to comparative opinion mining in a recent study (Younis et al., 2020).

i. Naïve Bayes
ii. Support Vector Machine
iii. K-Nearest Neighbors
iv. Decision Tree
v. Random Forest
vi. Gradient Boosting.

Beside these eight popular ML algorithms in classification, recent research papers showed these two algorithms have been used in opinion mining.

i. Multilayer perceptron (Kim, 2014) and
ii. Stochastic Gradient Descent (Ruder, 2016).

Both algorithms are becoming of great interest among experts in opinion mining. For this reason, these two algorithms were also experimented with in a bid to determine their suitability for use in the hybrid ML mode. Algorithms common to Direct Opinion Mining and COM were:

i. Naïve Bayes
ii. Logistic Regression
iii. Support Vector Machine
iv. K-Nearest Neighbor
v. Decision Tree
vi. Random Forest

From the above breakdown, it is evident that DL algorithms have not been widely used in most recent studies in opinion mining. It is similarly evident that for comparative opinion mining, only single / lone / independent algorithms have been used.

2.10 Deep Learning (DL) Algorithms

DL algorithms are efficient in opinion mining owing to their capability to learn robotically. They utilize multiple Neural Networks (Gulli & Pal, 2017), which is demonstrated in the contemporary uses of DL models, for instance, Multilayer Perceptron (MLP), Convoluted Neural Networks (CNN), and Recurrent Neural Networks (RNN) in opinion classification. The approach resolves various opinion mining issues facing popular ML algorithms for instance KNN models, and models based on rule mining. Particularly, DL algorithms can learn extensive features and entity relations from text, which improves overall classification and analysis of opinions in comparative opinion texts.

2.10.1 Multilayer Perceptron (MLP)

MLP is a powerful DL algorithm that through its gradient-based optimization strengths and the utilization of multiple hidden layers that possess hidden activations excels at learning the intricate non-linear relationships that exist in data (Bengio et al., 2013).

Among the various DL algorithms, MLP is easier to integrate with other machine learning models such as RF, DT, and SVM. Generally, MLPs have one hidden layer, at least one hidden layer, and one output layer. The number of hidden layers in the architecture has paramount effects on the performance of MLP on comparative opinion mining related tasks. Every hidden layer consists of nodes (neurons) which process the received input before feeding it into the next layers via weighted connections.

The hidden layers are essential in capturing intricate representations and patterns from the input. This allows MLPs to both learn and consequently model non-linear relationships (Goodfellow, Bengio, & Courville, 2016). To capture the nuances in comparative opinion mining more accurately, deep MLPs are necessary as they can learn hierarchical representations, which enables them to capture more intricate and abstract patterns in data. Thus, a higher number of hidden layers in MLP architecture translates to a higher power to better capture these hierarchical representations at the risk of model overfitting mainly for small datasets (Bengio et al., 2013). Generally, MLP has no intuitive parameterization given its many layers, it requires an intermediate amount of memory to work, has average model overfitting, takes longer to train but less time to predict outputs (i.e. make predictions).

2.10.2 Convolutional Neural Networks (CNNs)

CNN is a high performance classification model predominantly when applied to image classification and computer vision projects. It mimics how the visual cortex of a person's brain functions to recognize and classify objects. It can be made of convolutional layers, ReLU layers, and loss layers. Its other layers include pooling layer in addition to fully connected layers. A distinctive CNN design has numerous layers. Besides a convolutional layer, a Rectified Linear Unit, a Pooling layer, it must have as

a minimum one convolutional layer, and a fully connected layer (Gulli & Pal, 2017). A sole property of CNN is that it studies image structures when handling image processing or classification. On the other hand, basic neural networks normally convert their inputs into a 1-Demensional array that causes the trained classifier to endure a smaller amount of sensitivity to positional changes.

One of the greatest results from research using the MNIST dataset involved using multi-column based deep neural networks. In this, multiple maps in every single layer were exploited and every layer had more than a few layers consisting of nonlinear neurons (Gulli & Pal, 2017). The architecture typifying a CNN allows operational and resourceful processing even with the intricacy of CNNs that make them much more demanding to train even when using special code and graphical processors like GPUs. The network works because of the winner-take-all kinds of neurons depending on maximum pooling to limit the specific neurons win. In a dissimilar research, CNNs produced improved accuracy in computer vision.

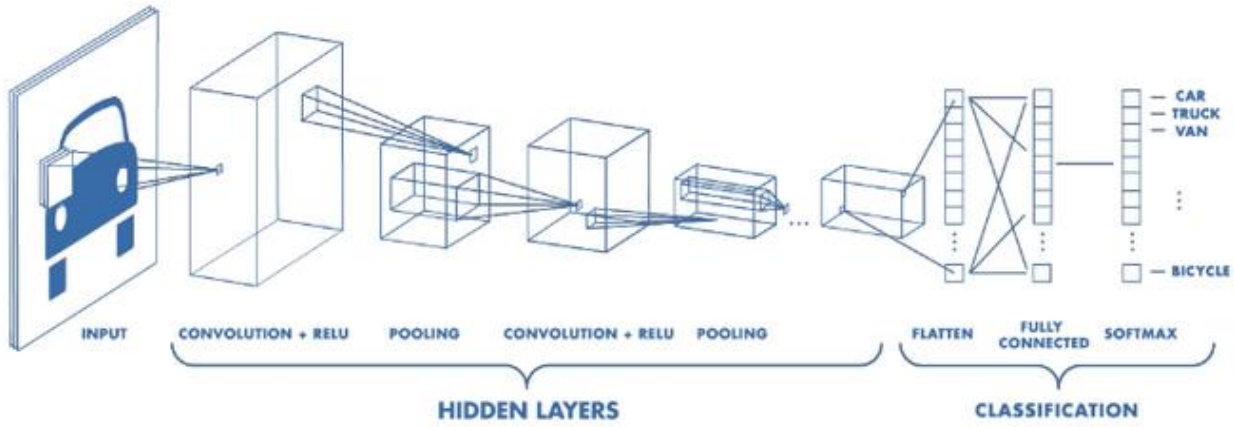

Figure 2. 6 Typical CNN Architecture

(Source: Chatterjee, 2019)

2.10.3 Recurrent Neural Networks (RNNs)

RNN is an improved CNN by utilizing an architecture that contrasts that of a CNN. In its design, analogous weights are repeatedly relating to specific data. Therefore, the common uses of RNN include handwriting recognition and speech recognition. RNNs have been tried in predicting the following word in a given sentence (Gulli & Pal, 2017) . However, the use of RNN layers alongside a couple max pool layers amid the layers and having at the culmination a global max pool layer grants a number of advantages.

To begin with, in each RNN layer, every unit factor contains in a somewhat progressively bigger way around it. A supplementary advantage is that RNN layers upsurge the network's complexity devoid of necessitating the application of additional parameters. This DL algorithm has proven effective in language translation and opinion mining among other uses. Their design permits them to remember historical information during processing (Malik & Kumar, 2018).

One of the RNN architecture is a network of neuron-like nodes, arranged in consecutive layers. The other RNN architecture is known as the Elman network architecture. In this architecture, there are three layers structured horizontally, which comprises a few context units.

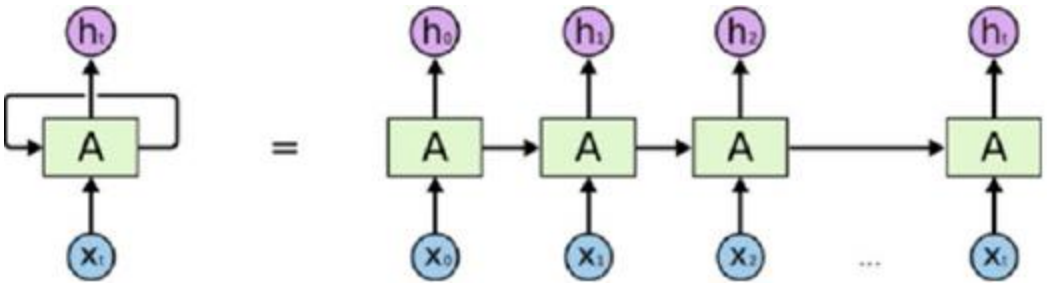

Figure 2. 7 An Unrolled Recurrent Neural Network

2.10.4 Transformer Models

Transformer Models have demonstrated great performance in NLP in the recent past. The models were introduced by Vaswani et al. (2017), showing remarkable success in handling contextual information, in addition to capturing long-term associations in text. Some of the most powerful transformer models according to Lu et al. (2022) include the Bidirectional Encoder Representations, which is known as BERT for short, and Generative Pre-trained Transformers, which his known as GPT for short. Both models have been applied in extracting opinion nuances and entity relationships.

A study by Devlin et al., (2019) utilized BERT model for comparative sentiment analysis, attaining state-of-the-art performance results in gaining insight into opinion and sentiment nuances in comparative texts. Transformer models have high performances because of their ability to model dependencies and context much more effectively, which enables them to advance a profounder understanding of the nuances in comparative opinion language. In this regard, they use attention mechanism to take into consideration every word in a sentence when performing predictions of opinion classes, thereby capturing complex relationships between entities on large corpora.

However, transformer models have their weaknesses too. They are computationally expensive, demanding substantial resources for their training. In addition, because of their huge sizes, the models are often difficult to deploy in resource-constrained computing environments (Devlin et al., 2019), which limits their adoption. In addition, it is laborious to fine-tune transformer models to make them perform particular tasks. This reason also has made their adoption slow. Token limitations is a major drawback of transformers in comparative opinion mining applications. This problem arises from the challenge of transformers struggling to process very lengthy sequences that are

often inherent in lengthy comparative opinion texts. Transformer performance degrades with a rising number of tokens in a comparative opinion (Tan et al., 2023). Lastly, for transformers to work effectively in different domains, the process is resource-intensive, requiring huge amounts of pre-training data and model fine-tuning.

2.10.5 Benefits of Deep Learning Models

There are many benefits for using DL models. First, deep learning models have the capability to learn important and relevant features from available raw data (Bengio et al., 2013). This makes the models to represent complex patterns better. Second, deep learning models handle larger datasets more effectively because of their hierarchical architectures and utilization of parallel processing (LeCun et al., 2015). This makes them more scalable hence better in handling big data based projects. Third, deep learning models have attained modern performance in performing tasks related to using unstructured data like speech recognition and image processing (Lu et al., 2022). Fourth, DL models have the ability to leverage pre-trained representations on high volume datasets and knowledge transfer to new tasks (Yosinski, Clune, Bengio, & Lipson, 2014). This has the potential to enhance performance when such models are working with small datasets. Finally, deep learning models have greater capabilities to capture complex relationships found in data for purposes of achieving improved model generalization and complexity (Zhang et al., 2021).

2.11 Existing Hybrid Machine Learning (ML) Models

The following hybrid ML models have been experimented with in areas other than COM. This study aimed at filling this research and application gap.

i. RF and LSTM: A study was carried to perform sentence classification using a hybrid model consisting of RF and LSTM. In this model, RF was used to extract features that LSTM used for classification. This hybrid model leveraged the strengths of both LSTM (a deep learning model) and RF - an ensemble of Decision Trees, which are machine-learning model (Karijadi & Chou, 2022). Tis model was applied to energy consumption prediction.

ii. CNN and LSTM: Using Qualitative User-Generated Contents, built a hybrid model consisting two deep learning models (CNN and LSTM) and applied it to consumer opinion mining (Jain et al., 2021). This hybrid model was applied to a single entity; it was only used to perform direct opinion mining.

iii. CNN and RF: A Convolutional Neural Network (CNN) was hybridized with Random Forest (RF) and applied to sentence classification (Kim, 2014). In sentence classification, this hybrid model was able to output the class that a sentence belonged. The two classes were comparative sentence and non-comparative sentence. Thus, the model was not used in comparative opinion mining.

iv. BERT + RF: this is a hybrid model created from combining Bidirectional Encoder Representation Transformer (BERT) with Random Forest. In this model, BERT generates word embedding as features for the RFt to train with an eventual classification of sentences. This model was used to perform direct opinion mining.

v. DMN and SVM: this hybrid model combines Deep Memory Network (DMN) for selection of features and SVM for classification (Yang et al., 2016).

vi. ANN and DT: This model consisted of Artificial Neural Network model for prediction and the Decision Tree model for classification for building energy consumption (Banihashemi et al., 2017).

2.12 Criteria for Selecting ML/DL Algorithms for Comparative Opinion Mining

The key to selecting a good combination of ML algorithms for inclusion in a hybrid model lies in the strengths and weaknesses of different single machine learning models. This was aimed at leveraging the strengths of the combined algorithms while addressing the challenges or limitations of the independent ML & DL algorithms. This section addresses common criteria for selecting algorithms for developing a hybrid ML model.

i. **Model Type Diversity**: A critical criterion in deciding on which machine learning algorithm to use is to combine deep learning algorithm (e.g. MLP or CNN) with a machine learning algorithm (e.g. RF or SGD) to aid leveraging the unique strengths of these two unique model types in a bid to boost the performance of the hybrid model (Sagi & Rokach, 2018). This criterion was adopted in this study especially in the development of four hybrid models with the MLP algorithm as the base model and ML algorithms as top-level models. In four other cases, the SGD machine learning algorithm was used as the based model with MLP as the top-level model. In both cases, leveraging the strengths of the ML and DL was an influencing factor.

ii. **Performance and Generalization**: Algorithms selected for developing a hybrid model should have reliable generalizations and performance in opinion mining tasks. Evaluating the performance of each model on benchmark datasets is required to help ascertain that the resulting hybrid would have accurate opinion classification and opinion class comparisons (Yang et al., 2016). In this study, different ML and DL algorithms were tested on the similar datasets under similar computing environments to ascertain the comparative performance. This allowed for selecting higher performing algorithms for use in building the final hybrid ML models.

iii. **Complementary Features and Feature Representations**: The algorithms selected to form a hybrid model should be able to capture features that are complementary and representative of the data. For instance, a deep learning model like MLP could be used to capture contextualized word embedding while ML model like RF may well be used to represent sparse data features in a more effective way (Devlin et al., 2019). This was a key consideration in this study, leading to the development of hybrids like MLP and RF.

iv. **Ensemble Methods**: It would be necessary to choose the method of ensemble learning for developing a hybrid model. Ensemble methods like boosting, bagging and voting can help identify the strengths of each independent model to help determine the overall performance improvement once the independent models have been combined (Breiman, 2001). The RF algorithm used in this study uses ensemble learning. It was used in the creation of a hybrid model with other algorithms like MLP using the same ensemble learning method.

v. **Adaptability and Flexibility**: In COM, data is from multiple sources and exists in different formats and structures with varying linguistic styles. Therefore, the selected machine learning techniques should be adaptable and flexible in handling such data types (Ruder et al., 2019). This study utilized data from social media platforms and consumer review websites, which are differently formatted.

vi. **Model Scalability and Complexity**: Hybrid models should exhibit computational resource manageability and efficiency particularly on real time applications or when dealing with large datasets. Consequently, algorithms that are less resource intensive and more scalable form a good choice in making a hybrid model (Bengio et al., 2013). In this study, the combination of DL algorithms with ML algorithms was aimed at reducing computational resource requirements attributed to using only DL algorithms in creating hybrid models. ML algorithms are less resource-intensive, reducing the resource requirements on a model based on DL and ML hybrid architecture.

vii. **Model interpretability**: To understand the predictions of the model in as far as COM is concerned; the algorithms used should be those with interpretable results. For ML models, techniques like DTs are easy to interpret. In deep learning, techniques like attention mechanisms are helpful in interpreting the model's results. Therefore, combing models whose results are easier to interpret would result in a model that is easy to interpret. The interpretation of the algorithms or models used in this study was simplified through the use of a standard presentation format: classification reports.

2.13 Ensemble Learning Method for Developing Hybrid ML Models

ML and DL are powerful tools for handling complex problems in diverse domains. A reliable effective method to enhancing the robustness in addition to performance of ML models is by means of ensemble learning. This section presents the process involved in the development of hybrid ML models based on the ensemble learning method for purposes of achieving model generalization and improved model predictive accuracy for satisfactory results in comparative opinion mining (Younis et al., 2020). This is because hybrid models leverage the strengths of the individual base learners to provide a more precise and robust model. The process for building a hybrid ML model using ensemble method may follow the steps below.

i. Literature Review: An extensive analysis of literature is needed to develop proper understanding of how ensemble learning works and its utilization in developing hybrid ML models. Bagging, stacking, and boosting are the three main techniques used in ensemble learning method. These three have had wide application in past studies with demonstrated results in developing powerful hybrid models using multiple base learners (Bergstra & Bengio, 2012). Many studies have demonstrated that hybrid ML models made using ensemble learning method outperform their respective single learners in handling prediction tasks (Sagi & Rokach, 2018).

ii. Data Preprocessing: Technically, this is the foremost step in developing a ML model. A diverse dataset that aligns with the problem to be solved are collected, cleaned, and normalized. This gives way to perform feature engineering in a bid to prepare the data so that that it can be fed to a ML algorithm for training and subsequent testing. Data preprocessing makes certain that the data is consistent and devoid of anomalies as well as presented in a suitable format (Kotsiantis, 2007).

iii. Base Model Selection: In this step, several factors are considered in selecting a base learner for the hybrid model. It is critical that the base learner is good at feature extraction so that the different patterns in the data are identified with the highest possible accuracy. Some of the notable good algorithms that can serves as base learners include Random Forest, SVM, and Gradient Boosting algorithm, which is capable of handling gradient-based optimization issues (Hastie et al., 2009). However, where the dataset is huge, deep learning algorithms would be the best choice as they can do self-learning of the features hence improve feature extraction accuracy. A powerful choice in this case, for example, is the MLP algorithm.

iv. Ensemble Method: This step involves merging the outputs from the base model(s) as is the case with algorithms like Random Forest that consist of Decision Trees as the base learners, or making predictions made from the base learner and feeding them into a final model for classification or prediction purposes. The base predictions may take the form of probabilities or actual classification predictions but they are fed into the top-level model for final classification, which can be achieved through boosting, bagging or stacking (Wolpert, 1992).

v. Model Training and Evaluation: At this step, the dataset is split into two: training set and testing set. The first set, which is usually about 70% of the whole dataset (Younis et al., 2020), is used in training the base model and the meta-learner. Then, the predictions made by the base model are used as inputs to train the meta-learner or final estimator. To evaluate the hybrid model, different metrics like accuracy and f1-score could be used to assess the model's performance.

vi. Hyper parameter Tuning: Since the hybrid model may not at first produce the desired performance, it would be necessary in that case, to tune its hyper parameters with the intent of improving its performance. Cross-validation and grid search technique are some of the popular and powerful techniques for determining the most optimal hyper parameters that could be combined to ensure enhanced model performance (Bergstra & Bengio, 2012).

vii. Model Deployment and Monitoring: Once the performance of the hybrid ML model has been confirmed as satisfactory, it may be the right time to deploy the model so that it can then be utilized to make predictions on unseen, unknown data. This step requires monitoring how the model performs on the new data for purposes of improving its performance in case it is lower than found during model testing. This step also involves maintaining the model once its performance has been found to be stable over time (Manouselis et al., 2014).

2.14 Feature Extraction Techniques in Opinion Mining

In feature engineering, there are different techniques for extracting features for model training. Each of the techniques has its strengths and weaknesses. This section presents a brief description of some of the more popular feature extraction techniques.

a) Feature Vector - this technique involves converting a given text into a particular matrix made up of token counts.

b) Term Frequency - this feature engineering technique involves counting the total frequency of each particular term in text (e.g. a review or tweet).

i. **Count Vectors**: The Count Vectorizer (CV) technique is a trivial representation of features that works by converting textual documents into a matrix containing counts of each token. In this technique, every row in the matrix represents exactly one single document while each row represents only one unique word from the whole corpus. The cells contain values of that correspond to the number of occurrences of a word in a respective document. This technique is considered basic but powerful in the conversion of text data into numeric values that are fed into a ML technique or algorithm as input (Joachims, 1998).

ii. **TF-IDF**: TF-IDF computes the importance of each word in a given document proportionate to a corpus of many documents. Its first part is the Term Frequency (TF) that measures the frequency of a term or word in a document. The second part is the IDF, which measures the uniqueness of a term or word in the corpus. Therefore, TFIDF gives additional weight to important words in a particular document while assigning less weight to words that are less common (Salton et al., 1975). Hence, TFIDF is considered more advanced than the CV technique. However, this advancement may not be evident in some applications.

iii. **CBOW**: The Continuous Bag of Words (CBOW for short), is one of the word embedding techniques that finds its use in areas like natural language processing. In its use, a model is specifically trained to make predictions of a target word based on the context of words surrounding that word. For applications where it is critical to extract the semantics of a word, this technique proves very useful. However, it requires dense vectors to perform well (Mikolov et al., 2013).

iv. **Skip Gram**: Skip Gram is also one of the feature extraction techniques under word embedding technique but working in the opposite way to the CBOW technique. For Skip Gram, the focus is to predict context words with respect to some target word. In applications where it is important to capture word semantics and relationships, this technique may find usefulness (Mikolov et al., 2013).

v. **Word2Vec**: This is an extension of the Skip Gram and CBOW techniques. It works by creating dense vectors that are then used to characterize words by bagging the words' semantic meaning (Mikolov et al., 2013).

vi. **Glove**: The Global Vectors for Word Representation is a feature extraction technique that combines the occurrence statistics of global words to help in learning word vectors (Lu et al., 2022).

vii. **BERT**: The Bidirectional Encoder Representations from Transformers (BERT) is a more powerful model for feature extraction with use cases in NLP. It is a pre-trained model for capturing word relationships and context (Devlin et al., 2019).

2.15 Model Performance Evaluation Metrics

There are numerous ways of finding the metrics for evaluating opinion classifier performance, which help in understanding the accuracy of an opinion-mining model.

2.15.1 Cross-validation

Feasibly, this is one of the most extensively used metric in model performance validation is cross-validation. Cross validation splits training data into several folds. In many cases, 75% of the data suits the training fold and the remaining 25% becomes

testing folds. This process is recapitulated several times after which an average for each metric is worked out. It overcomes overfitting because the number of folds to use depends on the amount of data available.

2.15.2 Accuracy

Accuracy measures how many texts from the whole corpus were predicted correctly, placing them in the correct class. In other words, accuracy is a measure of how frequently a technique correctly makes a positive prediction as a fraction of the total number of positive predictions made. This implies categorizing a few texts as having its place in a definite class unlike other texts. Eventually, at least two classes are involved in deciding accuracy. With imbalanced datasets, accuracy may not be suitable in measuring if a classifier is good or bad. Hence, precision and recall are applied (Sokolova & Lapalme, 2009; Malik & Kumar, 2018).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2.15.3 Precision

This metric is for defining the quantity of texts predicted appropriately as suitable in a certain category in relation to all predicted texts as suitable in the category. This refers to the number of texts that were predicted correctly divided by the total number of texts predicted as positive (Sokolova & Lapalme, 2009; Malik & Kumar, 2018).

$$Precision = \frac{TP}{TP + FP}$$

2.15.4 Recall

Recall is an algorithm performance evaluation metric that gauges the sum of texts predicted properly as fitting in a definite category relative to all texts belonging in that category. As the classifier gets more data, precision and recall grows.

$$Recall = \frac{TP}{TP + FN}$$

Comparative opinion mining is a classification problem since its goal is to classify an opinion, sentence, feature, entity, or relation correctly. In view of this, the recurrent measures are adapted from conventional classification field and they include accuracy, precision, recall, and F-score (Varathan et al., 2017). In each case, the output is a positive, negative, or neutral label (Sokolova & Lapalme, 2009; Malik & Kumar, 2018).

2.15.5 F-score

F-score is a combination of recall and precision hence more suitable in evaluating the performance of classification techniques. It relies on precision or recall alone is not enough (Sokolova & Lapalme, 2009). In the case of imbalanced datasets, F1-score, which is a harmonic average of both precision and recall generates a fair representation of both metrics (Powers, 2011; Malik & Kumar, 2018).

$$F\text{-}score = 2 \cdot \frac{precision\, recall}{precision + recall}$$

The above four measures are frequently used in evaluating the performance of classification techniques in comparative opinion mining in various research works. The use of a correct performance metric (measure) gives way for benchmarking with future research studies and against previous studies.

2.15.6 Kappa Static

This metric is used to measures the agreement levels between actual opinions and model predictions. It provides more insight about possible chance agreement. This helps understand model performance over and above the expectation because of chance (McHugh, 2012).

While the above metrics are applicable to binary classification tasks, multi-class classification tasks require slightly different performance metrics shown in Table 2.2.

Table 2.2. Multi-class Classification Performance Metrics

Measure	Formula	Evaluation focus
Average Accuracy	$\frac{\sum_{i=1}^{l} \frac{tp_i+tn_i}{tp_i+fn_i+fp_i+tn_i}}{l}$	The average per-class effectiveness of a classifier
Error Rate	$\frac{\sum_{i=1}^{l} \frac{fp_i+fn_i}{tp_i+fn_i+fp_i+tn_i}}{l}$	The average per-class classification error
$Precision_{\mu}$	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} (tp_i+fp_i)}$	Agreement of the data class labels with those of a classifiers if calculated from sums of per-text decisions
$Recall_{\mu}$	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} (tp_i+fn_i)}$	Effectiveness of a classifier to identify class labels if calculated from sums of per-text decisions
$Fscore_{\mu}$	$\frac{(\beta^2+1)Precision_{\mu}Recall_{\mu}}{\beta^2 Precision_{\mu}+Recall_{\mu}}$	Relations between data's positive labels and those given by a classifier based on sums of per-text decisions
$Precision_M$	$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i+fp_i}}{l}$	An average per-class agreement of the data class labels with those of a classifiers
$Recall_M$	$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i+fn_i}}{l}$	An average per-class effectiveness of a classifier to identify class labels
$Fscore_M$	$\frac{(\beta^2+1)Precision_M Recall_M}{\beta^2 Precision_M+Recall_M}$	Relations between data's positive labels and those given by a classifier based on a per-class average

2.16 Statistical Tools and Libraries for Comparative Opinion Mining

To develop the hybrid machine-learning model for comparative opinion mining, the following statistical tools and libraries (available in Python programming language) are important. This is more especially because Python is the most popular and powerful programming language for machine language projects today.

i. Natural Language Toolkit (NLTK): this is one of the most powerful and widespread libraries in processing natural language. It interfaces easily with numerous corpora as well as lexicons. It has other text processing libraries (Sundaram et al., 2023).

ii. Scikit-Learn: this is a popular ML library in Python for NLP. It has data analysis and data mining tools for text classification and sentiment analysis (Sundaram et al., 2023).

iii. Pandas: pandas library is used to clean and pre-processing of comparative textual data to make it fit for analysis (Sundaram et al., 2023).

2.17 Datasets

Datasets for comparative opinion mining are significantly fewer than datasets available for direction opinion mining because comparative opinions are 10% of user opinions. Among those comparative opinion datasets available, some of them have imbalanced classes, meaning, a certain opinion polarity has a high occurrence hence causing incorrect or biased classification , which leads to wrong pinion analysis (Alkharabsheh et al., 2022). For balanced classes where there is near-even distribution of opinions in the positive and negative opinion classes, model-overfitting issues do not easily arise and the accuracy metric is recommended for model performance evaluation. However, in cases where class imbalances occur, f1-score evaluation metric is recommended as it a harmonic representation of recall and precision. Credible sources of opinion mining datasets include Kaggle's datasets and UCI's datasets (Ghag & Shah, 2018). For example, For example, Younis and others (2022) obtained three comparative reviews datasets from kaggle.com for their research. The datasets were: Microsoft vs Google, Facebook vs Twitter, and Pearl Contintental vs Marriott.

2.18 Applications of Comparative Opinion Mining

Approximately 80% of data in this world remains unstructured. Common forms include chats, reviews, social media, or text documents. As a result, manually gaining insights from such data could prove both time-consuming and difficult. Conversely, with opinion mining systems, firms can quickly get insights through automation processes (Varathan et al., 2017).

2.18.1 Benefits of Opinion Mining

i. *Scalability* – it is conceivable to process data without incurring high costs and in a proficient fashion irrespective of the size of data (Sharma et al., 2017).

ii. *Real-time analysis* – for an interested entity to identify vital information instantaneously so that a state of affairs can be brought to the responsiveness of the business for instant action. This is key in handling public relationships and reputation in the wake of rising dependence on reviews (Malik & Kumar, 2018).

iii. *Consistent criteria* – opinion mining leads to objective opinion evaluation since humans vary in terms of opinion polarities on many subjects (Sharma et al., 2017).

2.18.2 Applications of COM in Brand Reputation Monitoring

Opinion mining has a wide range of real-life application areas. These include brand reputation monitoring, market analysis and research, customer service, social media monitoring, voice of customer, and product analytics (Malik & Kumar, 2018). Granting brands have access to vast sizes of data existing on online review websites like Amazon.com and social media platforms like X from where they can monitor mentions of their brands, analyzing the worth of such brand mentions. The ensuing are uses of opinion mining in brand reputation monitoring:

i. To examine online consumer reviews to obtain insight into user opinions.

ii. To mechanically categorize as critical all online mentions of a specific brand.

iii. To automatically alert selected employees of a brand about explicit online mentions linked to their work areas

iv. To perform brand monitoring tasks automatically through computing technologies.

v. To gain an enhanced appreciation of a precise brand's online manifestation over access to exciting business acumens besides data analytics.

A rising number of people and businesses have espoused review websites in current years hence forming an upward trend. Review websites are important in the boosting of the sum of interactions amid customers and businesses. Likewise, existing consumers entail people that enjoy sharing their views or opinions around what is of interest to them. The sites are now vital in corporate advertising and brand reputation management. Customers make available their opinions on, for example, Amazon review website freely, just as they do on other platforms like X and YouTube. The reviews may comprise consumer complaints and sentiments about the business. Companies, conversely, in a bid to leverage such online reviews, are interested in opinion mining to draw valuable insights (Sharma et al., 2017).

The existence of plentiful online sites accessed by means of electronic devices, for example, smartphones, and computers has caused it to become multifaceted for business intelligence players and salespersons on social media to handle brand reputation. In keeping with Perakakis et al. (2019), the application of machine learning eliminates the intricate challenges associated with scrutinizing the massive data volumes on internet-based podia and offers key benefits for instance improved accuracy, task computerization, compact operational costs, and minimized human power. With machine learning, businesses can make shrewder strategic decisions about how they position their brands before their clients by primarily attaining a superior appreciation of their consumers' opinions concerning their brand. By understanding

customers, brands increase the capability to regulate suitable promotion messages, advance their content marketing policies, discover apt influencers, and get instinctive statistics on their clients. Consequently, a painstaking comprehension of business clients guarantees a brand of improved social advertising tactics. Correspondingly, this makes room for making informed choices and endorsements. This is a key use of mentions (Perakakis et al., 2019).

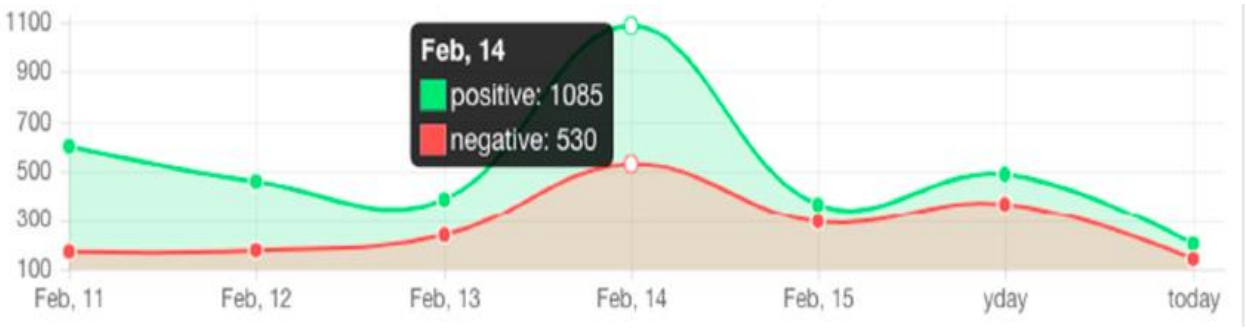

Figure 2. 8 Daily Opinion Analysis

(Source: Perakakis et al., 2019).

2.18.3 The process involved in brand reputation monitoring

The processes necessary in brand reputation monitoring has several stages. First, the brand deliberately analyzes available opinion reviews over a precise period to help it determine opinions targeted at the brand. Second, the urgency expressed in the mentions is categorized automatically, depending on the specific corporate interests. Third is the sending out specific mention-related notifications to designated teams of the brand. This process is automated to permit brand management teams to stay aware of what people say about the brand for purposes of initiating corrective measures. Fourth is to accomplish automation of opinion mining tasks and processes. The last stage is through action on a variety of business analytics and insights obtained through opinion mining; a better brand reputation is sustained (Istrati & Ciobotaru, 2022).

This approach of brand reputation monitoring could benefit a brand in various ways:

i. Gaining an understanding of the evolution of a brand's reputation over time.
ii. Understanding business rivals through mentions that include them.
iii. Identify probable public relations watersheds and take applicable actions. This necessitates that a brand prioritizes mentions that call for a speedy response
iv. Implement opinion mining for definite business events or in the course of time.

2.19 Challenges facing Opinion Mining

There are numerous unsettled issues in the field of opinion mining. These problems are consistent across the majority of studies in opinion mining (sentiment analysis). It should be noted that there is no single technique or approach for solving all these issues (Borele & Borikar, 2016). This section presents some of the common, important issues in opinion mining. Some of these issues necessitate comparative opinion mining while others are still problematic even when carrying out comparative opinion mining.

2.19.1 Named Entity Recognition (NER)

Named-entity recognition and extraction is aimed at determining entities named in unstructured texts. Examples of named entities include brand names, product names, service names, organizations, events, names of people, locations, and time. In the context of comparative opinion mining, a key element is entity detection. Typically, in comparative opinions, multiple rival entities are mentioned together, with the relationship between them. This shows how they are compared based on a certain feature or business aspect as is often the case in comparative reviews on Amazon Reviews and X (Jahanbin et al., 2019). In line with Kaur (2016), the different mentioned entities must be accurately identified to allow of determination of opinions or

sentiments attributed to each of the entities. The example below shows an opinion review that could be typical of what one would find on Amazon Reviews website or social media platforms like X and YouTube.

"I like Dell laptops. HP sucks!"

Considering the above instance, Kaur (2016) contends that using the simple BOW model would generate a neutral opinion class. However, the statement actually has both positive and negative opinions with respect to the two mentioned entities: Dell and HP correspondingly. Thus, applying entity detection or recognition could help brands to know when their rivals have been mentioned in the same review as the target (interested / inquiring) brand. Such mentions, when examined, can reveal business aspects or features upon which the comparative review was based, even further showing the preferred entity between the two mentioned entities. Such information is useful for brands interested in carrying competitor analysis to identify products and/or services that require improvements to meet customer needs.

One of the ways of handling named entity recognition is by the use of lexicons or dictionaries containing brand names, product names, or service names (Varathan et al., 2017). However, because many brands have alternative names, the lexicons must be created to encompass the most common alternative brand names by which customers refer to the business entities or their products or services. In this study, a lexicon was used for entity identification or detection, considering that entities are key elements in comparative opinion mining.

2.19.2 Entity Relation (Order Dependence)

Since there are at least two comparable entities in comparative opinion reviews, it is vital to establish the direction of the relationship between the entities. This issue is closely linked to the named entity detection issue. This work utilizes machine learning and/or deep learning techniques to learn how the subject and object entities are associated in text. ML or DL models that have been trained on large datasets containing multiple compared entities are capable of establishing a pattern that typifies entity relations in reviews. Using this pattern, the models are then capable of accurately determining the entity relations in comparative opinions. Again, entity relations are one of the key pillars of comparative opinion mining. Below are two examples comparative opinion reviews.

Review #1: "Apple is better than Samsung"

Review #2: "Samsung is better than Apple"

The first review shows Apple and Samsung have been linked by the phrase "better than," thereby revealing the direction of the relationship in terms of opinion preference. The reviewer expresses his / her preference for Apple, meaning, more positive opinion polarity towards Apple and a less positive polarity opinion towards Samsung. In short, the reviewer prefers Apple to Samsung. On the other hand, the second review shows reversed opinion direction because the two entities (Apple and Samsung) have switched their positions while the opinion words remain intact. In review #2, the reviewer prefers Samsung to Apple. Therefore, a change in the order of entities in opinion reviews requires accurate detection and handling to enable coming up with the correct opinion classes for each entity, and further, determining the preferred entities. Yet, this

challenge has not been adequately addressed, calling for further studies in this area to improve the accuracy of entity relation detection (Kaur, 2016).

2.19.3 Opinion Negation

There are cases when an opinion holder (a reviewer) negates an opinion devoid of using common negation words like “never,” “not,” or “no.” This is still a challenge in opinion mining (Kaur, 2016). Even though studies such as those carried out by (Sun et al., 2009) have tried to address negation in opinion mining, there are many non-standard words or phrases people use to express opinion negation. Trying to use a lexicon proves unfeasible in many domains. A better way to handle this is through ML and DL approaches as the algorithms in these approaches can learn the various patterns through which negation is expressed in reviews and be able to use this knowledge to make accurate opinion classification when given opinions containing negation. In the example below, negation is expressed implicitly, making it difficult to detect it using NLP and lexical approaches, as there is no standard negation word used in the review. “*This solution will overcome overspending habits in corporations*.

2.19.4 Domain Dependence

A word or phrase may have different meanings based on the knowledge domains in which it is applied. For instance, *'unpredictable'* bears a positive implication in the entertainment sector despite the fact that this very word has a negative connotation in the autos domain (Kumar & Sharma, 2017). That is why the opinion class derived from a sentence using it could be positive or negative depending on the application domain. This leads to subjective opinion statistics. Training ML and/or DL models on huge datasets in specific domains tends to solve this problem. Besides, transformers models help with knowledge transfer across domains but they require huge training datasets.

2.19.5 Non-English Language Limitations

Most studies have reported carrying out opinion mining using language-dependent datasets. Common languages that have been used include Chinese, Korean, and English. Unfortunately, a number of useful data sources like Amazon Reviews and social media platforms allow people to communicate using multiple languages. However, there is solution for automatically processing opinions in multiple languages. This is why opinion research usually involves datasets in specific languages that the researchers are familiar with, and/or which have been handled or researched upon before. There is, therefore, a growing interest in more research using other languages apart from English, Chinese, and Korean (Kaur, 2016). Besides, language dialects and language nuances like short forms that often appear in reviews on social platforms like X complicate opinion analysis (Kumar & Sharma, 2017). In this work, English-based datasets were used, being the language the researcher is more familiar with, and which comparative datasets were much more readily available in than other languages.

2.19.6 Context and Polarity

Written words find meaning depending on the surrounding words and phrases. Disregarding the context of words may lead to inaccurate classification or analysis of opinions, which consequently leads to errors in decision making when using such opinion results. Many approaches have been proposed for handling context. This work utilizes the higher-order n-gram model, Count Vectorizer and TF-IDF features to tackle this problem. But, transformer models are gaining ground in this regard but they face the challenge of requiring huge amounts of training data and dedicated computational resources to work, making them unfeasible in a number of real-life applications.

2.19.7 Opinion Spam

Social media allow people to express their opinions without restrictions, about any subject matter or entity, short of fearing any adverse consequences and without identifying themselves. Individuals with malicious intentions use this freedom to discredit other entities like business organizations and individuals, aware that their identity is undisclosed. These persons are referred to as opinion spammers while what their actions are called opinion spamming (Jindal & Liu, 2008). Opinion spamming represents a temporal dimension associated with how time affects comparative opinions (Thomas et al., 2011). Opinion spammers may be paid by a brand to discredit a rival brand. Therefore, a more accurate representation of the collective opinion pool from consumer reviews should be devoid of contributions obtained from opinion spams as their inputs cannot be trusted. This is why opinion spam detection as a data mining challenge needs to be addressed in comparative opinion mining (Liu, 2015).

User profiles have a vital role in comparative opinion in regards to achieving more effective and accurate detection of spam opinions. The incorporation of some information that is user-specific, such as engagement patterns, preferences, and historical behaviors, it is possible to develop a context-aware and robust approach to detection spam opinions. Specifically, user profiles yield incredible insights into the characteristics and typical behavior of legitimate users thereby allowing for ways to differentiate between genuine and possible spam opinions. The following sections describes some of the techniques in use in the detection of spam opinions while carrying out opinion mining.

i. **Engagement Patterns (Account Activity and Age)** – the use of user profile metrics including comments, shares and likes can help track user engagement online. A sudden increase or decrease on engagement rates with respect to certain opinions from a specific user could be used to indicate possible spam (Thomas et al., 2011) particularly when such behavior is in variance with historical engagement patterns (El-Mawass et al., 2020).

ii. **Behavioral Analysis** – past interactions, posting patterns, and content preferences of a user are part of a user's profile. The analysis of deviations of these behaviors from the expected norm that is often recognized in sentiment shift or spike could be a red flag for suspected opinion spamming (Castillo et al., 2007; Gao et al., 2010). Using this approach, the hybrid model developed in this research used a user's account ID to detect the presence of multiple, similar tweets from the same user and ignore duplicate tweets from that user to avoid the opinion bias during opinion classification.

iii. **Consistency and Content Relevance** – the expertise and interests of a user are in the user's profile. If there is a drastic shift in the topics of interest such that the user is rapidly sharing content on an unrelated sentiment or opinion subject matter, this may be a reason to suspect opinion spamming from such a user (Castillo et al., 2007). Irrelevant and repetitive content is an indicator of spamming from automated tools like bots (Davis et al., 2016). Due to the complexity involved in using this approach, this study did not attempt this solution.

iv. **Social Network Analysis** – in the case of social networks like X, the social network itself provides useful information about a user's connections, together with his/her followers and relationships. Unusual interactions in a user's interactions with users of low-quality accounts or unusual behaviors might be indicative of potential spam opinions (Meusel et al., 2014). Attempting to solve opinion spam through this method would have had significant time, skill, and financial implications to the researcher. These three, were limitations the researcher faced in attempting to explore this option.

v. **NLP Techniques** – analyzing a combination of opinion text with user profiles using NLP techniques can help detect anomalies in sentiments, language usage, and writing style. Spams often exhibit specific linguistic characteristics that make it dissimilar to classic language patterns. This information could be used to indicate the presence or absence of spam opinions (Potthast et al., 2018). The researcher found it computationally resource intensive to try using this approach using techniques like deep learning approach in opinion spam detection.

vi. **Bot Detection Algorithms** – Artificial Intelligence driven bots could be used to detect the use of bots in opinion spamming. The AI bots could study the features obtained from user profiles on how the users interact with others to help improve on the accuracy of detecting spam (Davis et al., 2016). To use this approach, the researcher needed to integrate the developed hybrid model with an existing AI bot specifically for spam opinion detection. Unfortunately, the researcher was not familiar with ways of integrating existing bots with the developed hybrid model.

2.19.8 Lack of a Universal Opinion Mining Algorithm

No existing algorithm is proficient in addressing all the problems in opinion mining. For this reason, different studies usually apply various approaches and algorithms to try to achieve improved opinion analysis. For instance, this study entailed the development of hybrid machine-learning model for comparative opinion mining. Theoretically and empirically, hybrid models outperform the single models from which the hybrid is created thus offering better performance (Saberi & Saad, 2017).

2.20 Challenges in Comparative Opinion Mining

COM is aimed at analyzing opinions about mentioned entities that are related by some aspect(s) or feature(s). It is a challenging task carrying out opinion mining involving two entities. Thus, extending the task of comparative opinion mining to three or more entities poses great additional complexities. These complexities include the following:

2.20.1 Data Annotation Effort

Creating labelled datasets that could be used in ML model training for COM involving multiple entities is very resource-intensive and time-consuming. This has a direct impact on the availability of training data containing three or more comparable entities (Maas et al., 2011). While this study created a few datasets for model training and validation, the datasets were relatively small. To create bigger datasets, a lot more time and financial resources are needed for data collection and annotation.

2.20.2 Increased Dimensionality

The dimensionality of the data increases when the researcher has multiple entities involved in opinion analysis. This is because each of the entities involved creates a new dimension making it difficult in attaining effective interpretation and visualization of

the results. Each entity requires three labels. Thus, the more entities one has to handle, the more the labels. For instance, three entities would need nine labels while five entities would need 15 labels. An increase in the number of labels causes a direct increase in dimensionality making it difficult to prepare data for training and testing models for such high dimensionality data. For this reason, the researcher chose to work with two entities. This study strictly selected datasets with two entities to ensure that challenges associated with three or more entities were avoided.

2.20.3 Pairwise Comparisons

With multiple entities, the number of probable pairwise comparisons increase exponentially. Analyzing and managing such comparisons is both resource-intensive and complex (Ghose & Ipeirotis, 2011). Working with two entities in this study involved nine different pairs, based on the opinion polarities assigned to each entity in the opinion reviews. Attempting to process more than two entities will make pairwise comparisons among mentioned entities even more difficult. Thus, the researcher restricted this study to comparative texts with two entities only.

2.20.4 Data Sparsity

It becomes rarer to get available data containing multiple comparisons. This leads to data sparsity. This is because comparative opinions form only 10% of the available opinionated content. This problem was evident when looking for comparative opinion datasets to train and validate the model. While datasets for direct opinion mining exist in bigger numbers and sizes, those for comparative opinion mining do not exist in great numbers or sizes. To overcome this challenge, the researcher attempted to create a new dataset. Unfortunately, due to time and computing resource constraints, the study could not collect a good number of records.

2.20.5 Entity-entity interactions

With multiple entities, the interactions and relationships between the entities become elaborate leading to dynamic sentiment complexities and technical challenges. While this study relied on the use of machine learning and deep learning algorithms to detect and learn the patterns on how entities interact in comparative texts, the presence of more than two named entities in the same comparative make it difficult to train a model on the interactions. However, deep learning approach may solve this problem but it is faced with the challenge of demanding greater computational resources, which were not available to the researcher during this study.

2.20.6 Contextual and Semantic Challenges

With multiple entities, additional nuanced analysis of how the entities semantically relate with each other and the contextual relationships involved is needed. Higher order n-grams like trigrams, used in this study, for instance, are simple solutions to the problem of contextualizing text for purposes of obtaining better semantics on the comparative opinions (Liu & Forss, 2014; Yan, 2022). Transformer models could offer a better solution to the problem of opinion contextualization through more effective feature extraction (Devlin et al., 2019). However, their use demand high computational resources, which were a constraining factor in this study.

2.20.7 Interpretable Visualization

Results obtained from comparative opinion mining become increasingly difficult to visualize as the number of entities involved, which is caused by complex multi-dimensional sentiment relationships. This problem was apparent right from dataset annotation to determination of preferred entities where multiple entities are mentioned

together in a review. This study focused on two entities to minimize visualization challenges for better results interpretation.

2.20.8 Entity Disambiguation

It is challenging to recognize and disambiguate the mentions of many entities in text content. Accurate entity reference resolution is critical in obtaining meaningful opinion analysis and this is affected negatively as the number of entities increases in text (Cucerzan, 2007). While transformer models could potentially offer a solution to this problem due to their powerful feature extraction capabilities and relationship mining (Devlin et al., 2019), this study did not use them because of the high computing resource requirements associated with them.

2.20.9 Scalability

Scalable computational resources are needed to handle comparative opinion mining tasks like feature extraction, sentiment analysis, and data processing for cases of multiple entities (Cambria & Hussain, 2012). Working with traditional machine learning algorithms like Random Forest is computational efficient and less resource intensive. However, for high dimensionality data and larger datasets, deep learning algorithms perform better. However, this performance comes at a cost of high computational resource requirements like specialized Graphical Processing Units (Chollet, 2017) that many researchers and application developers and may not have access to, which limits their use in real world applications.

2.20.10 Opinion Fusion

With multiple entities, summarizing, weighing, and aggregating diverse sentiments for every comparison is required. Performing these tasks is difficult in the case of multiple entities. Because multiple entities introduce multiple columns to represent opinion polarities (Yueyang & Wang, 2019) for each mentioned entity, it means that increasing the number of entities would result in increasing the number of columns to represent the polarities for each mentioned entity. The result of this is opinion fusion, where it becomes difficult to associate an opinion or sentiment with a particular brand entity mentioned in the text.

2.21 Research Gaps in Comparative Opinion Mining

This section presents the research gaps in COM identified in this study.

i. Hybrid machine-learning models have been developed to perform direct opinion mining, showing improved performance. However, to the best of the researcher's knowledge, there exists no hybrid machine-learning model for comparative opinion mining. This study had a primary purpose of filling this gap, which was achieved by combining two different algorithms to leverage their strengths while minimizing their weaknesses in the hybrid model (Sagi & Rokach, 2018).

ii. Comparative statements in general texts like web documents, news, or other scientific texts may be factual and not opinionated (Bos & Nissim, 2006; Wan & Xiao, 2011; Park & Blake, 2012; Chang & Jin, 2012). It is difficult to distinguish between factual and opinionated textual content posted online. To overcome this challenge, the researcher relied on secondary datasets to train and perform initial model validation. To ensure that the texts contained comparative opinions, the researcher employed three human experts to annotate the datasets obtained.

iii. There is scanty research on comparative sentence detection using supervised machine learning techniques. Most existing studies were conducted on Chinese and Korean languages. However, this study relied on the few existing English-based comparative opinion reviews. This gap needs to be filled to avoid relying on already collected datasets in specific languages.

iv. Unsupervised machine learning techniques have not been widely employed in comparative sentence detection. To the best of our knowledge, there is limited research in this area hence a need to explore this area in the future. This study experimented the application of K-Nearest Neighbor (KNN) algorithm in performing comparative opinion mining. However, the algorithm underperformed in all tests. Perhaps, improved versions of KNN need to be developed to attain higher classification accuracies.

v. The use of comparative and superlative words in COM is ineffective because some comparative words like “best” may carry no opinion in some sentences like “it is best to trust in God.” Therefore, identifying this sentence as comparative or applying the word to determine a preferred entity would be erroneous.

vi. Some sentences have no comparative words yet they are comparative opinion sentences. Lexical approaches fail to identify such sentences like that correctly.

vii. It is still difficult to handle COM involving multiple entities. While handling two comparable entities has been tried out, there is limited research on how to handle three or more entities. The challenge includes handling data imbalance, model

overfitting and dimensionality, increased complexity, limited training data, and a need for sufficient model evaluation metrics (Bengio et al., 2013).

viii. Recency Bias – when opinion-mining models are subjected to analyzing recent comparative opinions, the chance of relying on recent opinion result increases, introducing a bias against older or historical opinions. Thus, analyzing recent comparative opinions and ignoring older comparative opinions may affect opinion comparison due to time effect (Das et al., 2018).

2.22 Conceptual Model

The conceptual model in this work was derived from the research objectives of this study. It guided this study. Besides helping the research decide on the topics to cover in the literature review chapter, the framework was key in the design of the research methods and experiments, analysis of the results, discussions of the results, and the conclusion of the study. This is because a change in each of the elements of the conceptual model has a potential of changing the outputs of the developed model due to the interrelations or interactions among the elements. The framework described in this section, involves four key components: a learning algorithm, feature extraction technique, and dataset, which together help predict the opinion class of a comparative opinion text. The predicted class, based on accuracy, translates to a brand's reputation when applied to brand reputation monitoring.

2.22.1 Conceptual Elements

The framework presented here was used to establish the various comparative opinion elements and applying machine-learning algorithms to develop a hybrid-machine learning model that would effectively perform COM. However, to develop a hybrid ML

model, comparative datasets and feature selection/extraction techniques were a key consideration. Therefore, comparative opinion elements (opinion targets or brand entities, entity relations, brand aspect features, and opinion/sentiment words), feature selection techniques, and ML or DL algorithms are essential elements. The developed model requires to be evaluated to validate its performance. To achieve this, accuracy and f1-score were important evaluation metrics as they help determine the most optimal model for use. Algorithms represent the representation of knowledge in the model.

The association amid the independent and dependent variables, their operationalization, as well as the influence from the moderating variables were key. In Section 2.3, the researcher found that brand reputation is affected by comparative consumer reviews or other comparative user-generated content. Thus, there was need for a technique for detecting or identifying entities. With entities identified, opinion mining and user profile analysis were then used to complete the model. The framework addressed two comparable entities in each comparative opinion review.

2.22.1.1 Independent Variables

Machine Learning Model - different ML algorithms can be applied to comparative opinion with potentially varying accuracy levels in the manner in which they will classify opinions. These models are capable of detecting the following four elements of comparative opinion data: a comparative sentence, multiple entities, entity relations, and comparative feature. All these elements are found in comparative opinion data.

Table 2.3 Elements of Comparative Opinions

Comparative Opinion Element	***Comparative Opinion Mining Element***
Sentence (A comparative sentence)	Comparative Sentence Detection
Entity (Multiple Brand Entities)	Entity Detection
Relation (How the entities are compared)	Relation Detection
Feature (Brand aspect in the comparison)	Feature Detection

A comparative opinion sentence contains multiple entities, a relation between the entities, and a feature or aspect upon which the entities are compared. To determine the reputation of a brand as being positive, negative, or neutral, the four elements of a comparative opinion (mining) must be analyzed. Consider these two comparative opinion sentences, for instance:

i. Nokia phones have better signal quality than Samsung phones.

ii. Samsung phones have better signal quality than Nokia phones.

In the first sentence above, Nokia has a more positive brand reputation than Samsung. In the second sentence above, Samsung has a more positive brand reputation than Nokia. Hence, besides determining the entities compared in a text review, it is vital to detect how the entities are related (e.g. A is better than B) and the brand aspect (e.g. signal quality) compared.

2.22.1.2 Moderating Variables

The moderating variables in this framework have the function of adjusting the brand reputation class by altering the intensity of the relationship between the dependent and the independent variables. In this study, these were moderating variables:

a) Feature Extraction Technique / Model

To detect and extract the COM elements, a machine learning algorithm was considered suitable for use due to the nature of the data involved (unstructured) and the limitations of other approaches like association rule mining, natural language processing, and statistical methods.

1. Machine Learning Algorithm (or Deep Learning Algorithm) - different machine learning and/or deep learning algorithms present different attributes that cause machine-learning models to attain different performances. This has been observed in many studies including a study by Banihashemi et al., (2017) that led to the development of a hybrid model for building energy consumption.
2. Feature Extraction Technique – different feature extraction techniques result in different model performances (Ligthart et al., 2021; Ondara et al., 2022). The parameters used feature extraction influence feature extraction. For example, when using the N-gram model, the value of n (n-gram range value) affects the general accuracy of classified opinions (Ligthart et al., 2021).

b) Dataset Choice

The datasets required for comparative opinion mining must have at least two comparable entities. Each dataset had reviews whereby each review must mention the entities being reviewed. According to Kleinberg et al. (2018), different dataset yield varying accuracy levels when performing comparative opinion mining with the same machine learning model.

2.22.1.3 Dependent Variable

The dependent variable in this study is Accuracy. The accuracy was used to determine the *Opinion* Class. The reputation is computed from the opinion polarity (Positive, Negative, or Neutral) from each opinion analyzed. A positive opinion class relative to a brand implies a positive brand reputation. Conversely, a negative sentiment class relative to a brand implies a negative brand reputation. A neutral sentiment class implies that the analysis of the opinion holders' opinion reveals a lack of decisiveness on whether the brand's reputation is positive or negative. Opinion class visualization is a processed, visual form of the opinion classes. User feedback or satisfaction relates to feedback from users involved in ground truth testing.

2.22.2 Operationalization of Variables

A study by Varathan et al. (2017) identifies the accuracy of a machine learning model is a preferred model evaluation metric for measuring the performance of a model in comparative opinion mining. According to Liu, Xia, and Yu (2021), a comparative opinion sentence has four elements, which include subject (subject entity), object (object entity), comparative aspect, and comparative opinion. Thus, these four elements need to be detected first. But since we this study used comparative opinion datasets, there was no need to detect sentences. Also, the entities were manually input to reduce computational resource demands in case Named Entity Recognition were to be used, there was no need to detect entities. Therefore, the primary tasks for the machine learning model in relation to operationalization of variables was to detect features and relations between features to help determine the opinion label / class.

1. **Sentence:** A comparative sentence is the basic unit of user-generated content that holds comparative opinions. The other three elements of comparative opinion mining are identified and/or extracted from textual material.

2. **Entities:** This entails the brand entities identified in comparative opinion texts like reviews. The opinions were analyzed relative to the opinion polarities directed at the target brand. This may affect the opinion class of the brand.

3. **Relations:** **T**his refers to the type relationship linking the (brand) entities. These could be comparative or superlative connectors that could be mined through NLP, rule mining, or machine learning approaches.

4. **Feature:** This is the aspect within a comparative sentence. It is upon this feature that a comparison between the mentioned entities is based.

Dependent Variables

Accuracy: the two model evaluation metrics used predominantly on COM related research are: accuracy and f1-score. The accuracy of the model was used to infer the brand reputation class (positive, negative, or neutral). The values for both accuracy and f1-score were quite close. Brand reputation was directly mapped from the dependent variable (accuracy), representing the opinions held by clienteles concerning the brand's business products, or services. In this study, changes in the opinion labels for a given brand due to effect of opinions targeting rival brands gave an improved examination of a brand’s reputation. Based on the above variables, the following conceptual model was developed.

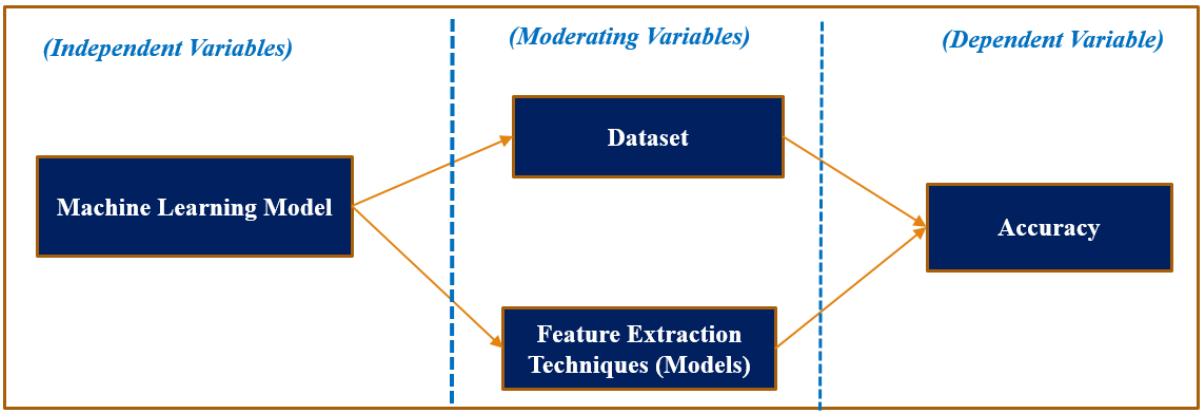

Figure 2. 9 The Conceptual Model

Current literature shows that a brand's reputation is influenced or impacted by the opinions of customers about the brand in relation to its products and/or services. Besides, a brand's reputation is unpleasantly impacted when the brand's customers compare its products or services with those of rival brands. Trust levels of consumers have a significant effect on how they (customers) understand and respond to their brand's business campaigns. However, some users may use fake content, resulting in low quality, erroneous, and defective opinion results.

2.23 Conceptual Process Model

This is a graphical depiction of the main constituents of a process. It is used to bring about the definite development of a process, or system. The realization of a conceptual process model is often the development of a complete system or system prototype. In this study, this was used in developing a prototype for COM. The conceptual process model shown in Figure 2.10 is for comparative opinion mining. The process starts with data pre-processing which involves either one or both of these two tasks: feature engineering and feature selection. For this study, only feature selection was done. Upon selecting features, the next task is model training using training datasets. After training, model testing / evaluation is done using the training dataset. Thereafter, a COM

classification model is applied on comparative opinion data that is unknown to the trained model to obtain classification results that could be applied to brand reputation monitoring. As data is obtained for classification by the trained model, additional tasks like spam opinion removal are performed to make the results more reliable.

The opinion class produced by the classification model is then used to indicate the reputation of a brand. For example, a higher number of positive opinions than negative opinions about a given brand could be interpreted to mean the brand has a positive reputation. The reverse is also true where the number of negative opinions exceed the number of positive opinions for a given brand. In a case where the positive and negative opinions are the same, then the brand's reputation is classified as neutral.

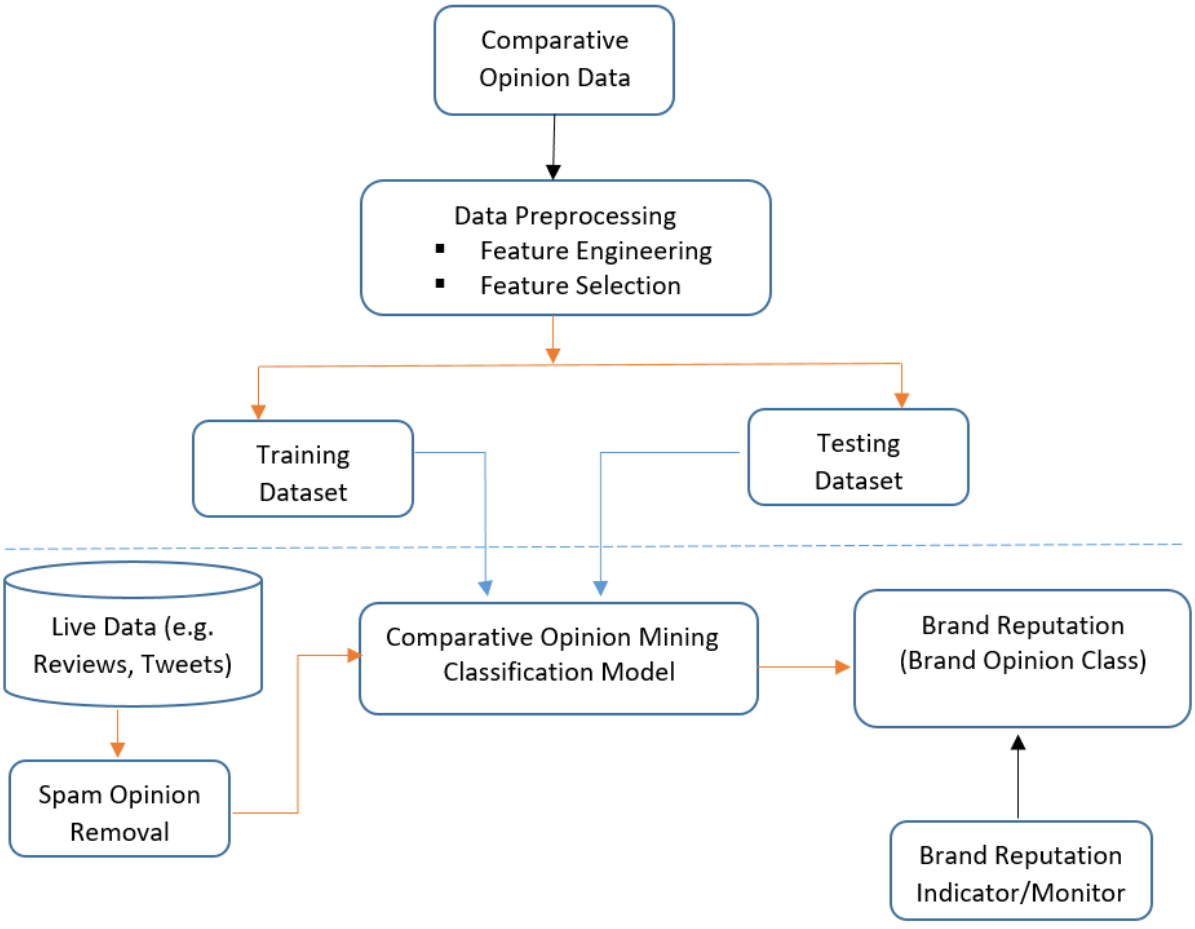

Figure 2. 10 Conceptual Process Model

2.24 Chapter Summary

This chapter described the theoretical framework for comparative opinion mining (COM) as it relates to brand reputation monitoring. Specific theories behind brand reputation monitoring, algorithm selection and feature selection criteria were explored. The chapter presented the background to COM, the elements of COM, the various approaches to COM, specific algorithms for opinion mining and comparative opinion mining, together with the relevant feature extraction techniques and model evaluation metrics. A key finding from this chapter was that there were limited studies on the use of hybrid ML models for COM. In the subsequent chapter, the state-of-the-art ML algorithms like Support Vector Machine, Naïve Bayes, Random Forest, and Stochastic Gradient Descent; and DL algorithms like CNN, RNN, LSTM, and Multilayer Perceptron described in this chapter would be validated or challenged through a pilot study. This would be followed by an empirical study to establish their suitability for use in building a hybrid ML model for COM. Feature extraction techniques like count vectors and TFIDF, which integrate well with the BOW feature model were adopted, as they are more suitable for COM due to data sparsity reasons. The ensemble learning method, being theoretically and practically effective in developing hybrid models was adopted for use in the empirical study. Reviewed literature showed that accuracy was the commonest metric in model performance evaluation for COM especially for balanced datasets while F1-score was preferred especially for imbalanced datasets. These metrics were used in evaluating the effectiveness of the developed hybrid ML models for COM and applying the model to determining the reputation of a brand using comparative opinion data.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter describes the methodology employed in this study. The mixed methods research methodology was adopted thus integrating qualitative research that involved analyzing of qualitative data with quantitative research that involved analyzing of quantitative data. This methodology leveraged the advantages of qualitative and quantitative methods thus increasing the validity of the research findings. This chapter is organized as follows; Section 3.2 presents the research philosophy adopted and a justification for its use. Section 3.3 provides an overview of the research design for both the pilot study and the main experiments. This is followed by 3.4 that describes the research instruments. Section 3.5 presents the pilot study followed by Section 3.6 on the sampling strategy and sample size. Section 3.7 covers how the research hypotheses were tested followed by 3.8 on model development methodology. Section 3.9 covers the Design of hybrid ML model. Section 3.10 entails prototype development followed by Section 3.11 on data analysis. Lastly, Section 3.12 presents the research ethics.

3.2 Research Philosophy

A research philosophy is a belief that a researcher maintains about how he/she will collect and scrutinize data about a specific research phenomenon in a bid to create new knowledge (Muhaise et al., 2020). For this study, the researcher employed the philosophy of pragmatism. This philosophy assumes that research concepts are simply germane in cases where they put up with action (Dixon, 2020). The researcher began this study with problem identification then proceeded to ascertaining his aim towards causing real-world solutions that enlighten and help future practice. Principally, the

researcher began this work with a sense of uncertainty that somewhat, something was amiss within the chosen area of study. At the conclusion of this work, the researcher had a conviction that the problem was solved. Pragmatic researchers concentrate on creating practical solutions to real-world problems (Muhaise et al., 2020).

The researcher selected this philosophy for two reasons. First, it is applicable in scientific research where an entity or artifact (e.g. a product or service) is to be developed to help improve lives (positive impact on life). In this study, a useful machine-learning model was developed and a prototype created as a proof of concept. Second, it incorporates more than a few concepts and additional research philosophies, consequently, being more typical. This consists of subjectivism and objectivism, among other contextualized experiences. Since the research objectives focused on developing a ML model, the pragmatic philosophy provided guidance.

3.3 Research Design

This section describes how the research was designed from the perspective of the research design process, the experimental design, how the main experiments were designed, and the data collection tools used.

3.3.1 Research Design Process

Considering the pragmatic philosophy guiding this study, the researcher employed mixed methods approach. With this approach, four methodologies were used. First, qualitative, and quantitative research designs were used to guide the research. Second, system prototype development was carried using the Integrated Desktop Development Process (IDP). Finally, to assist with system model evaluation, primary experimental research design and secondary experimental research design were used. The whole

research process was conducted in three key phases: problem identification, solution design, and evaluation, according to Offermann et al., 2009, as shown in Figure 3.1.

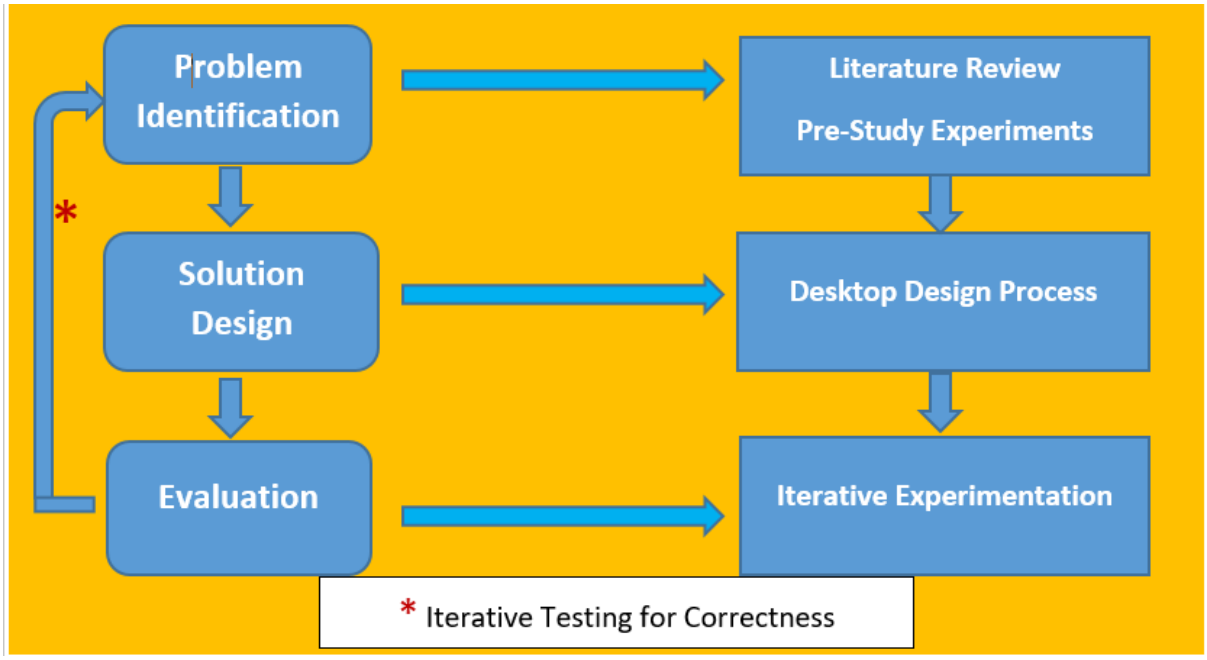

Figure 3. 1 Research Design Process

The problem identification phase required qualitative as well as quantitative research methods. Qualitative research was applied in document analysis to explore the state-of-the-art machine learning models, features, datasets, and model evaluation metrics used in opinion mining. The solution design phase involved developing a hybrid machine-learning model for comparative opinion mining. This leveraged the findings of the pilot study. In relation to Figure 3.1, this research had one primary problem to address: developing a hybrid machine learning model for comparative opinion mining. Once this need was clearly understood, the research design process required that a solution be designed, which was achieved through the desktop design process. Finally, the designed solution was evaluated iteratively using different variables including the ML or DL or Hybrid algorithm, dataset, and feature extraction technique. The intention of iteration is to determine the most optimal combination of algorithm, feature extraction technique, and feature parameters.

Pilot studies helped the researcher to gain insight, improve the research methodology, and validate if the approach was feasible or not (Malmqvist et al., 2019). In this study, for instance, the pilot study also helped to explore the data acquired for model training and testing in terms of its quality and volume, which is important in the development of machine learning models. Secondly, pilot studies help researchers to determine and select relevant features for various use cases including comparative opinion mining (Omar et al., 2014). Another key benefit of pilot study is the use of pilot study findings to select and tune algorithms for use in the main experiments. In this case, a pilot study gives a researcher an opportunity to evaluate the performance of a machine-learning algorithm prior to the main experiment, aiding in selecting and tuning the best performing algorithms from the pilot study.

In dataset labelling and annotation, a pilot study helps a researcher to test if collecting and manually annotating data is feasible or not (Khomsah et al., 2023), which is instrumental in the decision to use or not to use secondary datasets. In addition, pilot studies help researchers to evaluate a model's performance (González-Gonzalo et al., 2022). At the stage of pilot study, the researcher identifies the appropriate model evaluation metrics to apply in the main experiment based on realistic results from the pilot study. Finally, pilot studies help in the development of hybrid machine-learning models through the validation of the effectiveness and compatibility of the different approaches used in developing the hybrid model. This phase also benefitted from relevant theories concerning system design and development (Offermann et al., 2009).

For this study, IDP was used to design the system prototype following four key phases: needs analysis, conceptual design, prototype development, and prototype evaluation (Nam & Smith-Jackson, 2007). In the third phase of this research design process, the developed hybrid machine-learning model was evaluated through experimentation, using both secondary and primary data. This was done to assess the effect of comparative opinion analysis on the sentiment score of a brand and subsequently, the reputation of that brand. The developed system prototype was availed to three experts for one week.

This would help to gauge the effectiveness of the HML Model in monitoring the reputation of a brand based on comparative opinions generated by social media users. After this period, a discussion with the three experts was conducted to collect feedback on the usefulness of the HML Model. Ground truth data was obtained from these experts, having been given 100 comparative opinions to classify according to the respective opinion polarities for each entity mentioned in the text, and then comparing their results with those obtained using the HML Model on the same texts. This helped attain human validation of the effectiveness and efficiency of the HML Model for use in brand reputation monitoring. Figure 3.2 shows the mapping of research objectives, and research questions to the respective research methods used to address the research objectives and research questions.

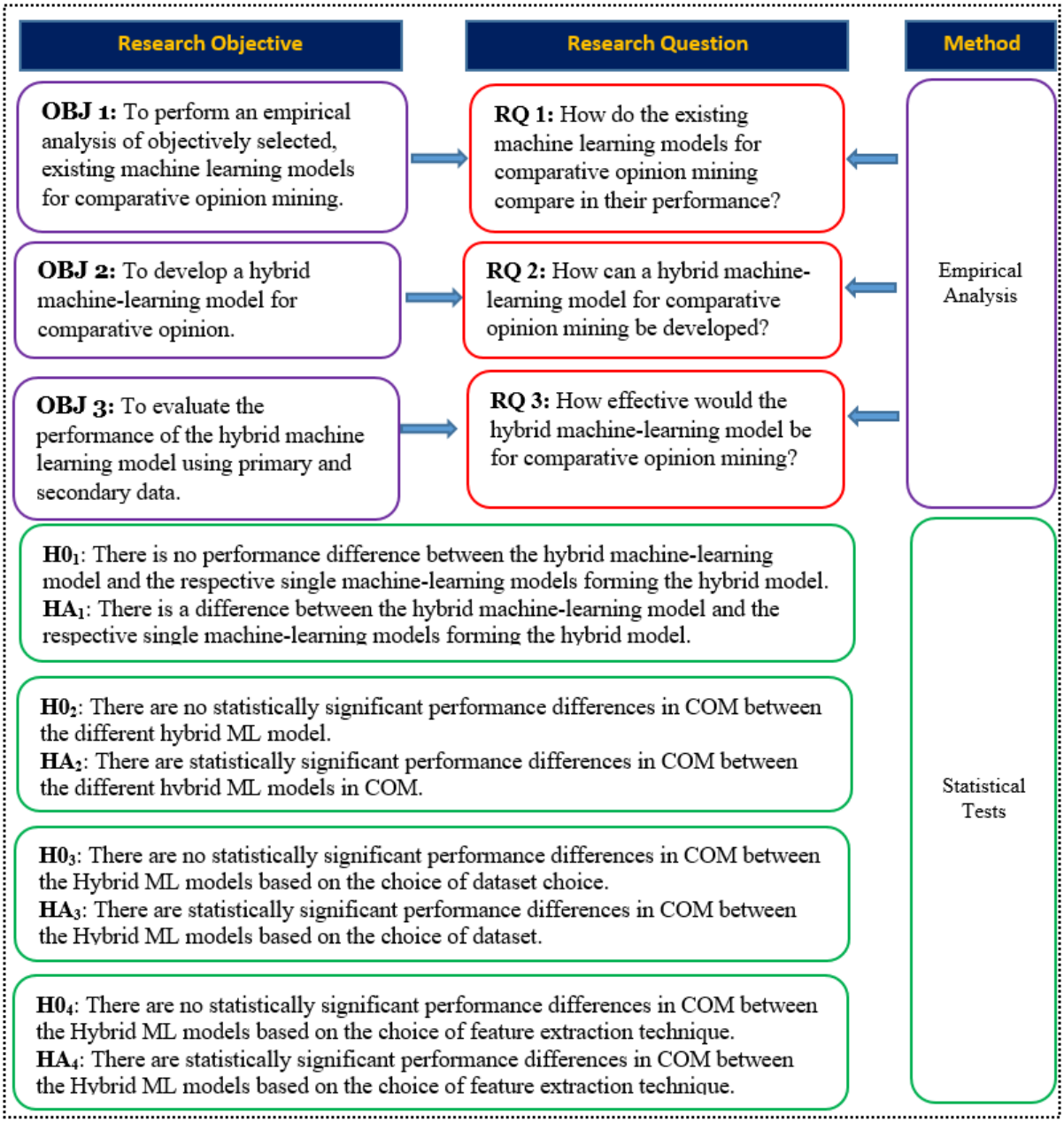

Figure 3. 2 Mapping of Research Objectives, Questions, and Hypothesis to Research Methods

3.3.2 Experimental Design

This study applied the Integrated Design Process (IDP) methodology in the development of the system prototype that was used to carry out experiments on monitoring brand reputation using comparative opinion mining. This methodology emphasizes the principle of dynamic system design principles (Nam & Smith-Jackson, 2007). In this methodology, Offermann, Levina, Schonherr, and Bub (2009) emphasize

the need to design systems that aim at meeting system requirements based on theory and experimentation (e.g. through pilot study). The process entails four phases: needs analysis, conceptual design, prototype development, and system testing (Nam & Smith-Jackson, 2007). Figure 3.3 illustrates this process.

This design was led by the interaction between / among the independent variable (brand reputation class), moderating variables (opinion classification algorithm, feature extraction technique, and N-gram range or window –size), and the independent variables drawn from the comparative opinion mining elements (sentence, entity, relation, and feature). Using machine learning and deep learning, the selected algorithms were trained using annotated data to learn the relationships among / between the various comparative opinion mining elements and the trained model was then exposed to new, unlabeled data for classification. Subsequently, after the model demonstrated satisfactory classification accuracy, a web-based prototype was developed to demonstrate the application of the hybrid machine learning model in comparative opinion for automating the process of brand reputation monitoring in different domains including banking, telecommunications, and automotive. Figure 3.3 shows the methodology used in designing our experiments.

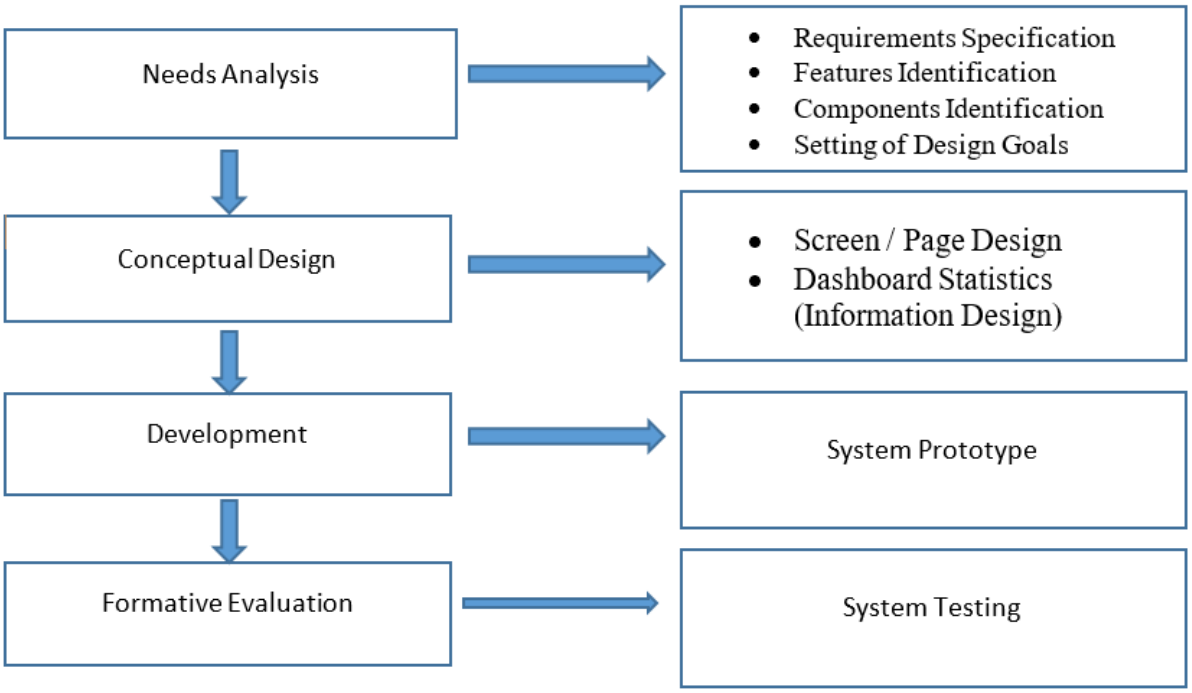

Figure 3. 3 Integrated Design Process (IDP) Methodology

(Nam & Smith-Jackson, 2007)

The contents obtained from the Integrated Design Process shown in Figure 3.3 were instrumental in understanding how the design of the system for this study would be handled. For example, the needs analysis phase led to the identification of necessary components for the system prototype. The conceptual design informed the choice of dashboard components to help in communicating brand reputation statistics over time. The development phase was used to inform the development of the prototype while the formative evaluation phase in the IDP was achieved through system testing of the prototype.

3.3.3 Experiments

For our main experiments, the conceptual process model outlined in section 2.23 was followed. First, comparative opinion datasets were obtained from Kaggle.com, which hosts free datasets for machine learning and data science projects. Each record in the dataset had two brands compared hence meeting the criteria for comparative opinion data. The features were extracted using a specific ML or DL algorithm, depending on the experiment as there were several algorithms to work with. Based on the algorithm and features extracted, the algorithm was trained on the training set of data, which consisted of 70% of the whole dataset. During training, the algorithm learned from the features in the labelled data on how to associate the output labels (opinion polarity) with the input labels (comparative sentences / reviews).

The ML or DL algorithm simply learns the patterns in the data, which involves detecting how specific comparative elements relate with opinion classes (positive or negative or neutral). The trained model was then subjected to the test set (30% of the dataset that was not used during training) to evaluate the performance of the trained model. Upon satisfactory classification of the training set data based on model accuracy, the model was deployed to classify live comparative data from X platform and YouTube product reviews. While carrying COM on live data, span opinions were minimized by ensuring only unique reviews were collected from the same opinion holder using account / profile IDs. The opinion classes (positive, negative, neutral) were then used to compute the reputation of the brand as a percentage. In this case, if the percentage of positive reviews exceeded those of the negative reviews with regards to a certain brand, then this was treated as positive brand reputation. Otherwise, the

reverse could mean a negative brand reputation. Where the percentages of both positive and negative reviews matched, then the reputation was treated as neutral.

The experiments in this study involved implementing each machine-learning algorithm with each feature extraction technique and each n-gram range or window size as shown in section 3.3.6 and table 3.1. This yielded many results showing how each algorithm performs with a variation from two factors: feature extraction technique and n-gram range or window-size. The n-gram range affects the performance of algorithms when using sparse vectors like in the case of Count Vectors and TFIDF feature extraction techniques while the window-size affected dense vectors used in the case of CBOW and Skip gram feature extraction techniques. Upon determining the best performing single machine learning techniques or algorithms, the study applied the algorithm selection criteria in section 2.12 to decide on which algorithms were fit for application in developing a hybrid model. This resulted in multiple hybrid machine-learning models. The researcher then evaluated the performance of the hybrid models using accuracy and f1-score metrics (see Section 2.15) to determine the most effective one.

Forty experiments were done to establish best combinations of machine learning algorithms, feature extraction techniques, and n-gram ranges for developing an effective hybrid model for application in comparative opinion mining for monitoring brand reputation. The experiments indicating a range in the SN column are those repeated but each time applying a different dataset. The researcher had three secondary datasets (see Section 3.3.4 Data Collection).

In each experiment, the researcher experimented with a set of machine learning algorithms to perform comparative opinion mining on three comparative datasets. This was to determine the accuracy of each algorithm in classifying comparative opinion data but with the effect of feature extraction techniques. The n-gram or window-size parameters were used to establish their impact on the performance of the various algorithms tested. The goal for these experiment was to establish the best or optimal combinations of machine learning algorithms, feature extraction techniques, and n-gram (window-size) parameter combinations that would be used to create an effective hybrid machine learning model for comparative opinion mining, which would find application in brand reputation monitoring. The experiments followed the conceptual process model in Figure 2.11.

The experiments were grouped into two: experiments involving single machine learning models and experiments involving hybrid machine learning models. In both cases, one-way ANOVA and T-tests were used to determine the significant difference in their performance.

Table 3. 1 List of Experiments

SN	***Algorithm***	***Feature Extraction Technique***	***n-gram range value or window size***	***Dataset***	***Model Evaluation Metric***
1 - 3	Single ML Algorithms: MNB, SVM, KNN, DT, RF, LR, SGD, MLP	Count Vectorizer	N = 1	D1: Microsoft vs. Google	▪ Accuracy ▪ F1-score
4 - 6	Single ML Algorithms: MNB, SVM, KNN, DT, RF, LR, SGD, MLP	Count Vectorizer	N = 2	D1: Microsoft vs. Google	▪ Accuracy ▪ F1-score
7 - 9	Single ML Algorithms: MNB, SVM, KNN, DT, RF, LR, SGD, MLP	Count Vectorizer	N = 3	D1: Microsoft vs. Google	▪ Accuracy ▪ F1-score
10 - 12	Single ML Algorithms: MNB, SVM, KNN, DT, RF, LR, SGD, MLP	TFIDF	N = 1	D1: Microsoft vs. Google	▪ Accuracy F1-score
13 - 15	Single ML Algorithms: MNB, SVM, KNN, DT, RF, LR, SGD, MLP	TFIDF	N = 2	D1: Microsoft vs. Google	▪ Accuracy F1-score
16 - 18	Single ML Algorithms: MNB, SVM, KNN, DT, RF, LR, SGD, MLP	TFIDF	N = 3	D1: Microsoft vs. Google	▪ Accuracy F1-score
19 - 21	Single ML Algorithms: MNB, SVM, KNN, DT, RF, LR, SGD, MLP	CBOW	W = 1	D1: Microsoft vs. Google	Accuracy F1-score
22 - 24	Single ML Algorithms: MNB, SVM, KNN, DT, RF, LR, SGD, MLP	CBOW	W = 5	D1: Microsoft vs. Google	Accuracy F1-score
25 - 27	Single ML Algorithms: MNB, SVM, KNN, DT, RF, LR, SGD, MLP	Skip gram	W = 1	D1: Microsoft vs. Google	Accuracy F1-score
28 - 30	Single ML Algorithms: MNB, SVM, KNN, DT, RF, LR, SGD, MLP	Skip gram	W = 5	D1: Microsoft vs. Google	Accuracy F1-score

31 - 33		Hybrid ML Algorithms: DT, MLP + RF, MLP + SGD, MLP + SVM, SGD + DT, SGD + RF, SGD + MLP, SGD + SVM	CV	N = 3	D1: Microsoft vs. Google	Accuracy F1-score
34 - 37		Hybrid ML Algorithms: DT, MLP + RF, MLP + SGD, MLP + SVM, SGD + DT, SGD + RF, SGD + MLP, SGD + SVM	TFIDF	N = 3	D1: Microsoft vs. Google	Accuracy F1-score
38		Hybrid ML Algorithms: DT, MLP + RF, MLP + SGD, MLP + SVM, SGD + DT, SGD + RF, SGD + MLP, SGD + SVM	CV	N = 3	Primary Dataset 1:	Accuracy F1-score
39		Hybrid ML Algorithms: DT, MLP + RF, MLP + SGD, MLP + SVM, SGD + DT, SGD + RF, SGD + MLP, SGD + SVM	CV	N = 3	Primary Dataset 2:	Accuracy F1-score
40		Hybrid ML Algorithms: DT, MLP + RF, MLP + SGD, MLP + SVM, SGD + DT, SGD + RF, SGD + MLP, SGD + SVM	CV	N = 3	Primary Dataset 3:	Accuracy F1-score

3.3.4 Data Collection

This section describes the process of data collection followed in this study. This includes the information about the Ground Truth underpinning the experiments. Additionally, information about Secondary Data and Primary Data used in the experiments is provided.

3.3.4.1 Ground Truth

In supervised ML, the use of human-annotated datasets provides the ground truth from which the machine-learning model is trained and the accuracy of the classification benchmarked or evaluated. Through purposeful sampling, three human annotators were identified and involved in performing manual verification of the sentiment polarity classes assigned to the datasets used for training the models. They were from three local brands in charge of corporate relationships with experience in brand reputation management on digital platforms. To determine the final sentiment polarity for each review, the researcher employed the commonest sentiment polarity assigned to each review by the human annotators. This w akin to the concept of majority votes as used in the parlance of elections. The inter-annotator agreement level was 82.7% with a Kappa (K) score of 0.81, which is satisfactory and indicative of a strong agreement among the annotators (McHugh, 2012).

SN	Reviews	Polarity
1	Android is innovation apple ios is just going downhill	pos_neg
2	Twitter only showed search results for the last seven days but Facebook stores results for long time	neu_pos
3	Android is way better than iOS	pos_neg
4	Android wins hands downï»¿ iOS losses	pos_neg
5	google is more effective then microsoft	pos_neg
6	There is deeper user engagement level in android as compared to iOS	pos_neu
7	iOS has becoming much more visual in its appearance than android	pos_neu
8	One of the most obvious differences between android and iOS is the allotted character count	neu_neu
9	Facebook wins this investing match up despite Twitter cheaper valuation	pos_neg
10	It is really true that microsoft is better than google	pos_neg

Figure 3. 4 Sample Human Expert Annotated Dataset (Screenshot)

After model training, the researcher carried out an additional ground truth test to perform additional human-validation of the hybrid model's performance with using a primary dataset of 100 reviews collected from X platform. In this ground truth experiment, three experts were tasked with classifying the 100 comparative reviews. This number of reviews was arrived at using purposive sampling (Kothari, 2015), considering the busy schedules of the experts and the fact that the reviews were retrieved from the first 100 records from what was collected, thus eliminating selection bias. They assigned a sentiment label to each entity mentioned in the review before declaring the preferred entity in the whole review. As such, each review had six classes –three classes for each entity. The reviews were comparisons between two entities. Finally, the experts then counted the number of times each entity was preferred in the 100 reviews to determine the preferred entity overall. Then, the trained hybrid machine learning model was also fed with the same 100 reviews to carry out comparative opinion mining on (i.e. classify the reviews).

The inter-annotator agreement level was 81.0, which is still satisfactory as the Kappa (K) score was 0.81, indicative of a strong agreement among the human experts on the sentiment labels assigned to each review. The accuracy of the predicted outcomes from the hybrid machine-learning model was compared with the ground truth for additional model validation of the accuracy of the hybrid model.

3.3.4.2 Primary Data

For purposes of model training, three datasets were collected from Kaggle, which is a credible source of datasets for ML and data science projects. Only comparative opinion datasets with two comparable brands per record were selected to align with the purpose of this study, which was comparative opinion mining for purposes of brand reputation

monitoring. The datasets (in .csv format) were given to three experts. Their task was to annotate the comparative texts according to opinion polarities (positive, negative, or neutral) and then determine the preferred entity for each record (tweet or review). This was used as the ground truth in training the model and validating the accuracy of the model. The results obtained from the performance of the trained machine learning models were used to validate the effectiveness of the models and determine the significance in the difference exhibited in the performance of the models. This data was also stored in .csv files and fed into Jamovi GUI platform for statistical analysis based on R programming language to perform ANOVA and T-tests to test our hypothesis.

Table 3. 2 Secondary Data

Dataset	Reviews	pos_pos	pos_neg	pos_neu	neg_neg	neg_pos	neg_neu	neu_neu	neu_pos	neu_neg
Microsoft vs Google	3011	360	1268	396	62	380	46	321	148	30
Facebook vs Twitter	3000	440	1208	447	54	307	59	310	143	32
Pearl Continental vs Marriott	1012	276	138	46	92	138	46	138	92	46

The datasets in Table 3.2 were downloaded at the link below and annotated by human experts: *https://www.kaggle.com/umairyounis/comparative-reviews-datasets*.

Secondary datasets from curated collections like Kaggle are reliable and suitable for machine learning and data science projects as they are cheaper, easier to collect, credible, permit for benchmarking and are readily available in high quality (Martins et al., 2018; Kaggle, 2021; Gonzalez et al., 2023). In the case of machine learning projects, curated datasets are refined for use in specific tasks. Each of the datasets had only two entities per record, to ease computational complexity associated with handling a high number of opinion classes in the cases of three or more comparative entities. Each dataset had comparative reviews as well as their corresponding sentiment labels corresponding to the two brands being compared.

Based on Table 3.2, the “pos” label means positive opinion class. The “neg” label means “negative” opinion class. The “neut” label means neutral opinion class. Since for every review there were to entities (brands) being compared, the corresponding opinion classes were used in a binary fashion. For example, pos_pos means that the review had a positive opinion class for both entities. Similarly, where pos_neg was applied, then the first entity had a positive opinion while the second entity had a negative opinion. This is the logic applied in interpreting the labels provided in the table. The three datasets presented in Table 3.1 were from three different domains: search engines, social media, and hospitality. This was to ensure that the developed model could work across knowledge multiple domains. There were two other sources of data in this study:

i. Data collected by the system prototype for model validating the effectiveness of the hybrid model and the system prototype implementing the hybrid model. This was collected from X platform and YouTube. The reviews were automatically extracted and fed into the hybrid model. The classification results were used to monitor brand reputation. A sample of this data is in Appendix III. A key challenge the researcher faced in collecting primary data was the size and quality of data. Obtaining comparative opinion data for different brands takes time, effort, data annotation skills, and financial resources. Moreover, X platform developer API for extracting tweets stopped working at the time of data extraction, making it impossible to collect more data without incurring extra costs. This resulted in switching to YouTube to harvest a few comments on some brands for purpose of this study. This provided a useful alternative source of data (Peters et al., 2020). since YouTube comments / reviews were freely accessible and had useful reviews. Overall, the

quality of the data was poorer than that of the secondary datasets, which predominantly relied on product reviews. Product reviews are better because the names of the brands and products are standardized.

ii. According to Kothari (2015), data obtained from experiments as primary data. Thus, the second source of primary data was the raw data collected during experiments. In each experiment, the researcher tested different algorithms on different datasets using variations of different feature extraction techniques and n-gram / window-size parameters. This data was used to evaluate the effectiveness of the models using statistical tests including ANOVA and T-test. This data is reported in Chapter 5 and was obtained from Twitter and YouTube.

Table 3. 3 Primary Data Details

SN	*Dataset Name (Comparative Brands)*	*Domain*	*No. of Records*
1	iPhone vs Samsung	Phone Gadgets	370
2	Nissan Patrol vs Toyota Land cruiser	Automobiles (Cars)	556
3	HP vs Dell	Computers (Technology)	60
4	Raila Odinga vs William Ruto	Politics	1000

The Experimental Process

For each experiment in Table 3.2, the steps 1 to 5 described below were followed because the process is the same for each experiment. The only variables in our experiments were the machine learning algorithm, feature extraction technique, n-gram range (or window size) parameter, and dataset. For purposes of understanding the description of experiments in this chapter, the following abbreviations were used.

- CBOW – Continuous Bag of Words
- CV – Count Vectorizer
- DT – Decision Tree
- KNN – K-Nearest Neighbors
- LR – Logistic Regression
- MLP – Multilayer Perceptron
- MLP+DT = Hybrid Model consisting of MLP base model and DT top-level model
- MLP+RF = Hybrid Model consisting of MLP base model and RF top-level model
- MLP+SGD = Hybrid Model consisting of MLP base model and SGD top-level model
- MLP+SVM = Hybrid Model consisting of MLP base model and SVM top-level model
- MNB – Multinomial Naïve Bayes
- RF – Random Forest
- SGD – Stochastic Gradient Descent
- SGD+DT = Hybrid Model consisting of SGD base model and DT top-level model

- SGD+MLP = Hybrid Model consisting of SGD base model and MLP top-level model
- SGD+RF = Hybrid Model consisting of SGD base model and RF top-level model
- SGD+SVM = Hybrid Model consisting of SGD base model and SVM top-level model
- SVM – Support Vector Machine
- TFIDF – Term Frequency Inverse Document Frequency

Step 1: Dataset Collection

The researcher used the datasets described in Table 3.2.

Step 2: Training Data

For model training, three human experts participated in annotating the datasets. Their task was to assign each record / review an opinion class (positive, negative, neutral) in relation to the entity referenced. The annotators agreed in 5808 records of the 7023 records. This is equivalent to a Kappa (K) score of 0.81, which implies a strong agreement among the annotators.

Step 3: Testing Set

In machine learning, testing sets perform the role of a benchmark during model testing. The testing set is applied to the model after the model has been completely training on the training data to establish if the model is working well or not. In this study, a random split method was adopted to split each dataset into 80% training set and 20% testing set. The researcher preferred this method since it produces more precisely partitioned sets. The resulting datasets were stored in .csv files for use during the experiments as

the machine learning models used could easily read .csv file formats from different development environments like Juypter Notebook.

Step 4: Dataset Preprocessing

Datasets collected are not fit to be fed directly into a machine-learning algorithm for training before such data is cleaned up of the inherent noise. Data cleaning helps achieve data consistency. The following tasks were performed during data cleaning:

1. Removal of Unimportant Characters: the columns in the datasets contained white spaces. These were removed to ascertain uniformity and ease of replication.

2. Unimportant characters including special characters were removed. Moreover, parts of speech (POS) tagging was done to associate each word with the part of speech it belongs. Other tasks involved in this process were:
   - *Tokenization*: this task involved using the NLTK Tokenizer in Python to help break down the words in the text small chunks known as tokens.
   - *Stop Word Elimination*: this process involves removing non-opinion words and characters, which are known as stop words. The researcher used a Python script contained predefined stop words like "the", and "is", which carry no opinion.

Step 5: Apply Machine Learning Algorithms

The researcher experimented first with single ML algorithms followed by hybrid ML algorithms. The single machine learning algorithms were MNB, SVM, KNN, DT, RF, LR, SGD, and MLP. The hybrid machine-learning models were MLP+DT, MLP+RF, MLP+SGD, MLP+SVM, SGD+DT, SGD+RF, SGD+MLP, SGD+SVM. In the case of the first four-hybrid ML above, the researcher had the MLP algorithm as the base

estimator while the other algorithms were each used as the final estimator (top-level estimator). In this phase, the partitioned datasets consisting of both the training and test sets were used. Already, the review had been annotated with the correct classes (positive, negative, or neutral) for each entity per row. The feature extraction models like CV, TFIDF, CBOW, and Skip gram, which are described in Section 2.14, were used to extract features from the data, which were input to the algorithm. The algorithm learned the patterns to help it make predictions on unknown data. The algorithm finds patterns by associating the inputs with the outputs (labels). A generic diagram for this kind of supervised machine learning technique is shown in Figure 3.5.

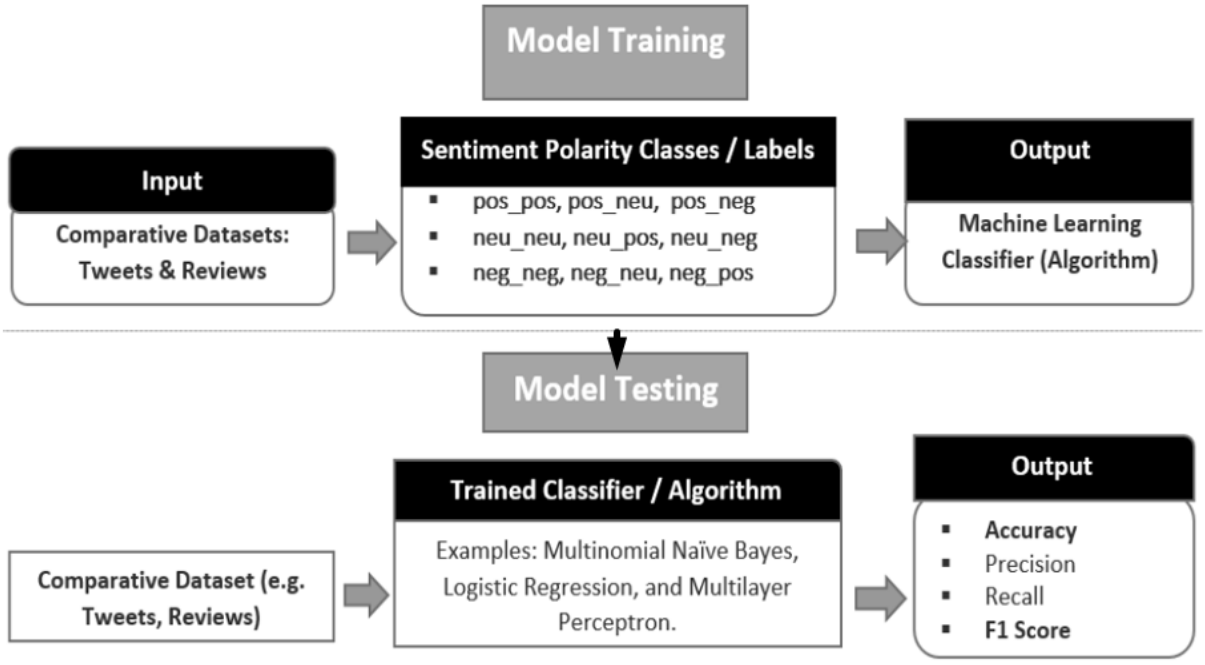

Figure 3. 5 ML-based Process for COM (Ondara et al., 2023).

3.4 Research Instruments

For this study, experimental checklists were used as a research instrument to ensure that all necessary variables were tested for purposes of obtaining comprehensive results. The checklists shown in Table 3.4 and Table 3.5 were applied three times, to account for the three datasets used in our experiments. The checklists were labelled according to the dataset used.

In the checklist in Table 3.4., the researcher employed a single machine learning algorithm such as Random Forest to select features from one of the datasets (e.g. Dataset 1) using one of the feature selection techniques (e.g. Count Vectorizer) with a parameter of, say n-gram value of 1). This was recorded. This experiment was then repeated for different values on n-gram ranges. Then, the experiment was repeated using the same algorithm but with a different feature selection technique (e.g. TFIDF) with variations of the n-gram range. The table indicates what values were changed in each experiment. The primary goal of using the checklist was to ensure that the researcher tests different combinations of algorithm, feature selection technique, and parameter for feature extraction based on a common dataset to establish the most optimal combinations that would be used in developing a hybrid model for COM.

Table 3. 4 Experimental Checklist for Single Classification Models

SN	Classification Algorithm	Feature Selection Technique	N-Gram Range (Window-Size)	Status (Done = ☑)	Comments
1	Multinomial Naïve Bayes (MNB)	CV	N = 1	☐	
		CV	N = 2	☐	
		CV	N = 3	☐	
		TFIDF	N = 1	☐	
		TFIDF	N = 2	☐	
		TFIDF	N = 3	☐	
		CBOW	W = 1	☐	
		CBOW	W = 5	☐	
		Skip Gram	W = 1	☐	
		Skip Gram	W = 5	☐	
2	Support Vector Machine (SVM)	CV	N = 1	☐	
		CV	N = 2	☐	
		CV	N = 3	☐	
		TFIDF	N = 1	☐	
		TFIDF	N = 2	☐	
		TFIDF	N = 3	☐	
		CBOW	W = 1	☐	
		CBOW	W = 5	☐	
		Skip Gram	W = 1	☐	
		Skip Gram	W = 5	☐	
3	K-Nearest Neighbor (KNN)	CV	N = 1	☐	
		CV	N = 2	☐	
		CV	N = 3	☐	
		TFIDF	N = 1	☐	
		TFIDF	N = 2	☐	
		TFIDF	N = 3	☐	
		CBOW	W = 1	☐	
		CBOW	W = 5	☐	
		Skip Gram	W = 1	☐	
		Skip Gram	W = 5	☐	

4	Decision Tree (DT)	CV	N = 1	☐
		CV	N = 2	☐
		CV	N = 3	☐
		TFIDF	N = 1	☐
		TFIDF	N = 2	☐
		TFIDF	N = 3	☐
		CBOW	W = 1	☐
		CBOW	W = 5	☐
		Skip Gram	W = 1	☐
		Skip Gram	W = 5	☐
5	Multi-layer	CV	N = 1	☐
	Perceptron (MLP)	CV	N = 2	☐
		CV	N = 3	☐
		TFIDF	N = 1	☐
		TFIDF	N = 2	☐
		TFIDF	N = 3	☐
		CBOW	W = 1	☐
		CBOW	W = 5	☐
		Skip Gram	W = 1	☐
		Skip Gram	W = 5	☐
6	Random Forest (RF)	CV	N = 1	☐
		CV	N = 2	☐
		CV	N = 3	☐
		TFIDF	N = 1	☐
		TFIDF	N = 2	☐
		TFIDF	N = 3	☐
		CBOW	W = 1	☐
		CBOW	W = 5	☐
		Skip Gram	W = 1	☐
		Skip Gram	W = 5	☐

7	Logistic Regression	CV	N = 1	☐
		CV	N = 2	☐
		CV	N = 3	☐
		TFIDF	N = 1	☐
		TFIDF	N = 2	☐
		TFIDF	N = 3	☐
		CBOW	W = 1	☐
		CBOW	W = 5	☐
		Skip Gram	W = 1	☐
		Skip Gram	W=5	☐
8	Stochastic Gradient Descent (SGD)	CV	N = 1	☐
		CV	N = 2	☐
		CV	N = 3	☐
		TFIDF	N = 1	☐
		TFIDF	N = 2	☐
		TFIDF	N = 3	☐
		CBOW	W = 1	☐
		CBOW	W = 5	☐
		Skip Gram	W = 1	☐
		Skip Gram	W = 5	☐

Table 3. 5 Experimental Checklist for Hybrid Classification Models

SN	Classification Algorithm	Feature Selection Technique	N-Gram Range (Window-Size)	Status (Done = ☑)	Comments
1	MLP + DT	CV	N = 3	☐	
		TFIDF	N = 3	☐	
2	MLP + RF	CV	N = 3	☐	
		TFIDF	N = 3	☐	
3	MLP + SGD	CV	N = 3	☐	
		TFIDF	N = 3	☐	
4	MLP + SVM	CV	N = 3	☐	

		TFIDF	N = 3	☐
5	SGD + DT	CV	N = 3	☐
		TFIDF	N = 3	☐
6	SGD + RF	CV	N = 3	☐
		TFIDF	N = 3	☐
7		CV	N = 3	☐
	SGD + SVM	TFIDF	N = 3	☐
8		CV	N = 3	☐
	SGD + MLP	TFIDF	N = 3	☐

3.5 Pilot Study

3.5.1 Sample Selection Strategy

Many machine learning algorithms and deep learning algorithms can be used in carrying comparative opinion mining. For each algorithm selected, different feature extraction techniques could be used to extract features that are fed to algorithm for model training. Feature selection is a technique used in dimensionality reduction in COM (Anjaria & Guddeti, 2014) so that machine learning or deep learning algorithms can work better. For each feature extraction technique, there may be different n-gram ranges or window sizes to determine specific aspects guiding feature selection. Furthermore, different datasets could be used. To simplify all of this, a pilot study was necessitated, purely, to check the feasibility of the main experiments but using a small sample from our population of study. Consequently, the following pilot checklist of experiments was developed to assist in handling the pilot study in a structured and consistent manner. Feature selection achieves dimensionality reduction, which reduces data complexity for enhanced opinion analysis (Anjaria & Guddeti, 2014).

From a population of eight classification algorithms, the researcher piloted with two algorithms, representing a 25% sample. The researcher then piloted our study with all feature selection techniques, representing a 100% sample based on the population. For N-gram ranges and Window Sizes, the researcher piloted with a range of 1 (33.3% sample size) and window size of 1 (50% sample size). The selected algorithms, feature extraction techniques, and n-gram range or window sizes were representative enough to establish the workability of the experiments. Since the researcher already had three datasets, the researcher piloted with one of the datasets, representing a sample size of 33.3%). All these samples selected for piloting followed a purposive sampling strategy since the population was small in size (Kothari, 2015).

Table 3. 6 Experimental Checklist for Pilot Study

SN	Classification Algorithm	Feature Selection Technique	N-Gram Range (Window-Size)	Status (Done = ☑)	Comments
1	DT (ML Classification Algorithm)	o CV	▪ N = 1	☐	
		o TFIDF	▪ N = 1	☐	
		o CBOW	▪ W = 1	☐	
		o Skip Gram	▪ W = 1	☐	
2	MLP (Deep Learning Classification Algorithm)	o CV	▪ N = 1	☐	
		o TFIDF	▪ N = 1	☐	
		o CBOW	▪ W = 1	☐	
		o Skip Gram	▪ W = 1	☐	

3	MLP + DT (Hybrid Classification Algorithm)	○	CV	▪	N = 1	☐
		○	TFIDF	▪	N = 1	☐
		○	CBOW	▪	W = 1	☐
		○	Skip Gram	▪	W = 1	☐

3.5.2 Observations from the Pilot Study

While conducting the pilot study, the following observations were made:

- Working with Jupyter Notebook on a personal computer worked fine for simple experiments especially those involving DT algorithm. However, when implementing MLP, a DL algorithm, it was observed that the model took very long to train and make predictions. This is because DL models require specialized computational resources like graphical processing units (GPU) and/or Tensor Processing Units (TPU), which were not available in my computing environment on Jupyter Notebook that was running on my personal laptop computer as a research tool or resource for model development. This informed the decision to seek a cloud-based computing environment when carrying out the main experiments.

- The researcher also observed that developing a hybrid classification model using MLP and DT could take two approaches:

i. DT as the base model and MLP as final model – in this case, DT could perform feature extraction while MLP performs opinion classification.

ii. MLP as the base model – in this case, MLP performs feature extraction while DT performs opinion classification.

However, considering the benefits of deep learning models especially in the task of feature selection where they outperform traditional machine learning models, it was decided that during the main experiments, a deep learning model would be a good fit for the base model in developing a hybrid model for COM (Varathan et al., 2017).

3.6 Sampling Strategy and Sample Size

3.6.1 Algorithm Selection Strategy

From a systematic literature review conducted by Ondara et al. (2022), eight ML algorithms were found to be the most popular in direct opinion mining research. Popular algorithms have proven track records in various applications in different domains, which confirms their reliability (Kourentzes et al., 2020). Such algorithms are often used as benchmarks in research to allow for comparative analysis of their empirical performance (Hodge & Austin, 2019) before their application in a project. Finally, popular algorithms have well-supported frameworks and libraries, making them much easier to optimize and implement (Zhang et al., 2019).

In a different study by Younis and others (2022), seven ML algorithms were used to carry out comparative opinion mining. By finding the common algorithms in both direct opinion mining and comparative opinion mining, the researcher ended up with six algorithms. These were Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), and Random Forest (RF). The Artificial Neural Network (ANN) algorithm was used in direct opinion mining only. The research used a purposive sampling strategy to select MLP for use in our experiments, being a variant of the ANN. This is because MLP outperforms basic ANN especially because of the benefits of multiple hidden layers used in MLP. Besides, MLP integrates much more easily with ML algorithms like RF (Bengio et al., 2013).

Gradient Boosting (GB) algorithm was used only in Comparative Opinion Mining study. The researcher selected its variant, Stochastic Gradient Descent (SGD) as it is easier to implement and has better performance than GB (Mikolov, Chen, Corrado, & Dean, 2013). Overall, purposive sampling strategy, informed by expertise and knowledge (Guest et al., 2006), theoretical analysis, diversity of perspective (Morse, 2015) and pilot study was adopted in selecting the sample given the fairly small number of algorithms (Kothari, 2015). Thus, from a population of 15 algorithms, the sample size was eight algorithms. This sample size represents 53% of the population.

Table 3. 7 Sampling of Classification Algorithms

DOM Algorithms	COM Algorithms	Common Algorithms	Deep Learning Algorithms	Sample Size
▪ NB	▪ NB	▪ NB	▪ -	▪ NB
▪ SVM	▪ SVM	▪ SVM	▪ -	▪ SVM
▪ KNN	▪ KNN	▪ KNN	▪ -	▪ KNN
▪ DT	▪ DT	▪ DT	▪ -	▪ DT
▪ RF	▪ RF	▪ RF	▪ -	▪ RF
▪ LR	▪ LR	▪ LR	▪ -	▪ LR
▪ ANN	▪ -	▪ -	▪ -	▪ MLP (an implementation of ANN)
▪ ME	▪ GB	▪ -	▪ ANN	▪ SGD (an implementation of GB)

3.6.2 Feature Extraction Techniques Selection Strategy

From theoretical analysis, and pilot study, and the researcher's expert judgement (Guest et al., 2006), and diversity of perspective (Morse, 2015), a few feature extraction techniques were identified and selected using purposive sampling strategy, given their relatively small number (Kothari, 2015). Feature extraction techniques like Word2Vec,

Glove, and BERT were not popular; hence, we dropped from the original population of seven techniques. The result was four techniques, representing 57% of the population.

Table 3. 8 Sampling of Feature Extraction Techniques

Sparse Vectors	Dense Vectors	Population
▪ Count Vectorizer (CV) ▪ Term Frequency Inverse Document Frequency (TFIDF)	▪ CBOW ▪ Skip Gram	▪ CV ▪ TFIDF ▪ CBOW ▪ Skip Gram

3.6.3 Datasets Selection Strategy

Based on theoretical analysis, diversity of perspective (Morse, 2015) expert judgement (Guest et al., 2006), and pilot study, purposive sampling strategy was adopted in the selection of datasets for use in this study. Two key factors were that the dataset had to be comparative (has two brand entities compared in each record) and be in English language. While the researcher was able to find freely published comparative datasets on credible platforms like Kaggle, the number was small (Kothari, 2015) and many of them had imbalanced classes. The researcher selected the three of four datasets (75% sample size) as shown in Table 3.9. The iOS vs Android dataset was removed because it was small and hence subject to overfitting.

Table 3. 9 Sampling of Comparative Opinion Datasets

Dataset	Dataset Domain	No. of Records
▪ Microsoft vs. Google	○ Search Engines	3011
▪ Facebook vs. Twitter (X)	○ Social Media	3000
▪ Pearl Continental vs. Marriott	○ Hospitality	1012

3.6.4 N-Gram Range and Window Size Selection

Based on expert judgement (Guest et al., 2006), literature analysis, diversity of perspective (Morse, 2015), pilot study and the relatively small number of values for n-gram range and window-size (Kothari, 2015), the researcher adopted purposive sampling strategy. The value of *n* in the *n-gram* model affects the performance of a classification model. This research shows that beyond n = 5, the performance of a classification model remains constant or in some cases starts dropping. Therefore, our study used a purposive sampling strategy to select n-gram ranges of 1 to 3 for experimentation, representing a sample size of 60%. For the window size parameter when working with other models, the researcher used 5, representing a sample size of 20%. This decision considered the complexities involved when working with CBOW and Skip Gram models. Dense vectors such as CBOW face the challenge of data sparsity and reduced accuracies particularly where data is not adequate.

Table 3. 10 Sampling of N-gram and Window Size Parameters in Feature Extraction

Feature Extraction Technique	*N-gram Range (or Window Size) Parameter*	*Sample*
CV or TFIDF	N = 1, N= 2, N=3, N=4, N=5	N=1, N=2, N=3
CBOW	W=1, W=5	W=5
Skip Gram	W=1, W=5	W=5

From existing research, the value on n in the n-gram model affects the performance of a classification model in opinion mining. This research shows that beyond n=5, the performance of a classification model remains constant and in some cases starts dropping. For this reason, our study used a purposive sampling strategy and selected the n-gram ranges of 1 to 3 for experimentation, representing a sample size of 60%. For

the window size parameter when working with the models, the researcher tested values of 1 to 5 and determined that 5 was good fit. The researcher therefore deliberatively chose to work with 5, representing a sample size of 20%. This decision was arrived at considering the complexities involved when working with CBOW and Skip Gram feature extraction techniques. Dense vectors such as CBOW face the challenge of data sparsity and often have reduced accuracies in some applications where data is not adequate for extracting adequate representation of features.

3.6.5 Algorithm Performance Evaluation Metrics Selection

Literature analysis revealed that multiple algorithm evaluation metrics exist. Based on expert judgement (Guest et al., 2006), pilot study, diversity of perspective (Morse, 2015), theoretical analysis, and coupled with the relative small number of metrics to select from (Kothari, 2015) for purposes of COM, the researcher adopted the purposive sampling strategy. However, there are many metrics in opinion mining as described in section 2.15. Literature analysis showed that two of them (accuracy and F-score) are predominantly used to report the performance of classification models. Accuracy was selected because it is the most widely accepted measure, while f-score is recommended for imbalanced datasets. This represents 50% of the population.

Table 3. 11 Sampling of Model Evaluation Metrics

Population (Model Evaluation Technique) = 4	Sample Size = 2
Accuracy	Accuracy
Precision	F1-score
Recall	
F-score	

3.7 Research Hypothesis Testing Design

This research had one hypothesis, which was mapped from a research objective:

H_0: There is no significant difference between the performance of the hybrid machine-learning model and the individual machine learning models.

H_A: There is a significant difference between the performance of the hybrid machine-learning model and the individual machine learning models.

The collected data was analyzed using appropriate statistical techniques (in this case, ANOVA and T-Test), which helped to decide on whether to accept the null hypothesis or reject it. Depending on the analyzed data, the researcher drew conclusions regarding the research hypothesis. The conclusion was based on the results. Where the results clearly supported the hypothesis, the researcher accepted the hypothesis. Where the results did not support the hypothesis, the researcher rejecting the hypothesis.

To prove this hypothesis, statistical tests were conducted using one-way ANOVA and T-test. The researcher used Jamovi, a GUI application for statistical research using R programming language. This research was anchored on determining a good combination of machine learning algorithms and deep learning algorithms for creating a hybrid machine-learning model that would be effective in performing comparative opinion mining. The core hypothesis, therefore, was about testing the effectiveness of the hybrid model through its performance in terms of opinion classification accuracy. To carry out statistical tests on this hypothesis, two statistical tests were used: Analysis of Variance (ANOVA) and T-tests.

ANOVA is used when a researcher is interested in comparing the averages of at least three groups (Raschka & Mirjalili, 2019). As is with the T-tests, ANOVA requires a categorical independent variable and a continuous dependent variable (Montgomery, 2017). But, with ANOVA, the independent variable needs at least two levels of treatment. In this study, a One-Way ANOVA was used to establish the difference in the accuracies of different ML models across various datasets and feature extraction techniques. It was used to find the variance in the accuracy of different models so that the researcher could establish if there was any significant statistical difference in how the various models performed across different datasets and with various feature extraction techniques and n-gram or window-size parameters. T-tests are applicable when one needs to compare the averages of two different groups, with the assumption that data is normally distributed (Field, 2018; Bland & Altman, 2019).

Normality tests were done, utilizing Shapiro-Wilk Test $p >= 0.5$ for normality non-violation. For instance, this study used T-tests to compare the accuracies of two best performing hybrid ML models with varying configurations of feature extraction techniques or datasets. This statistical measure was used to determine if there was any statistically significant difference between the best two performing hybrid models for comparative opinion mining. This is because the experiments showed that two hybrid models performed exemplary well. There was need to establish if their performance (in terms of accuracy) had a statistically significant difference.

3.8 Model Development

This section describes how the hybrid machine-learning model for comparative opinion mining was developed in the course of this study.

3.8.1 Process Model

The development of the hybrid ML model in this work followed the opinion mining process represented in Figure 3.6. The ensemble learning methodology was used. First, the data source for modeling consisted of online reviews obtained from acquired datasets. After development, the model was tested using primary data extracted from X platform and YouTube.

Step 1: comparative opinion data was collected from the source. During model development, secondary dataset from Kaggle.com was used. Manual verification was done on the datasets by the help of human experts to ascertain that the data had comparative opinions. This was the first step in the opinion mining process. The data used to develop and test this model is described in section 3.3.4.3.

Step 2: this involved extracting features from the comparative opinions using a specific feature extraction model. This involved various extraction techniques, which included Count Vectorizer, TF-IDF Vectorizer, Continuous Bag of Words, Skip grams for machine learning models. Tokenizers were used in the case of DL models.

Step 3: With the features extracted, the opinions in each review or sentence, in the form of vector data, were passed into a machine-learning algorithm like Random Forest or dep learning algorithm like Multilayer Perceptron for classification into opinion polarities, which consisted of positive, negative, or neutral. The overall net opinion

class is the result, which can be used to infer the reputation of a specific brand. In this study, if the total sentence opinion class calculated from all the reviews or sentences was positive, the researcher determined the result as positive reputation for the target brand. This approach was used in cases where the net reviews were negative or neutral.

Step 4: Overall Opinion Score Calculation: based on the distribution of opinion classes across the whole data, an overall opinion class is determined by taking the class with the highest count. Thus, for example, if most records were positive in opinion, then this was taken to imply a positive opinion on average for the entire data that was analyzed. Similarly, if most of the records showed negative opinion classes, the result was that the opinions expressed in the data were negative. This same principle applied for the neutral opinion class.

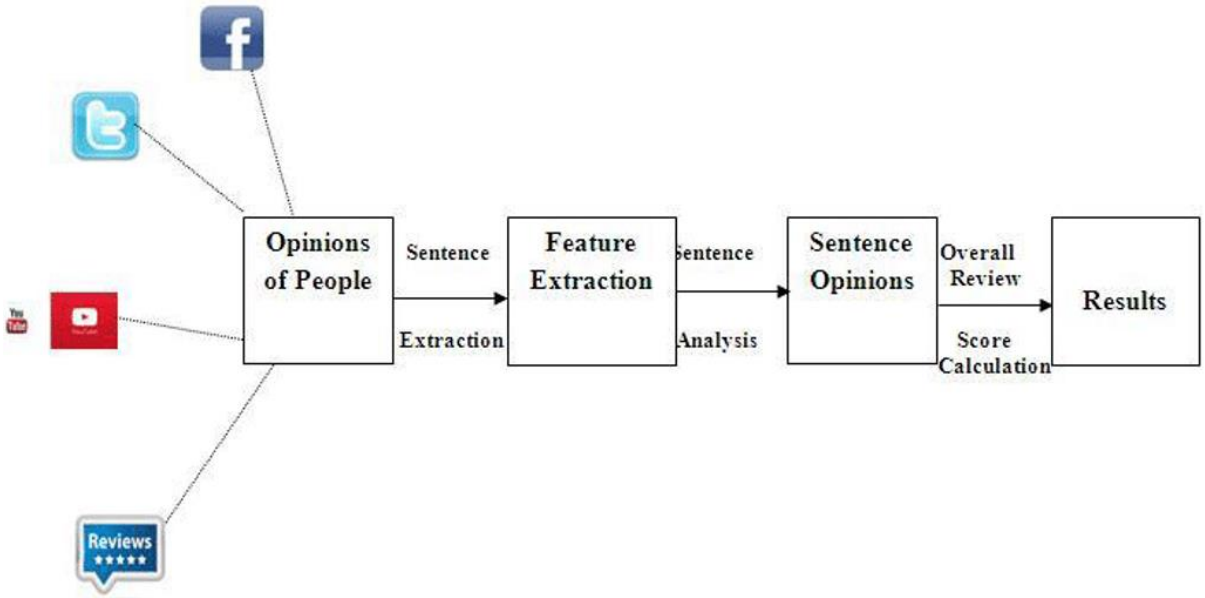

Figure 3. 6 Opinion Mining Process Nimbhore & Siledar, 2014).

This process model aligns with the conceptual model. The opinions of people are captured in comparative sentences, which when extracted and analyzed, the other three elements (entities, entity relations, and features) are inherent. To detect these elements,

the feature extraction is done using a feature extraction technique. The sentence classification is done using a machine-learning model. These elements contribute to overall sentence opinion class and opinion score, which inform the brand reputation class (positive, negative, or neutral).

Since the model developed in this study was a hybrid machine-learning model for comparative opinion mining, the above opinion mining process was followed, except that there were six opinion classes. This was done to ensure each entity mentioned in the opinion sentence or review had the three opinion classes: positive, negative, and neutral. Table 3.12 shows how any two entities compared in a dataset were handled during opinion classification.

Table 3. 12 Multi-label Classification for Comparative Opinion Mining

Entity (Brand) 1	***Entity (Brand) 2***
Positive	Positive
Positive	Neutral
Positive	Negative
Negative	Positive
Negative	Neutral
Negative	Negative
Neutral	Positive
Neutral	Neutral
Neutral	Negative

As noted from the Table 3.12, there are nine possible classes of opinion polarities when analyzing opinions towards two entities mentioned together in each record in a comparative opinion dataset. To obtain a preferred entity between the two, the developed model computes the total count of all positive classes for each entity and

decides upon the entity with a higher positive score (count), otherwise, there is no preferred entity, which implies both entities had equal number of positive polarity opinion classes.

For purposes of applying this model to perform brand reputation monitoring for each entity, the model makes decisions as shown in Table 3.13:

Table 3. 13 Reputation Class Matrix for Comparative Opinion Classification

Positive vs Negative Opinions for each Entity	***Reputation Class***
IF (Positive > Negative) THEN	Positive Reputation
IF (Positive == Negative) THEN	Neutral Reputation
IF (Positive < Negative) THEN	Negative Reputation

For purposes of determining the preferred entity, where two entities are involved, the model makes decisions as shown in Table 3.14:

Table 3. 14 Preferred Entity Matrix for Comparative Opinion Classification

Positive vs Negative Opinions for Entity 1 vs Entity 2	***Reputation Class***
IF (Positive (1) > Positive (2)) THEN	Entity 1 is more Preferred; higher positive score for 1; lower positive score for entity 2
IF (Positive (1) < Positive (2) THEN	Entity 2 is more Preferred; higher positive score for 2; lower positive score for entity 1
IF (Positive (1) == Positive (2) THEN	There is no preferred entity. Both entities have the same reputation (i.e. positive with no preference).

3.8.2 Single ML Model Design for COM

The overall design of the model in this study followed seven stages. The first stage involved extracting data from a data source such as Amazon Reviews or X. in our case; we used a dataset of product reviews from Kaggle.com. Kaggle.com curates datasets for machine learning and data science projects. Section 3.3.4.3 describes the data used in designing the mode. The datasets used contained the comparative opinion elements identified in the conceptual model in Section 2.22. The second stage involved cleaning of the raw data collected from the data source. This included removing opinion spams and business ads. Duplicate opinions (reviews, tweets, comments) from the same user ID were treated as opinion spam as they lacked uniqueness and were aimed at adding opinion bias. For data from X platform, cleaning involved removing Ads. Since Elon Musk took over Twitter and rebranded it to X, the platform started promoting advertisements, making it difficult to directly use tweets collected from the platform for COM. The cleaned data was stored in .csv files on local storage.

The cleaned, stored data was preprocessed for modeling purposes. The modeling phase involves training the model on the features to properly associated entities with features in order to predict the correct opinion classes. Pre-processing identifies key features in the data, using a ML algorithm like Stochastic Gradient Descent or a DL algorithm like Multilayer Perceptron. Continuous deployments and improvements are then done, each time aiming at identifying the best performing model for the COM task.

The continuous improvement phase involved experimenting with different combinations of machine learning or deep learning algorithms utilizing different feature extraction techniques on different datasets. This helped the researcher to identify the best machine learning or deep learning algorithm for comparative opinion mining given

the different feature extraction techniques. The output from this design process led to the identification of the best machine single learning or best single deep learning models for use in developing a hybrid machine-learning model for comparative opinion mining. With the best model selected and improved, it is then used to classify comparative opinion sentences into various classes based on key features like opinion words used, and comparative relationship between entities. This analysis leads to the determination of a brand's reputation class (positive, negative, or neutral).

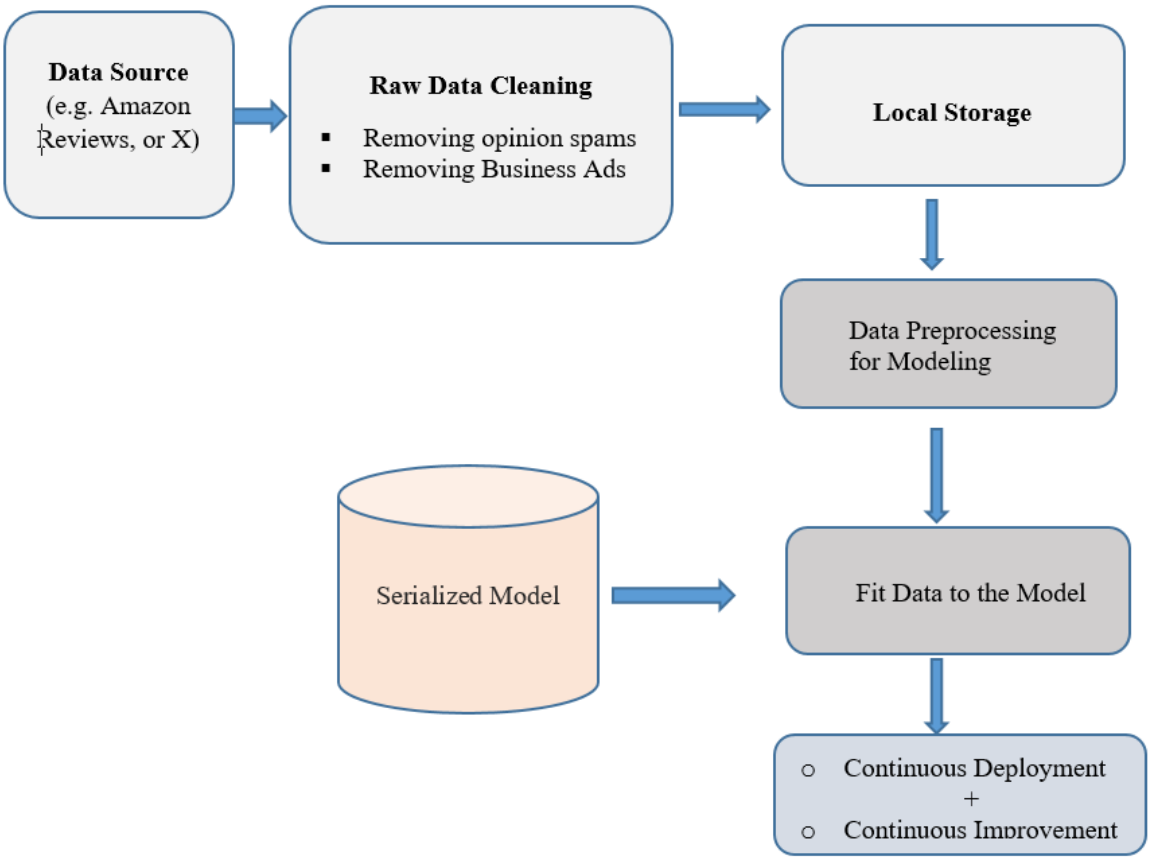

Figure 3. 7 Model Design: Single Machine Learning Model for Comparative Opinion Mining

The serialized model is produced through a process called serialization. This involves converting the ML Model into a special format which enables the developer to transmit or store data about the model and then use this data to recreate the model upon demand through a process called deserialization.

3.8.3 Model Selection Process

The following model selection process was used in the development of the best performing hybrid model for comparative opinion mining, illustrating the case for Deep Learning as the base model.

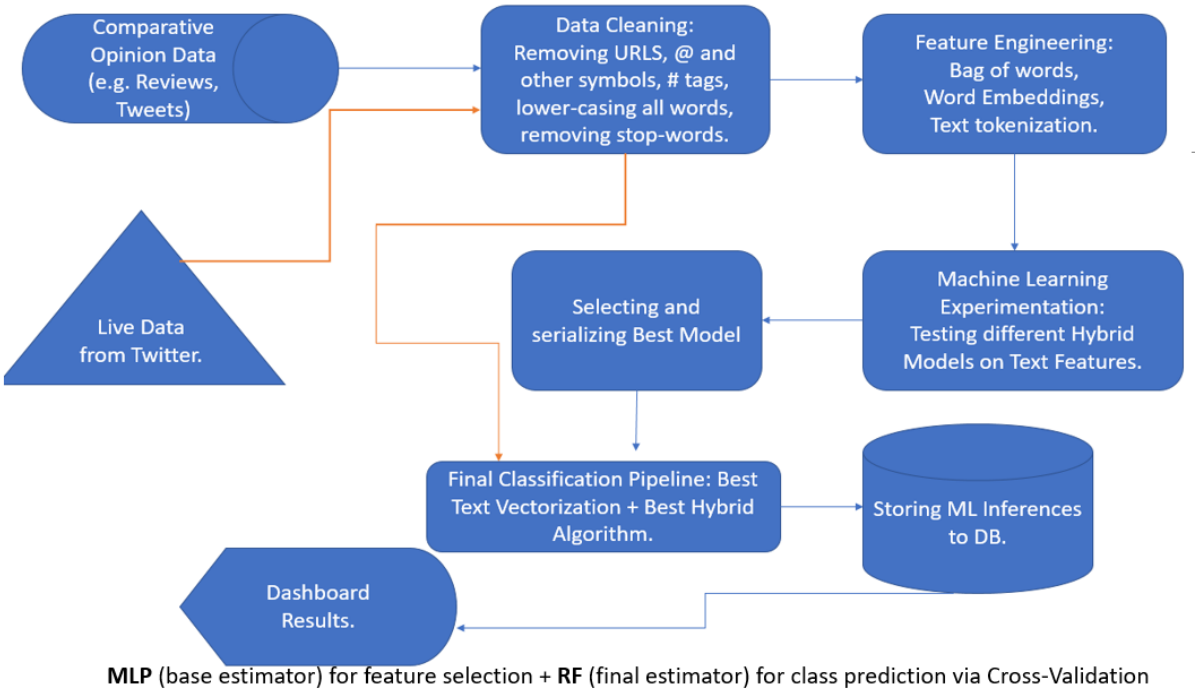

Figure 3. 8 Model Selection Process

3.9 Design of the Hybrid ML Model for COM

3.9.1 Hybrid ML Model Design

First, based on the literature reviewed in section 2.13, the steps for creating an ensemble ML model were established and followed. The first component of the architecture for the hybrid ML model required cleaned data from datasets described in Section 3.3.4.3. The second step was data preprocessing. Text preprocessing was done by removing white spaces, unwanted special characters, stop words, and parts of speech tags. Tokenization, which breaks down a review into small chunks called tokens was also done. For this, the researcher employed the NLTK Tokenizer in Python. The third step

was base model selection. To accomplish this, experimentation on feature extraction techniques using different single ML algorithm (e.g. RF) and DL algorithm (e.g. MLP) was conducted and the selection was based the hybrid model's performance in classifying comparative opinions. A change in the base algorithm that led to improved opinion classification accuracy meant that the new algorithm performed better in feature selection. The fourth step was the ensemble method of learning. In the hybrid design, the base algorithm selects features that generate classification probabilities which are then forwarded to the top-level model for opinion classification. The base model was used to select features while the top-level model was used to classify the opinions according to different classes. This ensemble method involved the use a base learner's outputs as inputs into the final classification algorithm to create the hybrid model as explained in section 2.13.

The fifth step was model training and evaluation. Model training was done on the pre-processed training set, which consisted of 70% of the whole dataset. This was iteratively done to determine the best combination of feature extraction and classification models to create a hybrid model for comparative opinion mining. The best hybrid model was determined through model performance evaluation using accuracy and F1-score metrics. The seventh step was hyper-parameter tuning. The study utilized the default parameters for both the base model and the classification model on the basis that the hybrid models outperformed the single models. The final step was model deployment and monitoring. This was attained by serializing the hybrid model to make it ready for deployment. Upon deployment, the hybrid model was used to monitor brand reputation, while also extracting key brand aspects to inform the brand reputation manager or

officer of the primary aspects captured by the model, forming the subject of comparisons in the comparative opinions analyzed by the hybrid model.

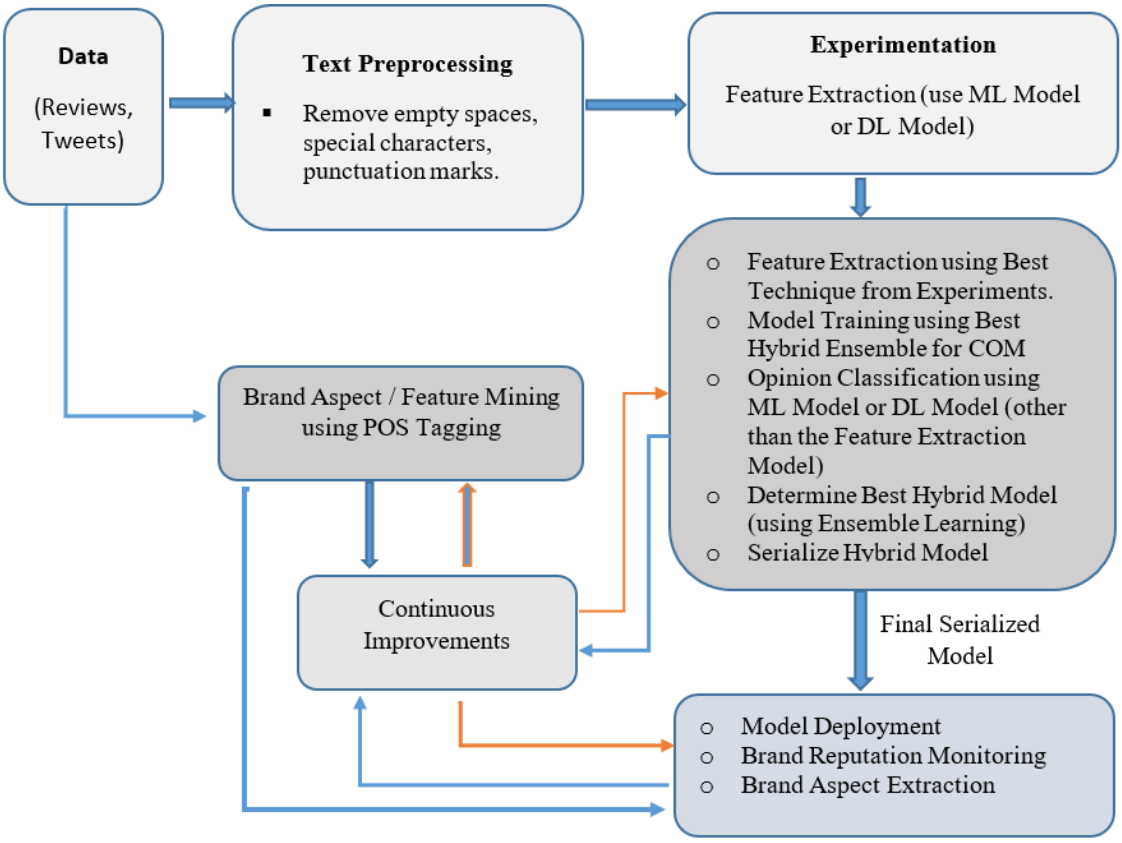

Figure 3. 9 Hybrid Machine Learning Model for Comparative Opinion Mining

3.9.2 Architecture of the Hybrid ML Model for COM

In the architecture of the hybrid ML model, there were two variations arising from the choice of the algorithm that formed the base model and the top-level model. In one instance, a DL technique was used. In this case, the Multilayer Perceptron (MLP) was chosen due to its proven high performance in feature extraction given its use of multiple deep neural networks. It also performs well on huge amounts of data (Goodfellow, Bengio, & Courville, 2016).

The top layer of this architecture was a traditional machine leaning algorithm such as Random Forest or Stochastic Gradient Descent. In the second variation, the Stochastic Gradient Descent (a traditional machine learning algorithm) was used as the base estimator because it produced the highest classification accuracy when used alone. The top-layer of this architecture had other high performance ML algorithm such as Random Forest. These architectural aspects are shown in Figure 3.9. In the architecture, opinions (reviews) capture the elements of comparative opinions, which are analyzed to produce opinion classes that define brand reputation classes.

The base model makes predictions known as probabilities, regarding the final classification. For example, the base model could classify an opinion as positive but without certainty. The probabilities are fed into the topo-level model for final determination of the opinion class.

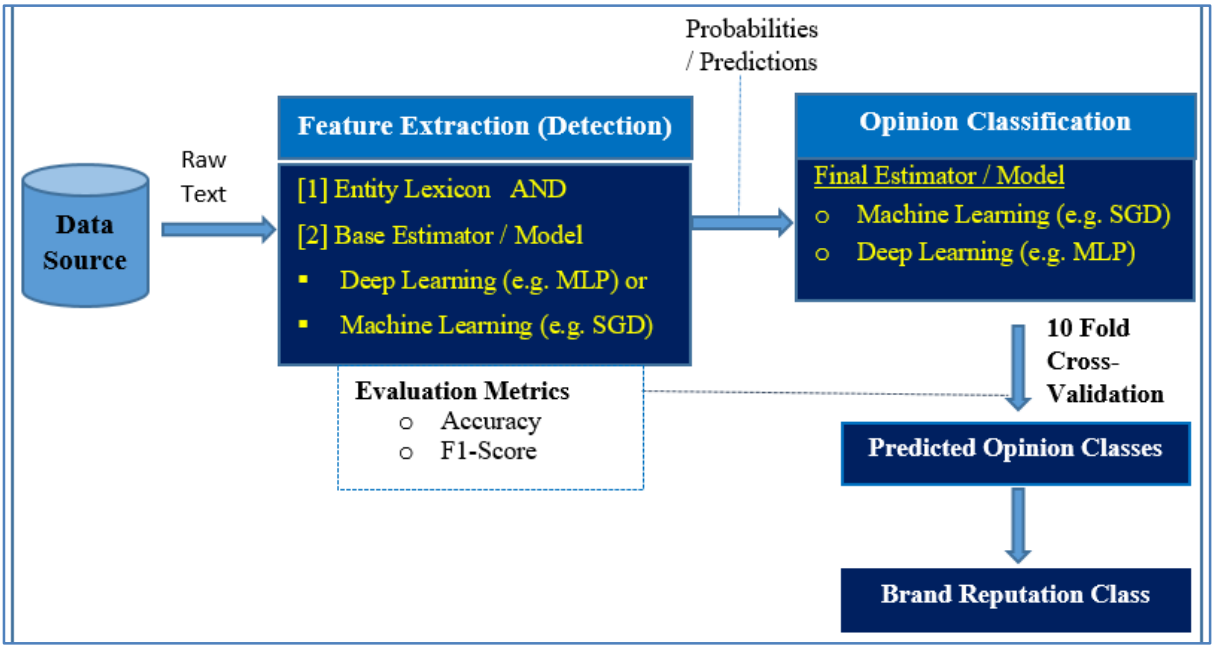

Figure 3. 10 Architecture of the Hybrid Machine Learning Model for COM

The researcher considered using a deep learning model in the hybrid architecture because of the many benefits of deep learning. As a base model, the key benefits could include enhanced feature extraction, capacity to handle different types of data, improved resource utilization, domain adaptation and transfer learning, improved model generalization, and model explainability and interpretability. These benefits are explained in detail in Section 2.8.2. Among the benefits that machine learning presents in a hybrid model, include reduced computational resource utilization, better model interpretability and explainability, and improved model generalization on smaller datasets. Often, the drawbacks of deep learning models are the strengths of machine learning models and vice versa. Thus, combining the two in a hybrid model leverages the strengths of both approaches while reducing the impact of the drawbacks that the two approaches. Section 2.8.1 covers in more detail these benefits by means of contrasting the benefits of deep learning models with machine learning models.

3.9.3 Hybrid ML Model Validation

The developed hybrid model for COM was validated using two approaches. First, an empirical method was used, which involved running the hybrid models on the testing data from the secondary datasets. To measure the performance of the model, accuracy and f1-score metrics were used. Since there were multiple datasets used, the performance was computed as an average of the three datasets. These performances were compared with those obtained from the single ML and DL models applied to the same datasets. The second model validation method was implemented using the ground truth method described in 3.3.4.1. In this method, 100 records that were already labelled by human experts were fed into the hybrid ML model and the results compared to validate the agreement levels between human and machine classification approaches.

3.9.4 Hybrid ML Algorithm Analysis

This section presents the analysis of the hybrid ML algorithm(s) developed in this study. The focus is on the space and time complexity of the algorithm to help determine its feasibility in practical use in use cases like comparative opinion mining.

3.9.4.1 Compute Needs

The hybrid algorithm, if deployed in a production setting or environment, has its own set of computational resources which are critical in ensuring that it performs optimally. It is important to point out that the computational resources are significantly tied to the volume of data that is to be handled by the algorithm. The more the volume of data in terms of disk-space requirements, the higher the resource needs for the algorithm.

However, the minimum requirement needed for the hybrid algorithm to work is a server with 4 CPUs, 8 GB of RAM and a disk-space storage size of 160 GB. In high-throughput settings, increasing both the number of CPUs and RAM chips of the servers would be desirable as this would imply that parallelization of the tasks could be easily performed, hence improving the speed of the algorithm even further. Finally, in cases where the volume of data to be processed per instance is significantly large, it would be worth considering setting up a cluster environment that would allow for the distributed running of tasks and processing of the texts.

3.9.4.2 Complexity Analysis

While there are various components of the entire pipeline that should be investigated for their Big O complexity, the hybrid classifier component is the focus of this section. The hybrid classifiers have four main variables that can be used to evaluate the algorithmic complexity. These are: the number of samples in the training set, the

number of samples in the test set, the number of features in each sample, and the number of hybrid models in the hybrid combination. Collectively, these can be denoted as *n, m, k* and *h* respectively.

3.9.4.3 Complexity Analysis

1. When using a single hybrid ML algorithm (e.g. MLP and RF hybrid or SGD and RF hybrid, a run time of O(h) applies. For the case where multiple hybrids are involved, since the hybrid ML algorithms have to be looped over, the algorithm will have to run *O(h)* times since it has to run every time for each hybrid combination.

2. For the model training where the hybrid ML algorithm is being fitted on the training features and labels, the running time is approximately O(n.k) for each hybrid, h, which is the dot product of the number of training samples and training features in each sample.

3. Similarly, in the prediction phase, the running time is approximated as the dot product of the testing samples and the number of test features, which becomes *O(m.k),* for each hybrid.

4. For the calculation of the performance metrics, the running time is done over a single hybrid model and so for each model, the running time is approximated as *O(m).*

5. Finally, as the hybrid models were also monitored for the time they took to compute, the time monitoring operation is defined using *O(1*) since it is a time-constant operation.

Putting the 5 aspects above together, the combined complexity becomes:

O(h.(n.k + m.k)) which further simplifies to *O(h.k.(n+m))*

The Big O notation above implies that the algorithmic complexity of the hybrid algorithm / model is linearly dependent on the number of hybrid models, the number of features being processed by the algorithm and the respective sizes of the training and test dataset. Therefore, the use of unigrams would result in a less complex algorithm as the number of features per input are less as compared to bigram and trigrams. This difference subsequently explains why unigram models had the shortest runtimes while trigram models had the highest runtimes as seen in the shared notebooks.

3.10 Prototype Development

3.10.1 Prototype Development Methodology

To develop a system prototype as a proof of concept in support of the developed HML model for comparative opinion mining, the agile methodology was used because it has wide applications in software development, focusing on promoting responsiveness to customer needs, adaptability, and improved collaboration. This is imperative, as the agile principles are central to tackling the unique challenges associated with machine-learning projects (Elman & Turk, 2017). The Agile methodology has ten phases:

i. **Project Initiation** - defining the scope and goals of the project to align them with organizational objectives (Doshi-Velez & Kim, 2017). Success criteria is defined.

ii. **Sprint Planning** - involves prioritization of tasks as well as the performance of the ML model. Complex tasks are broken down into manageable units (Cohn, 2014).

iii. **Sprint Execution** - in this phase, ML models are developed incrementally depending on the tasks. Development challenges and progress are regularly discussed in meetings to evaluate the progress (Humble & Farley, 2010). This is

vital because ML components need to be tested and integrated to for seamless working (Fowler & Foemmel, 2012).

iv. **Customer Feedback -** this task is key in adjusting ML algorithms as well tackling issues with the quality of data. The methodology encourages provision of feedback during the development process hence improving collaboration (Beck et al., 2001).

v. **Incremental Delivery -** this gives room for user-testing, collection of valuable data, and the validation of assumptions (Kale & Karamcheti, 2018).

vi. **Retrospectiveness and Sprint Review** - this helps evaluate the development process, identify challenges, and adapt key practices (Derby & Larsen, 2006).

vii. **Iterative Development** - this enables ML project teams to respond to requirement changes and/or evolving data during project execution (Lam et al., 2021).

viii. **Continuation Integration and Deployment (CI/CD) -** this is done to achieve automation in testing, validation, and model deployment. In turn, this ensures that changes can be reliably and rapidly deployed to the live or production environment.

ix. **Monitoring and Maintenance** – this is done to confirm the models aligned with the goals of the business.

x. **Scaling and optimization -** for larger ML projects, scaling, and optimization is key as the project attains maturity. This improves model usage; adjustment to evolving user requirements besides making the model capable of handling larger datasets to achieve better performance without influencing the accuracy of the ML model.

3.10.2 Prototype Development Steps

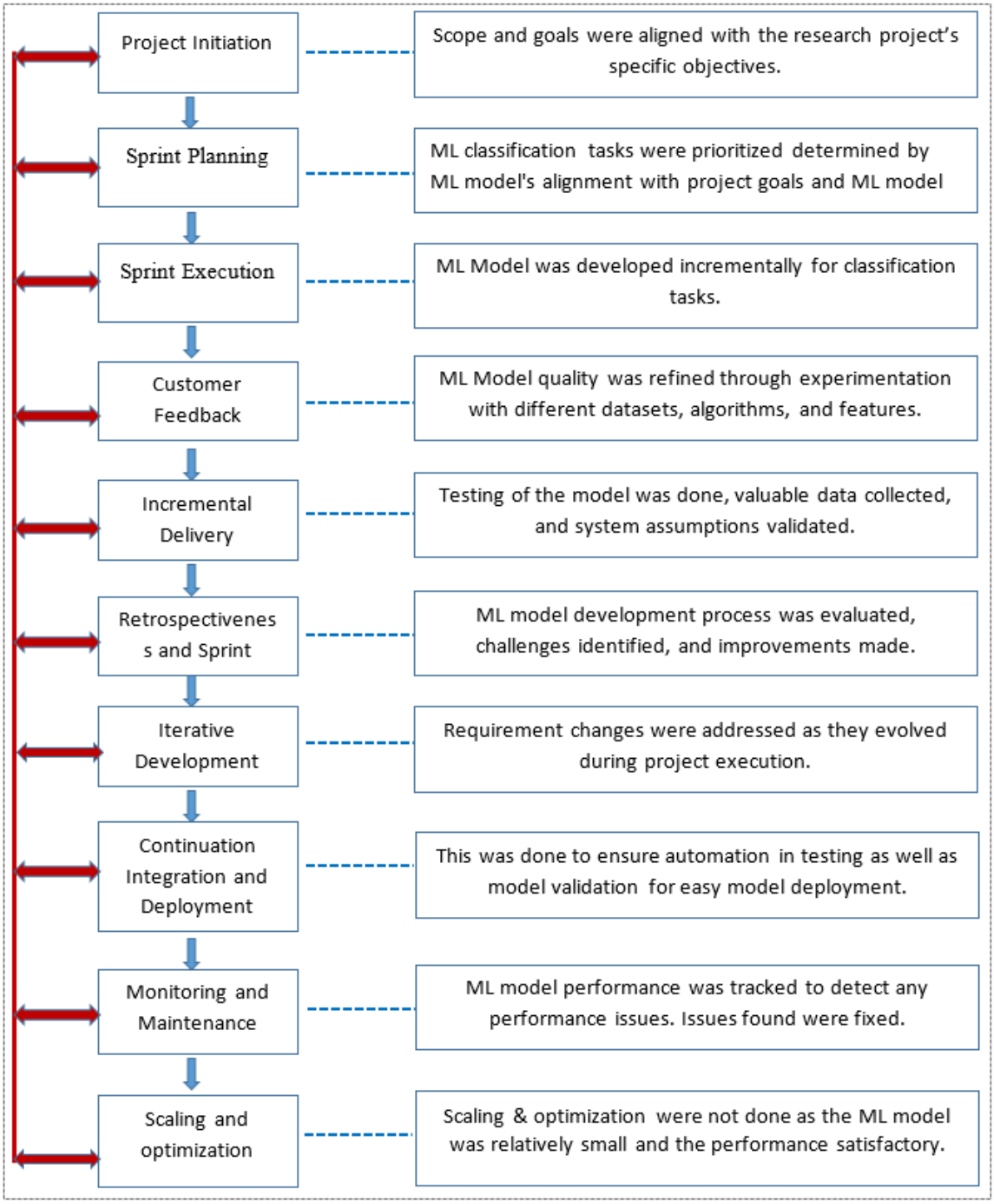

Figure 3. 11 Software (Prototype) Development Methodology

(Source: Elman & Turk, 2017).

The following is the Pseudocode of the Hybrid Machine Learning Model for Comparative Opinion Mining.

```
Input:
B.O.W vectors set X;
Stacked ensemble F:
Base estimator model f;
Final estimator model g;
F = f + g #stacked ensemble
Process F(X):
base_preds = f(x);
final_preds = cross_validate(g(base_preds))
return final_preds
Output:
final_preds.
```

Figure 3. 12 Pseudocode for the Hybrid ML Model for COM (Ondara et al., 2023).

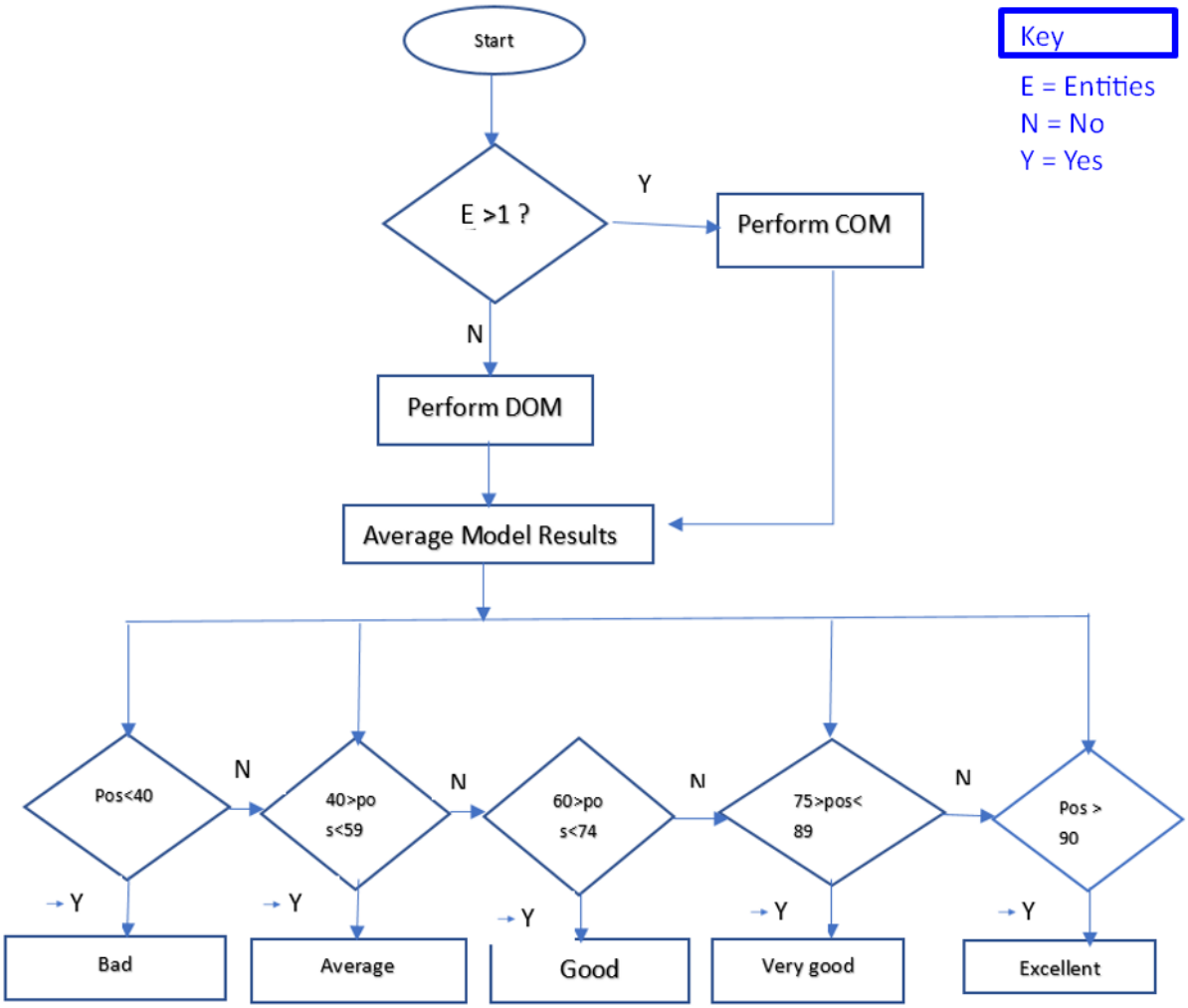

Figure 3. 13 Flowchart for the Hybrid System Prototype Development

3.10.3 Prototype Components

In terms of the standard process for machine learning, this section presents the general process followed in prototype development. The process involved the following phases: Data Collection, Data pre-processing, Development of the Hybrid Machine Learning Model for COM, Evaluation of the performance of the Hybrid Machine Learning Model for COM, and Model hyper parameter tuning. From the review of existing literature in comparative opinion mining, the findings from the pilot study conducted, this study established that a hybrid machine-learning model for comparative opinion mining would need several features. Table 3.15 presents the list of system features and their roles in the hybrid machine-learning model.

Table 3. 15 Pilot study Outcomes that Influenced System Development

Finding	**Role in the Hybrid ML Model for COM**
Data Extractor	Enables the model to extract data from a specific data source for use in brand reputation monitoring (Arboleda et al., 2017).
Distinct Brand Entities	This is the filter for data extraction from the various data sources for purposes of COM (Baziotis et al., 2017).
Opinion / Sentiment Classifier	This component is responsible for assigning opinion polarity to each brand, given a dataset (Yueyang & Wang, 2019).
Brand Preference Indicator	The basis for brand reputation monitoring. With this, the system would be able to show the preferred brand in the analyzed data (Varathan et al., 2017).
Brand Aspects Module	The basis for brand entity comparisons in opinionated comparative data / content. The aspects are key markers of what the brand is strong at or weak and would help brands to gain a competitive advantage if used well (Diamantini et al., 2019).

Brand Reputation Trend Indicator	This component is used to show the evolution of a brand's reputation over a specific period. This will help interested brands to know how their brand's positivity or negativity is changing over time (Perakakis et al., 2019).

3.10.4 System Prototype Development Process

The following steps were involved in the development of the system prototype for COM.

i. Identification of features from comparative opinionated texts
ii. Development of Hybrid Machine Learning Algorithm / Technique for COM
iii. Database Design
iv. User Interface (Dashboard) Design
v. Brand Reputation Indicator Development
vi. Testing of the Hybrid Machine Learning Model for COM

3.11 Data Analysis

The quantitative data collected was subjected to descriptive analysis. Qualitative data collected was subjected to content analysis to establish perceived challenges among industry experts in the access to and utilization of automated tools in the monitoring of brand reputation. To analyze the quantitative data, specialize software tools were used. First, the outputs from the machine learning models were analyzed in Microsoft Excel 2016. This was basic data where experimental results obtained in CSV files was transferred to MS Excel for easy of tabulation and analysis of the algorithm performance metrics such as accuracy and f1-score. This involved sorting according to accuracy and identifying the best performing model.

In computing the statistical significance of the various machine learning models experiment with across different datasets, R programming language was used. For ease of statistical processing, Jamovi, a GUI for R specially created for statistical research was used in performing the analysis of statistical variance (ANOVA) and T-tests. ANOVA was used to determine the statistical significance in the accuracy of the different ML models while T-test was used to determine the best performing model based on two models that had the highest accuracy. ANOVA is useful when comparing multiple groups for purposes of evaluating if there are statistically significant differences in their averages. Besides this, where several factors are interacting to establish the most influential features or factors (Davis et al., 2020).

3.11.1 One-way ANOVA

This was used to determine if there were statistically significant differences in model accuracy between multiple groups. This method is applicable when there are at least three groups. An example of its usage was in testing if there was a significant difference in the accuracies of the machine-learning models used to perform COM on multiple datasets, using multiple feature extraction techniques like Count Vectorizer and TF-IDF (Smith et al., 2020).

3.11.2 Two-way ANOVA

This was used to determine if there were significant interactions between two groups in predicting the outcome of opinion classes, among multiple groups. This method is applicable when there are at least three groups. An example of its usage was in testing if there was a significant difference in the accuracies of the eight single machine-learning models used to perform comparative opinion mining on multiple datasets, using multiple feature extraction techniques like Count Vectorizer and TF-IDF (Smith

et al., 2020). The aim was to identify whether there were statistically significant interactions between ML models and datasets.

3.11.3 T-tests

Independent sample T-tests as well as paired t-tests can be used to evaluate the difference in opinion classes between two particular conditions or groups. Independent sample t-tests are used if two independent groups are being compared. An example of this application is comparing the opinions towards tow specific brands. On the other hand, paired T-tests are employed when comparing opinions prior to an intervention and after the intervention; or when comparing a person's opinions on two different occasions (Chen & Wang, 2019).

Table 3. 16 Mapping Research Objectives to Data Analysis Methods / Techniques

Independent Variable	**Moderating Variable**	**Independent Variable**	**Data Analysis Method**
Comparative Opinion Elements	Machine Learning Technique	Brand Reputation	▪ Accuracy ▪ ANOVA Test ▪ T-Test
Comparative Opinion Elements	Vectorization Technique (Feature Extraction Technique)	Brand Reputation	▪ Accuracy ▪ ANOVA Test ▪ T-Test
Comparative Opinion Elements	N-Gram Range or Window Size	Brand Reputation	▪ Accuracy ▪ ANOVA Test ▪ T-Test

Table 3. 17 Data Analysis Methods for the Experiments 1 - 40

SN	*Algorithm*	*Feature Extraction Technique*	*n-gram range value or window size*	*Dataset*	*Data Analysis Method*
1 - 9	Single ML Algorithms: MNB, SVM, KNN, DT, RF, LR, SGD, MLP	Count Vectorizer	N = 1 to 3	D1 – D3	One-Way ANOVA: Testing for significant statistical differences in accuracies of the Single ML algorithms.
10 - 18	Single ML Algorithms: MNB, SVM, KNN, DT, RF, LR, SGD, MLP	TFIDF	N = 1 to 3	D1 – D3	
19 - 24	Single ML Algorithms: MNB, SVM, KNN, DT, RF, LR, SGD, MLP	CBOW	W = 1	D1: Microsoft vs. Google	• IV – Comparative Sentences • DV – Accuracy • MV – Hybrid Algorithm, Feature Extraction Technique, Dataset, Feature Extraction Parameters
25 - 30	Single ML Algorithms: MNB, SVM, KNN, DT, RF, LR, SGD, MLP	Skip gram	W = 1	D1: Microsoft vs. Google	
31 - 33	Hybrid ML Algorithms: DT, MLP + RF, MLP + SGD, MLP + SVM, SGD + DT, SGD + RF, SGD + MLP, SGD + SVM	CV	N = 3	D1: Microsoft vs. Google	One-Way ANOVA: Testing for significant statistical differences in accuracies of the Hybrid ML algorithms.
34 - 37	Hybrid ML Algorithms: DT, MLP + RF, MLP + SGD, MLP + SVM, SGD + DT, SGD + RF, SGD + MLP, SGD + SVM	TFIDF	N = 3	D1: Microsoft vs. Google	• IV – Comparative Sentences • DV – Accuracy • MV – Hybrid Algorithm, Feature

38 - 40	-	Hybrid Algorithms: DT, MLP + RF, MLP + SGD, MLP + SVM, SGD + DT, SGD + RF, SGD + MLP, SGD + SVM	ML CV	N = 3	Primary D1 - D3	Extraction Technique, Dataset, Feature Extraction Parameters

3.12 Research Ethics

This section presents the ethical considerations that guided this study. This included seeking authorization or approvals to conduct the research, avoidance of plagiarism, and proper citations and referencing of publications used in the study. The Postgraduate School of Kenyatta University granted research Authorization for this study after approving the research proposal presented to them. The National Commission for Science, Technology, and Innovation (NACOSTI) licensed this study.

To comply with research ethics, all materials referred to in this study are properly cited using APA referencing style. Due diligence was followed to ensure minimal plagiarism. This was achieved through in-text citations, referencing, and re-writing of useful texts. Thus, copyrights of other researchers were respected. In the case of the focused group discussion, the experts involved consented to participate in this study with a condition of remaining anonymous.

API Guidelines and Ethical Standards – information obtained using developer API was strictly for purposes of aiding this study. Data mining (extraction) was done in a manner to ensure only user opinions were mined for analysis. No user profile or account details beyond that which aided this study were collected and/or used in this study.

CHAPTER FOUR

RESULTS

4.1 Introduction

This chapter presents the results of this study from the following three components: (1) pilot study, (2) model development, and (3) experimentation. The purpose of this study was to develop a hybrid machine-learning model for comparative opinion mining. As a proof of the concept, the prototype using the hybrid model was applied to brand reputation monitoring in multiple domains. Guided by five research questions and one hypothesis, the following section presents the findings obtained from this study.

4.2 Experimental Results

In this section, the following words, symbols, and their working definitions are applied.

Statistic: this is the numerical measurement or summary for describing a specific characteristic of a given dataset. It is a computed value whose significance is in understanding the variability and central tendency of data. Examples of statistic values include the standard deviation, media, range, and mean. These statistics help researchers to condense data for better interpretation (Wasserstein & Lazar, 2016).

Degrees of Freedom (df): these are concepts with fundamental value especially in hypothesis testing where they represent the count of values that can freely vary. For this reason, df is important in determining the correct distribution for hypothesis testing, such as t-tests and chi-square distributions. In the case of t-test used in this study, df is associated with the size of the sample used with respect to determining the actual shape of the t-distribution (Khuri, 2013).

P-value: this is a critical measure in statistical analysis as it is used to quantitatively reveal the evidence available against accepting the null hypothesis. As a probability value, the p-value shows the likelihood of getting more extreme or observed results, assuming that the proposed null hypothesis is actually true. A p-value below 0.05 implies that the results observed are not likely to occur by random chance hence the reason to reject the null hypothesis and accept the alternative hypothesis. Conversely, when the p-value is greater than 0.05, the implication is that data does not show strong evidence to reject the null hypothesis (Fisher, 1992).

4.2.1 Performance of Existing Machine Learning Models in COM

RQ1: How do the existing machine learning models for comparative opinion mining compare in their performance?

Eight ML techniques were independently applied to the same datasets to evaluate their performance based on accuracy and f1-score. The experiments were repeated for different feature vectorization techniques, including Count Vectorizer, TFIDF, and Word2Vec (both CBOW and Skip Gram). In the case of Count Vectorizer and TFIDF, the n-gram ranges experimented with included unigrams, bigrams, and trigrams. The results shown in this section are based on average performance of each ML model, DL model, and/or hybrid machine-learning model (where deep learning is understood in the context of it being a sub-field of ML).

4.2.1.1 Machine Learning Models

Tables 4.1 to 4.3 present the accuracy of the different single ML models in performing COM using Count Vectorizer for feature extraction and for Unigrams, Bigrams, and Trigrams respectively. The D1, D2, and D3 represent the three datasets used. The

results for D3 seem to have suffered model overfitting due to the small number of reviews in this dataset.

Table 4. 1 Average Performance of ML Models: CV + Unigrams

ML Technique	D1	D2	D3	Avg
Random Forest	84.2	93.6	100.0	92.6
Multilayer Perceptron	83.6	92.3	100.0	92.0
Decision Tree	83.2	92.6	100.0	91.9
Stochastic Gradient Descent	79.6	89.0	100.0	89.5
Logistic Regression	70.0	83.2	100.0	84.4
Support Vector Machine	62.4	77.6	100.0	80.0
Gaussian Naïve Bayes	58.4	74.3	100.0	77.6
K-Nearest Neighbors	50.3	57.1	100.0	69.1

Table 4. 2 Average Performance of ML Models: CV + Bigrams

ML Technique	D1	D2	D3	Avg
Stochastic Gradient Descent	85.4	92.8	100.0	92.7
Random Forest	85.1	92.8	100.0	92.6
Multilayer Perceptron	*85.3*	*92.4*	*100.0*	*92.6*
Decision Tree	85.1	91.7	99.7	92.1
Logistic Regression	84.3	91.4	99.7	91.8
Gaussian Naïve Bayes	75.3	87.9	99.7	87.6
Support Vector Machine	74.3	84.9	99.7	86.3
K-Nearest Neighbors	51.0	57.7	100.0	69.6

Table 4. 3 Average Performance of ML Models: CV + Trigrams

ML Technique	D1	D2	D3	Avg
Multilayer Perceptron	86.4	92.4	100.0	92.9
Logistic Regression	86.0	92.0	99.7	92.5
Random Forest	85.2	92.4	99.7	92.4
Stochastic Gradient Descent	84.8	92.4	99.7	92.3
Decision Tree	84.2	92.0	99.0	91.7
Gaussian Naïve Bayes	79.8	89.7	99.7	89.7
Support Vector Machine	76.5	86.3	99.7	87.5
K-Nearest Neighbors	47.1	53.3	100.0	66.8

Tables 4.4 to 4.6 present the accuracy of the different single ML models in performing COM using TFIDF for feature extraction and for Unigrams, Bigrams, and Trigrams respectively. The D1, D2, and D3 represent the three datasets used.

Table 4. 4 Average Performance of ML Models: TFIDF + Unigrams

ML Technique	**D1**	**D2**	**D3**	**Avg**
Random Forest	83.4	92.1	99.7	91.7
Multilayer Perceptron	82.0	91.3	100.0	91.1
Decision Tree	81.2	92.0	99.3	90.8
Stochastic Gradient Descent	78.0	89.0	100.0	89.0
Support Vector Machine	74.1	86.9	100.0	87.0
Gaussian Naïve Bayes	59.5	74.4	100.0	78.0
Logistic Regression	61.4	70.0	100.0	77.1
K-Nearest Neighbors	53.2	52.4	100.0	68.6

Table 4. 5 Average Performance of ML Models: TFIDF + Bigrams

ML Technique	**D1**	**D2**	**D3**	**Avg**
Random Forest	85.5	92.3	99.7	92.5
Multilayer Perceptron	85.2	92.1	100.0	92.4
Stochastic Gradient Descent	84.2	92.8	100.0	92.3
Decision Tree	84.0	91.8	99.0	91.6
Support Vector Machine	80.8	90.0	99.7	90.1
Gaussian Naïve Bayes	75.3	87.6	100.0	87.6
Logistic Regression	65.4	74.3	100.0	79.9
K-Nearest Neighbors	53.8	52.8	100.0	68.8

Table 4. 6 Average Performance of ML Models: TFIDF + Trigrams

ML Technique	**D1**	**D2**	**D3**	**Avg**
Random Forest	85.8	93.0	100.0	92.9
Stochastic Gradient Descent	86.1	92.8	100.0	92.9
Multilayer Perceptron	85.6	92.1	100.0	92.6
Decision Tree	84.5	92.4	99.0	92.0
Support Vector Machine	81.2	90.7	99.7	90.5
Gaussian Naïve Bayes	80.0	89.3	100.0	89.8
Logistic Regression	66.2	75.0	100.0	80.4
K-Nearest Neighbors	54.3	53.3	100.0	69.2

Tables 4.7 to 4.8 present the performance of the different single ML models in performing COM using CBOW for feature extraction and for Window Sizes of 1 and 5 respectively. The D1, D2, and D3 represent the three datasets used.

Table 4. 7 Average Performance of ML Models: CBOW + Window Size 5

ML Technique	D1	D2	D3	Avg
K-Nearest Neighbors	41.7	40.1	48.7	43.5
Gaussian Naïve Bayes	41.7	39.9	48.7	43.4
Decision Tree	41.5	39.9	48.7	43.4
Stochastic Gradient Descent	41.5	39.9	48.7	43.4
Random Forest	41.7	39.1	48.7	43.2
Multilayer Perceptron	14.3	40.3	51.6	35.4
Support Vector Machine	38.9	30.2	34.5	34.6
Logistic Regression	4.1	24.4	39.1	22.6

Table 4. 8 Average Performance of ML Models: Skip gram + Window Size 5

ML Technique	D1	D2	D3	Avg
K-Nearest Neighbors	41.7	40.2	48.7	43.5
Decision Tree	41.5	39.9	48.7	43.4
Multilayer Perceptron	41.7	39.7	48.7	43.4
Stochastic Gradient Descent	41.7	39.2	48.7	43.2
Logistic Regression	41.7	39.1	48.7	43.2
Random Forest	38.9	30.2	34.5	34.6
Gaussian Naïve Bayes	26.1	28.0	44.1	32.7
Support Vector Machine	5.0	24.7	39.1	22.9

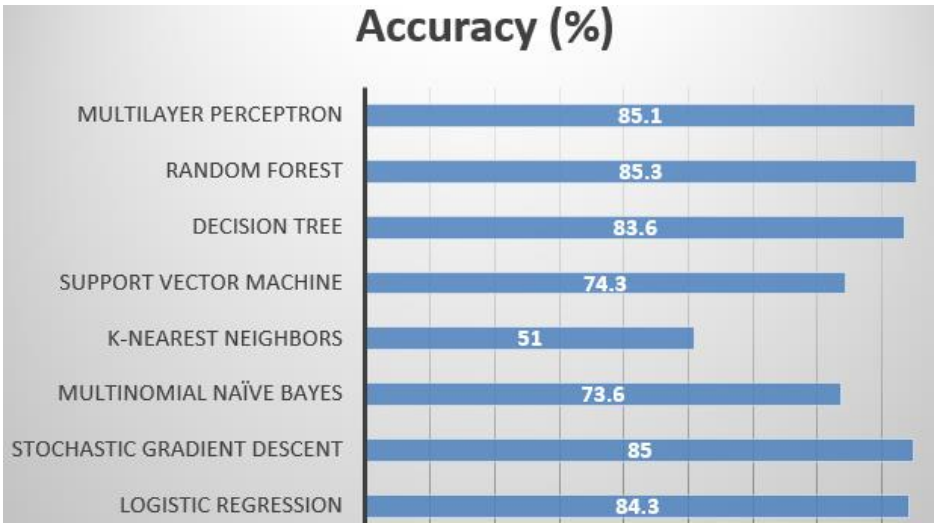

Figure 4. 1 Accuracy levels of the ML Models Using CV + Unigrams on Dataset #1

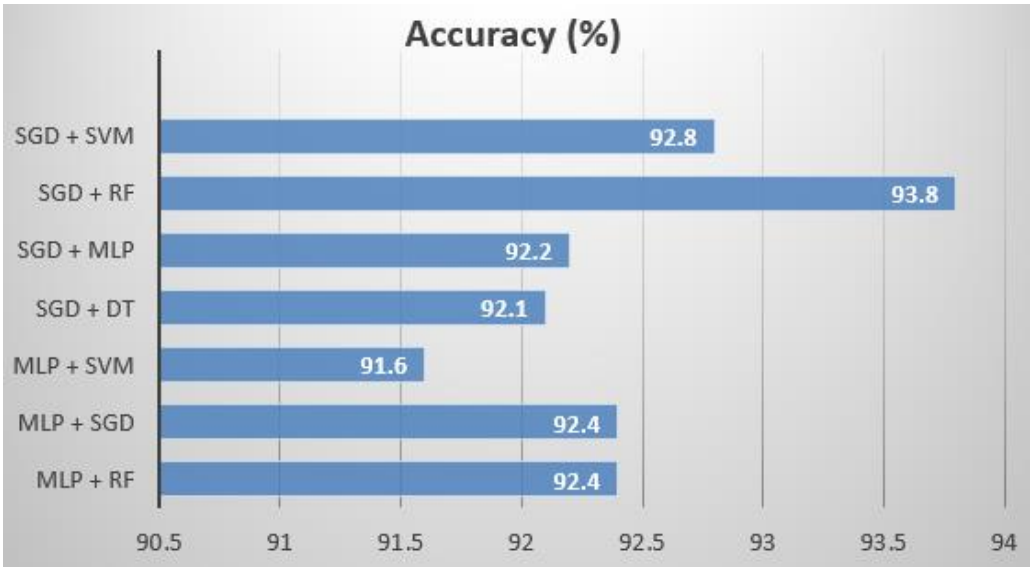

Figure 4. 2 *Accuracy levels of the ML Models Using CV + Unigrams on Dataset #1*

Statistical Test of Significance

Table 4.9 shows One-Way ANOVA results from testing if there was a statistically significant difference in the accuracies of the eight single machine learning models when applied to three different comparative opinion datasets using Count Vectorizer and trigram features. The p-value (0.04) is below 0.05, suggesting a strong evidence against the null hypothesis, thereby confirming that there is a significant statistical difference in the accuracies of the ML models across the three datasets.

Table 4. 9 One Way ANOVA (Welch's)

	F	df1	df2	p
Accuracy	9.57	7	6.83	0.004

Table 4.10 shows One-Way ANOVA test results from testing if there was a statistical difference in the accuracies of the eight single machine-learning models when applied to three different comparative opinion datasets using TFIDF and trigram features. The p-value (0.077) is above 0.05, indicating mild evidence against the null hypothesis. This

suggests that there is insignificant statistical difference in the accuracies of the ML models across the three datasets.

Table 4. 10 One Way ANOVA (Welch's)

	F	df1	df2	p
Accuracy	3.20	7	6.76	0.077

Table 4.11 shows One-way ANOVA test results from testing if there was a significant statistical difference in the accuracies of the eight ML models using CV and trigram features. The p-value is 0.265, indicating there is no statistically significant difference between the accuracy of algorithms based on the datasets used.

Table 4. 11 One-way ANOVA – ML Algorithms vs Accuracy

	Sum of Squares	df	Mean Square	F	p
Algorithm	1648	7	235	1.42	0.265
Residuals	2655	16	166		

Table 4. 12 Normality Test for ML Algorithms vs Accuracy

Normality Test (Shapiro-Wilk)	
Statistic	**p**
0.921	0.061

The normality in table 4.12 shows a Shapiro-Wilk p-value of 0.061, which is above 0.05, showing a non-violation of the normality assumption.

4.2.1.2 Deep Learning Models

In the results shown in tables 4.13 to 4.15, the runtime (ms) metric indicates how long the model took to train while the prediction latency shows the duration the model took to predict the opinion classes. Tables 4.13 to 4.15 present results for the application of DL models in COM using datasets 1 to 3. The results show that the across all the datasets, the accuracy of CNN exceeds that of LSTM and MLP.

Table 4. 13 Performance of DL Techniques on Dataset #1

Deep Learning Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Runtime (ms)	Prediction Latency (ms)
CNN	87.4	87.4	87.4	87.3	175.6	1.8
LSTM	80.6	81.5	80.6	80.8	2711.3	41.4
MLP	85.6	86.3	85.6	85.8	42.9	0.0

Table 4. 14 Performance of DL Techniques on Dataset #2

Deep Learning Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Runtime (ms)	Prediction Latency (ms)
CNN	93.1	93.5	93.1	93.2	176.8	2.0
LSTM	88.4	89.3	88.4	88.5	2714.1	41.2
MLP	92.4	92.9	92.4	92.5	35.7	0.0

Table 4. 15 Performance of DL Techniques on Dataset #3

Deep Learning Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Runtime (ms)	Prediction Latency (ms)
CNN	99.7	99.7	99.7	99.7	61.5	0.7
LSTM	99.3	99.4	99.3	99.3	951.6	13.9
MLP	77.5	78.9	77.5	77.9	155.8	0.0

4.2.2 Developing a Hybrid Machine Learning Model for COM

RQ4: How can a hybrid machine-learning model for comparative opinion mining be developed based on the experimental results?

From literature analysis and empirical study, the researcher established that one of the best ways to develop a hybrid ML model is through ensemble learning method. The approach of using this method in developing hybrid ML models was described in Section 2.13. Indeed, studies show that models using ensemble method of learning are known as hybrid models. In our method, the researcher employed two scenarios:

i. The base model being MLP (deep learning algorithm) as the base model for extracting features from text while the traditional machine learning algorithm like RF performs the classification using the probability features that MLP feeds it with. Top-level models included other algorithms like RF, DT, and SVM. This method worked so well, producing the best performing hybrid machine-learning model for comparative opinion mining. This model was the MPL + RF.

ii. Using SGD (traditional machine learning algorithm) as the base model for extracting features from text and feeding the same to the top-level model for classification. Top-level models included other algorithms like RF, DT, and MLP in this case. This option gave a high performance hybrid model, SGD + RF, whose performance was very close to that of the first option.

In both scenarios, the hybrid model received entities features through a lexical approach. The researcher created a lexicon consisting of entity names (brand names) and their common brand name variations to reduce computational complexity if the researcher were to use the named entity extraction approach. Since the brands the

researcher were interested in were already known, identified using a purposive sampling strategy, a simple dictionary of brand names was needed. This was the hybridization approach achieved in our model at feature extraction level.

4.2.3 Effectiveness of the Hybrid Machine Learning Model for COM

RQ5: How effective would the hybrid machine-learning model be in performing comparative opinion mining?

4.2.3.1 Accurate Prediction of Opinion Classes

For the tables 4.16 to 4.23, D represents the dataset. For example, D1 represents dataset 1. The *Avg. accuracy* column represents average accuracy of the model across the datasets 1 to 3. This metric was used to capture the smoothed accuracy given multiple datasets considering that different datasets yielded varying accuracies for purposes of ensuring model robustness (Kleinberg et al., 2018).

Table 4. 16 Average Performance of the Hybrid ML Models for CV + Unigrams

Hybrid ML Model	D1: Accuracy (%)	D2: Accuracy (%)	D3: Accuracy (%)	Avg. Accuracy (%)
MLP + DT	***83.4***	***92.4***	***100.0***	***92.0***
MLP + RF	***83.6***	***92.4***	***100.0***	***92.0***
MLP + SGD	78.4	90.4	94.7	87.9
MLP + SVM	75.7	88.2	100.0	88.0
SGD + DT	80.2	92.9	100.0	91.0
SGD + MLP	69.8	89.2	100.0	86.3
SGD + RF	82.4	91.8	99.7	91.3
SGD + SVM	59.2	86.6	100.0	81.9

Table 4. 17 Average Performance of the Hybrid ML Models for CV + Bigrams

Hybrid ML Model	D1: Accuracy (%)	D2: Accuracy (%)	D3: Accuracy (%)	Avg. Accuracy (%)
MLP + DT	86.2	92.4	99.3	92.7
MLP + RF	***86.1***	***92.6***	***99.7***	***92.8***
MLP + SGD	87.1	93.1	95.1	91.7
MLP + SVM	83.8	90.6	100.0	91.5
SGD + DT	83.6	92.6	99.7	92.0
SGD + MLP	81.3	92.1	96.7	90.0
SGD + RF	80.5	92.1	99.7	90.8
SGD + SVM	78.9	91.7	100.0	90.2

Table 4. 18 Average Performance of the Hybrid ML Models for CV + Trigrams

Hybrid ML Model	D1: Accuracy (%)	D2: Accuracy (%)	D3: Accuracy (%)	Avg. Accuracy (%)
MLP + DT	86.0	92.0	98.4	92.1
MLP + RF	***86.6***	***92.6***	***99.7***	***92.9***
MLP + SGD	86.9	92.3	91.8	90.4
MLP + SVM	85.6	91.9	99.7	92.4
SGD + DT	63.1	92.0	99.7	84.9
SGD + MLP	84.6	91.0	95.7	90.4
SGD + RF	85.6	91.3	100.0	92.3
SGD + SVM	85.7	92.1	100.0	92.6

Table 4. 19 Average Performance of the Hybrid ML Models for TFIDF + Unigrams

Average Accuracy of the Hybrid ML Models: TFIDF + Unigrams				
Hybrid ML Model	**D1: Accuracy (%)**	**D2: Accuracy (%)**	**D3: Accuracy (%)**	**Avg. Accuracy (%)**
MLP + DT	82.9	91.0	99.7	91.2
MLP + RF	***81.3***	***91.9***	***100.0***	***91.1***
MLP + SGD	77.7	89.0	100.0	88.9
MLP + SVM	78.7	89.1	100.0	89.3
SGD + DT	78.5	91.0	100.0	89.8
SGD + MLP	77.2	86.0	100.0	87.7
SGD + RF	81.2	92.0	100.0	91.1
SGD + SVM	64.5	89.7	100.0	84.7

Table 4. 20 Average Performance of the Hybrid ML Models for TFIDF + Bigrams

Average Accuracy of the Hybrid ML Models: TFIDF + Bigrams				
Hybrid ML Model	**D1: Accuracy (%)**	**D2: Accuracy (%)**	**D3: Accuracy (%)**	**Avg. Accuracy (%)**
MLP + DT	85.3	92.2	99.3	92.3
MLP + RF	85.7	92.4	99.7	92.6
MLP + SGD	85.8	92.4	100.0	92.8
MLP + SVM	85.0	92.1	100.0	92.4
SGD + DT	***86.4***	***92.6***	***99.7***	***92.9***
SGD + MLP	84.7	92.3	99.7	92.2
SGD + RF	86.4	93.0	100.0	93.1
SGD + SVM	81.2	92.9	100.0	91.4

Table 4. 21 Average Performance of the Hybrid ML Models for TFIDF + Trigrams

Average Accuracy of the Hybrid ML Models: TFIDF + Trigrams				
Hybrid ML Model	**D1: Accuracy (%)**	**D2: Accuracy (%)**	**D3: Accuracy (%)**	**Avg. Accuracy (%)**
MLP + DT	86.6	92.2	99.0	92.6
MLP + RF	86.1	92.4	99.7	92.7
MLP + SGD	86.3	92.4	100.0	92.9
MLP + SVM	85.4	91.6	100.0	92.3
SGD + DT	85.3	92.1	99.7	92.4
SGD + MLP	***86.6***	***92.2***	***100.0***	***92.9***
SGD + RF	86.4	93.8	100.0	93.4
SGD + SVM	64.7	92.8	100.0	85.8

Eliminating Dataset #3 Due to Potential Model Overfitting

Table 4.22 presents the accuracy of the hybrid ML models on Dataset 1 and Dataset 2, ignoring Dataset 3 which had caused model overfitting due to its small size. The results are for when Count Vectorizer technique was used for feature extraction, with n-gram range of 3.

Table 4. 22 Average Performance of the Hybrid ML Models for CV + Trigrams

Average Accuracy of the Hybrid ML Models: CV + Trigrams			
Hybrid ML Model	**D1: Accuracy (%)**	**D2: Accuracy (%)**	**Avg. Accuracy (%)**
MLP + DT	86.0	92.0	89.0
MLP + RF	***86.6***	***92.6***	***89.6***
MLP + SGD	86.9	92.3	89.6
MLP + SVM	85.6	91.9	88.8
SGD + DT	63.1	92.0	77.5
SGD + MLP	84.6	91.0	87.8
SGD + RF	85.6	91.3	88.5
SGD + SVM	85.7	92.1	88.9

Table 4. 23 Average Performance of the Hybrid ML Models for TFIDF + Trigrams

Average Accuracy of the Hybrid ML Models: TFIDF + Trigrams			
Hybrid ML Model	**D1: Accuracy (%)**	**D2: Accuracy (%)**	**Avg. Accuracy (%)**
MLP + DT	86.6	92.2	89.4
MLP + RF	86.1	92.4	89.3
MLP + SGD	86.3	92.4	89.4
MLP + SVM	85.4	91.6	88.5
SGD + DT	85.3	92.1	88.7
SGD + MLP	86.6	92.2	89.4
SGD + RF	***86.4***	***93.8***	***90.1***
SGD + SVM	64.7	92.8	78.7

Table 4.24 presents the performance of Hybrid ML Models when using Count Vectorizer for feature extraction and n-gram range of 2 on Dataset 2. The results show that MLP and SGD hybrid had the highest accuracy among the hybrid models.

Table 4. 24 Performance of Hybrid ML Techniques for CV + Bigrams for Dataset #2

Hybrid ML Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Runtime (ms)	Prediction Latency (ms)
MLP + DT	92.4	92.8	92.4	92.5	33.0	0.0
MLP + RF	92.6	92.9	92.6	92.6	41.2	0.1
MLP + SGD	93.1	93.5	93.1	93.2	31.1	0.0
MLP + SVM	90.6	91.3	90.6	90.6	35.1	0.2
SGD + DT	92.6	93.0	92.6	92.6	1.1	0.0
SGD + MLP	92.1	92.5	92.1	92.1	32.6	0.0
SGD + RF	92.1	92.5	92.1	92.1	8.3	0.1
SGD + SVM	91.7	92.1	91.7	91.6	2.8	0.1

Table 4.25 presents the performance of Hybrid ML Models when using Count Vectorizer for feature extraction and n-gram range of 3 on Dataset 2. The results show that MLP and SGD hybrid had the highest accuracy among the hybrid models.

Table 4. 25 Performance of Hybrid ML Techniques for CV + Trigrams for Dataset #2

Hybrid ML Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Runtime (ms)	Prediction Latency (ms)
MLP + DT	92.0	92.4	92.0	92.0	52.2	0.0
MLP + RF	92.6	92.9	92.6	92.6	63.4	0.1
MLP + SGD	92.3	92.7	92.3	92.3	50.2	0.0
MLP + SVM	91.9	92.6	91.9	91.9	57.8	0.2
SGD + DT	92.0	92.3	92.0	92.0	1.5	0.0
SGD + MLP	91.0	91.9	91.0	90.6	50.2	0.0
SGD + RF	91.3	91.9	91.3	91.2	10.6	0.1
SGD + SVM	92.1	92.8	92.1	92.1	3.3	0.2

Tables 4.26 to 4.29 present the performance of Hybrid ML Models when using CBOW feature extraction technique with window sizes of 1 an d5 respectively. The results show that MLP and RF hybrid model had the highest accuracy among the hybrid models. All the models underperformed, with accuracies below average (50%). This is because CBOW and Skip-Gram models require huge amounts of data to capture the meanings of words according to their contexts Skip Gram models (Mikolov et al., 2013). In this study, datasets used for COM were relatively small. Also, these models are computationally expensive particularly in handling huge vocabularies. Due to this, the models suffer inefficiencies that cause sub-optimal model training on complex tasks (Levy & Goldberg, 2014).

Table 4. 26 Performance of Hybrid ML Techniques for CBOW + Window Size = 1

Average Accuracy of the Hybrid ML Models: CBOW + Window Size = 1				
Hybrid ML Model	**Dataset 1**	**Dataset 2**	**Dataset 3**	**Avg. Accuracy (%)**
MLP + LR	41.7	40.2	48.7	43.5
MLP + SVM	41.2	39.7	48.7	***43.2***
MLP + SGD	15.4	31.1	35.5	27.3
MLP + MNB	4.3	6.7	39.1	16.7
MLP + RF	41.5	39.9	48.7	43.4

Table 4. 27 Performance of Hybrid ML Techniques for CBOW + Window Size = 5

Average Accuracy of the Hybrid ML Models: CBOW + Window Size = 5				
Hybrid ML Model	**Dataset 1**	**Dataset 2**	**Dataset 3**	**Avg. Accuracy (%)**
MLP + LR	41.7	39.9	48.7	43.4
MLP + SVM	41.2	39.7	48.7	***43.2***
MLP + SGD	41.7	39.0	32.6	37.8
MLP + MNB	3.5	4.7	34.5	14.2
MLP + RF	41.5	38.8	48.7	43.0

Tables 4.28 and 4.29 present the performance of Hybrid ML Models when using Skip Gram feature extraction technique with window sizes of 1 and 5 respectively. The results show that MLP and SVM hybrid model had the highest accuracy among the hybrid models.

Table 4. 28 Performance of Hybrid ML Techniques for Skip gram + Window Size = 1

Average Accuracy of the Hybrid ML Models: CBOW + Window Size = 1				
Hybrid ML Model	**Dataset 1**	**Dataset 2**	**Dataset 3**	**Avg. Accuracy (%)**
MLP + LR	41.7	40.1	48.7	43.5
MLP + SVM	41.2	39.8	48.7	***43.2***
MLP + SGD	31.1	33.6	33.9	32.8
MLP + MNB	3.0	4.0	39.1	15.4
MLP + RF	41.5	39.0	48.7	43.1

Table 4. 29 Performance of Hybrid ML Techniques for Skip gram + Window Size = 5

Average Accuracy of the Hybrid ML Models: CBOW + Window Size = 5				
Hybrid ML Model	**Dataset 1**	**Dataset 2**	**Dataset 3**	**Avg. Accuracy (%)**
MLP + LR	41.7	40.1	48.7	43.5
MLP + SVM	41.2	39.4	48.7	***43.1***
MLP + SGD	41.2	19.3	34.5	31.7
MLP + MNB	3.7	4.9	39.1	15.9
MLP + RF	41.3	39.0	48.7	43.0

Statistical Test of Significance

Table 4.2 shows One-Way ANOVA test results from testing if there was a statistical difference in the accuracies of the hybrid machine-learning models applied to three different comparative opinion datasets using Count Vectorizer model and trigram features. The p-value (0.996) is above 0.05, signifying mild evidence against the null hypothesis and hence there is no significant statistical difference in the accuracies of the ML models across the three datasets.

Table 4. 30. One-Way ANOVA (Welch's)

One-Way ANOVA (Welch's)

	F	**df1**	**df2**	**p**
Accuracy	0.106	7	6.72	0.996

Tables 4.31 and 4.32 present the results on the accuracy of the hybrid ML models using TFIDF models for feature extraction on dataset 2, with bigrams and trigrams respectively. The results show that the SGD and RF hybrid outperformed other models in both cases. Its performance was followed closely by that of MLP and RF in the case of models that used a DL model as the base model in their architectures.

Table 4. 31 Performance of Hybrid ML Techniques for TFIDF + Bigrams for Dataset #2

Hybrid ML Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Runtime (ms)	Prediction Latency (ms)
MLP + DT	92.2	92.8	92.2	92.3	36.5	0.0
MLP + RF	92.4	93.0	92.4	92.5	42.0	0.0
MLP + SGD	92.4	92.9	92.4	92.5	33.9	0.0
MLP + SVM	92.1	92.6	92.1	92.2	35.1	0.1
SGD + DT	92.6	92.7	92.6	92.6	1.1	0.0
SGD + MLP	92.3	92.7	92.3	92.4	28.6	0.0
SGD + RF	93.0	93.4	93.0	93.0	7.9	0.0
SGD + SVM	92.9	93.2	92.9	92.9	2.9	0.1

Table 4. 32 Performance of Hybrid ML Techniques for TFIDF + Trigrams for Dataset #2

Hybrid ML Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Runtime (ms)	Prediction Latency (ms)
MLP + DT	92.2	92.8	92.2	92.3	59.9	0.0
MLP + RF	92.4	92.8	92.4	92.5	65.9	0.1
MLP + SGD	92.4	93.0	92.4	92.5	55.0	0.0
MLP + SVM	91.6	92.1	91.6	91.6	63.4	0.2
SGD + DT	92.1	92.4	92.1	92.1	1.4	0.0
SGD + MLP	92.2	92.8	92.2	92.2	55.5	0.0
SGD + RF	93.8	94.0	93.8	93.8	11.7	0.1
SGD + SVM	92.8	93.2	92.8	92.8	3.4	0.2

Figure 4.5 shows a bar chart of the accuracy of the hybrid ML models for TFIDF feature extraction technique and Trigrams (nn = 3) on Dataset 2.

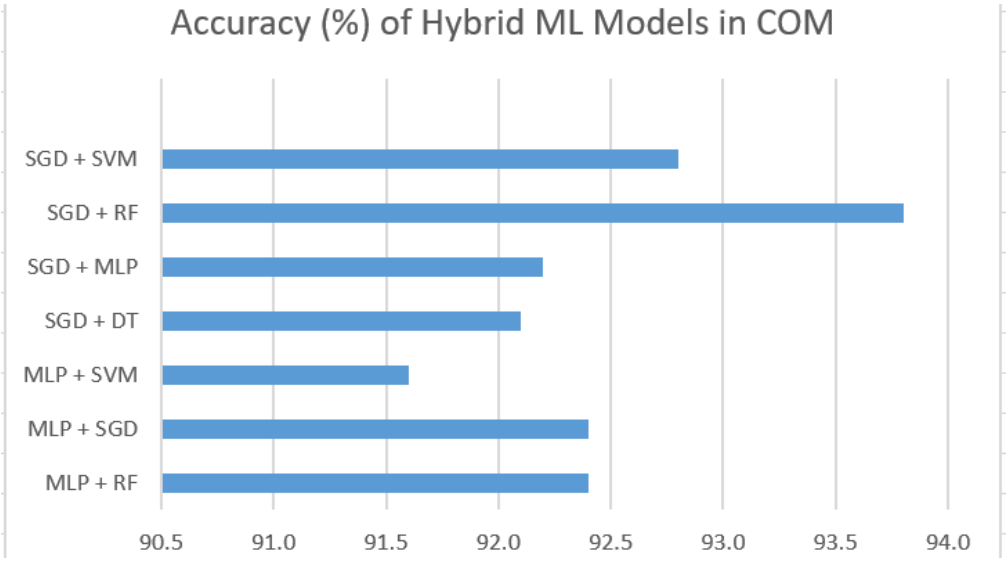

Figure 4. 3 Accuracy of the Hybrid ML Models: TFIDF + Trigrams; D#2

Table 4. 33 Best Performing Hybrid ML Model for Comparative Opinion Mining

Classifier	Accuracy (%)			F1-Score (%)			Averages (%)	
Hybrid ML Technique	D1 (Microsoft vs Google)	D2 (Facebook vs Twitter)	D3 (Pearl vs Marriott)	D1 (Microsoft vs Google)	D2 (Facebook vs Twitter)	D3 (Pearl vs Marriott)	Accuracy	F1-Score
MLP + DT	86.0	92.0	98.4	86.1	92.0	98.3	92.1	92.1
MLP + RF	*86.6*	*92.6*	*99.7*	*86.8*	*92.6*	*99.7*	*93.0*	*93.0*
MLP + SGD	86.9	92.3	91.8	87.1	92.3	91.6	90.3	90.3
MLP + SVM	85.6	91.9	99.7	85.8	91.9	99.7	92.4	92.5

In determining the most effective hybrid machine-learning model for comparative opinion mining, the researcher applied accuracy and f1-scores as the model evaluation metrics recommended for comparative opinion mining. Based on Table 4.33, the best performing hybrid ML model was MLP + RF with an accuracy of 93.0% and F1-score of 93.0%. The minimum accuracy and f1-score was 90.3% across three datasets. This is a reliable performance that is suitable for use cases like brand reputation monitoring (Ondara et al., 2023).

Table 4. 34 Performance of Top Two Hybrid ML Models for N-Gram Range 1 - 3

Hybrid ML Model	N-Grams	CV Accuracy (%)	TFIDF Accuracy (%)	Avg Accuracy (%)	CV F1-Score (%)	TFIDF F1-Score (%)	Avg F1-Score (%)
MLP + RF	Unigrams	92.0	91.1		91.1	91.4	
	Bigrams	92.8	92.6		92.6	92.7	
	Trigrams	93.0	92.7		92.7	92.8	
		Avg = 92.6	Avg = 92.1	**92.4**	Avg = 92.6	Avg = 92.3	**92.5**
SGD + RF	Unigrams	91.3	91.1		91.5	91.1	
	Bigrams	90.8	93.1		90.3	93.2	
	Trigrams	92.3	93.4		92.3	93.5	
		Avg = 91.5	Avg = 92.5	**92.0**	Avg = 91.4	Avg = 92.6	**92.0**

To conclude this section, the summary table 4.34 shows four key things:

i. That the MLP + RF hybrid machine learning model outperformed the SGD + RF hybrid model on average across the n-gram range of 1 to 3.

ii. The difference between the performance of MLP + RF and SGD + RF is insignificant (see Table 4.44 for T-test results). One can use either hybrid ML models for COM.

iii. MLP + RF has a slightly higher accuracy than SGD + RF when using Count Vectorizer technique for feature extraction. On the other hand, SGD + RF has a slightly higher accuracy than MLP + RF when using TFIDF feature extraction technique.

iv. In both MLP + RF and SGD + RF, the highest performance was recorded when using trigrams (n-gram with a range of 3).

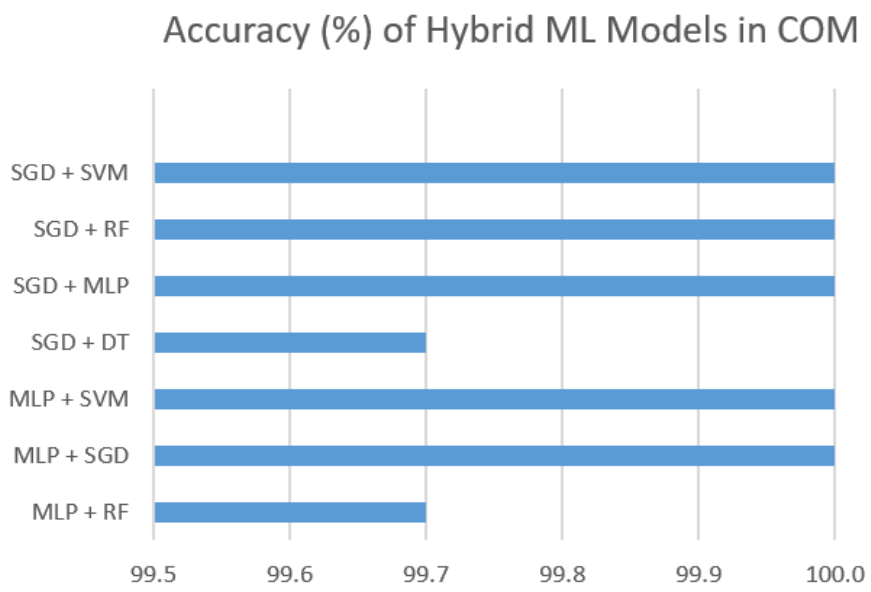

Figure 4. 4 Accuracy of the Hybrid Machine Learning Models: TFIDF + Trigrams: D#3

Statistical Test of Significance

Table 4.35 shows One-Way ANOVA test results from testing if there was a statistical difference in the accuracies of the eight hybrid ML models when applied to three different comparative opinion datasets using Count Vectorizer model and trigram features. The p-value (1.00) showing a lack of evidence against the null hypothesis. This suggests that there is no significant statistical difference in the accuracies of the ML models across the three datasets hence the researcher accept the null hypothesis. This means that any of the eight Hybrid ML Models could be used without a significant loss in the accuracy from opinion classification.

Table 4. 35. One-Way ANOVA (Welch's)

One-Way ANOVA (Welch's)

	F	df1	df2	p
Accuracy	0.0412	7	6.84	1.000

4.2.3.2 Reliable Brand Reputation Indicator

The results shown in this section are the screenshots from the web-based prototype system that implemented the developed hybrid model for performing COM using live data from X platform and YouTube. The results are aimed at demonstrating that the hybrid ML model was capable of performing COM using primary / live data. The results for brands whose data were easily available on X platform and/or YouTube for the prototype to use in monitoring brand reputation using the developed hybrid ML model that served as the engine for the prototype.

Telecommunications Brands

The reputation of a brand is directly related with the sentiment positivity at any given time. Figure 4.7 and 4.8 shows that Safaricom has a higher opinion positivity compared with Airtel (a competitor in the Telecommunications sector). Safaricom is thus more preferred than Airtel.

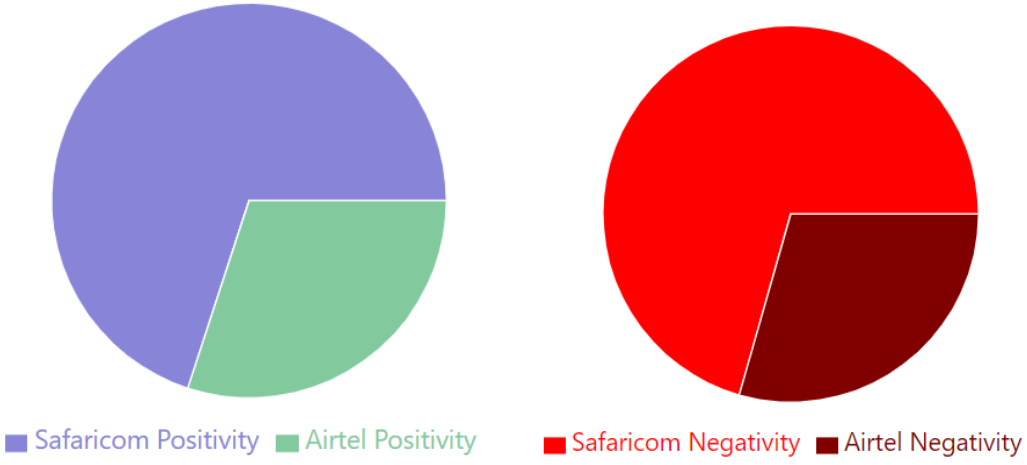

Figure 4. 5 Brand Positivity Comparison 1: Safaricom Vs Airtel

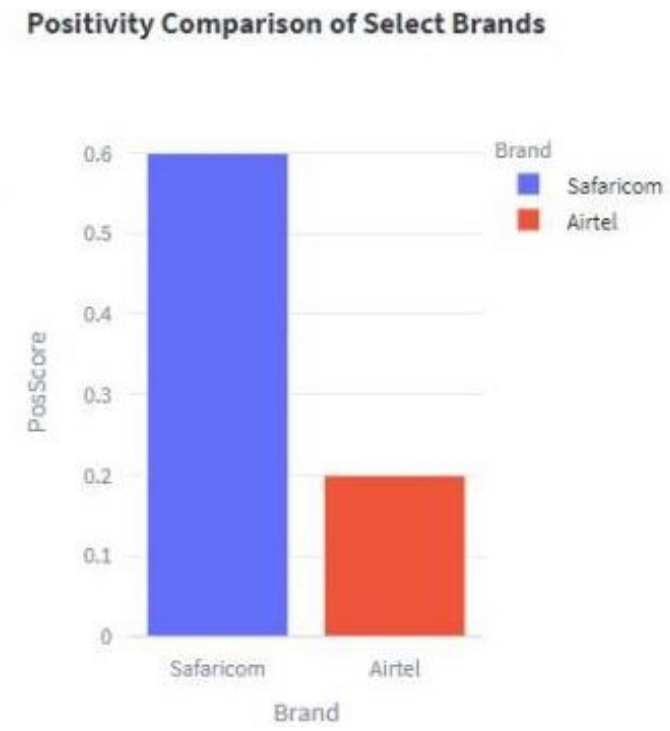

Figure 4. 6 Brand Positivity Comparison 2: Safaricom vs Airtel

Figure 4.9 shows the prototype's web interface. On this interface, the user has keyed in "Safaricom" as an entity / brand whose opinion / sentiment analysis is needed. The horizontal bar chart shows the distribution of tweets according to sentiment classes, while showing the statistics in terms of the number of tweets classified as positive, negative, and neutral. In this case, Safaricom has more positive tweets than negative and neutral tweets.

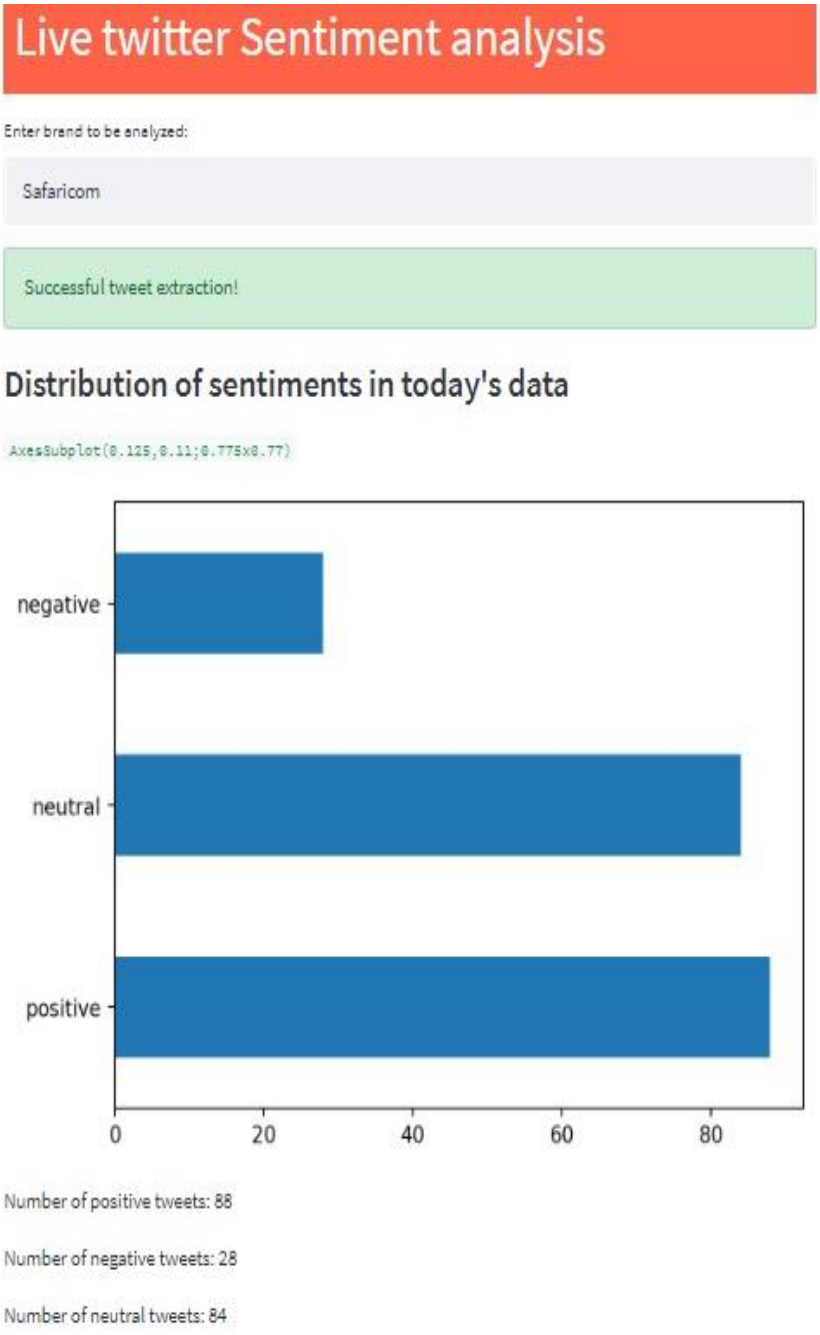

Figure 4. 7 Brand Reputation: Safaricom

Figure 4.10 shows the prototype's web interface. On this interface, the user has keyed in "Airtel" as an entity / brand whose opinion / sentiment analysis is needed. The horizontal bar chart shows the distribution of tweets according to sentiment classes, while showing the statistics in terms of the number of tweets classified as positive, negative, and neutral. In this case, Airtel has more positive tweets than both negative and neutral tweets.

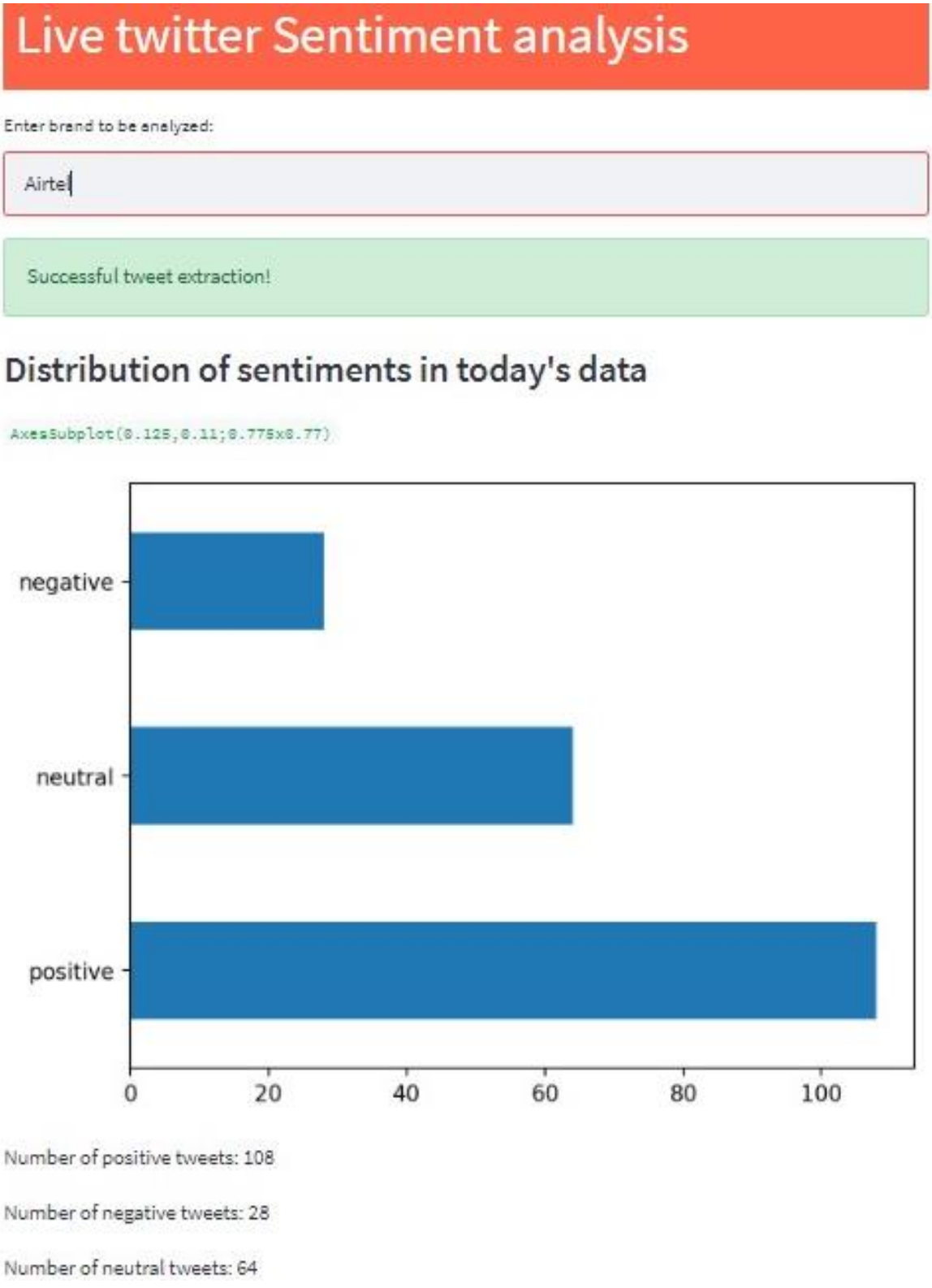

Figure 4. 8 Brand Reputation: Airtel

Banking Brands

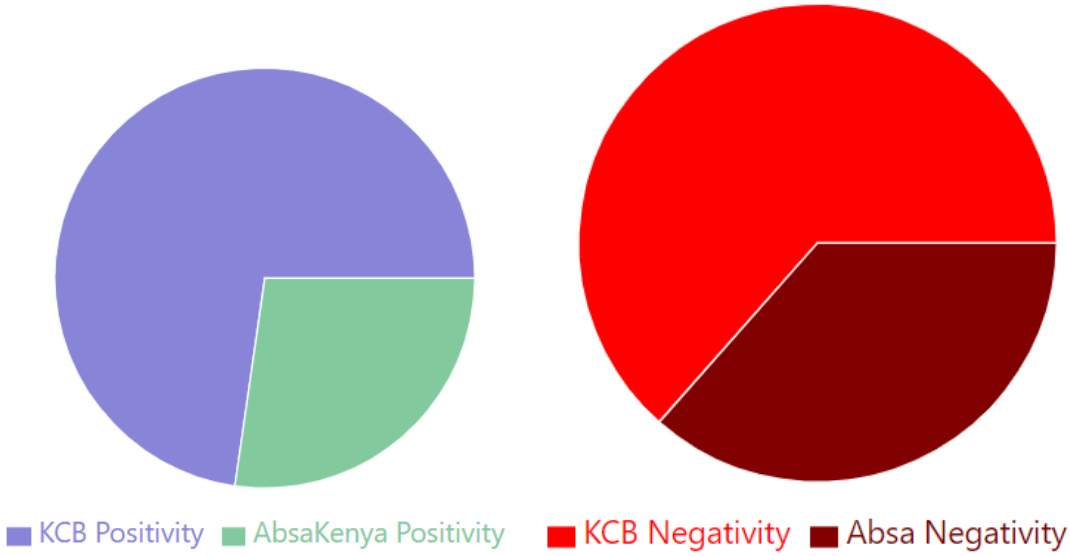

Figure 4. 9 Brand Positivity Comparison 1: KCB Vs ABSA

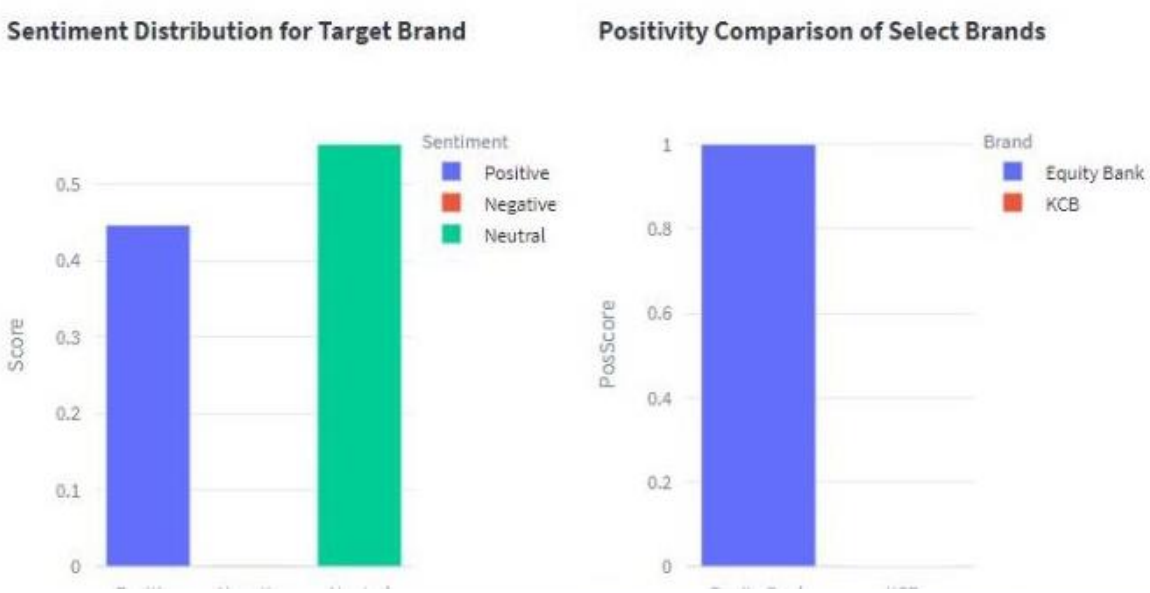

Figure 4. 10 Brand Positivity Comparison 2: Equity Bank Vs KCB

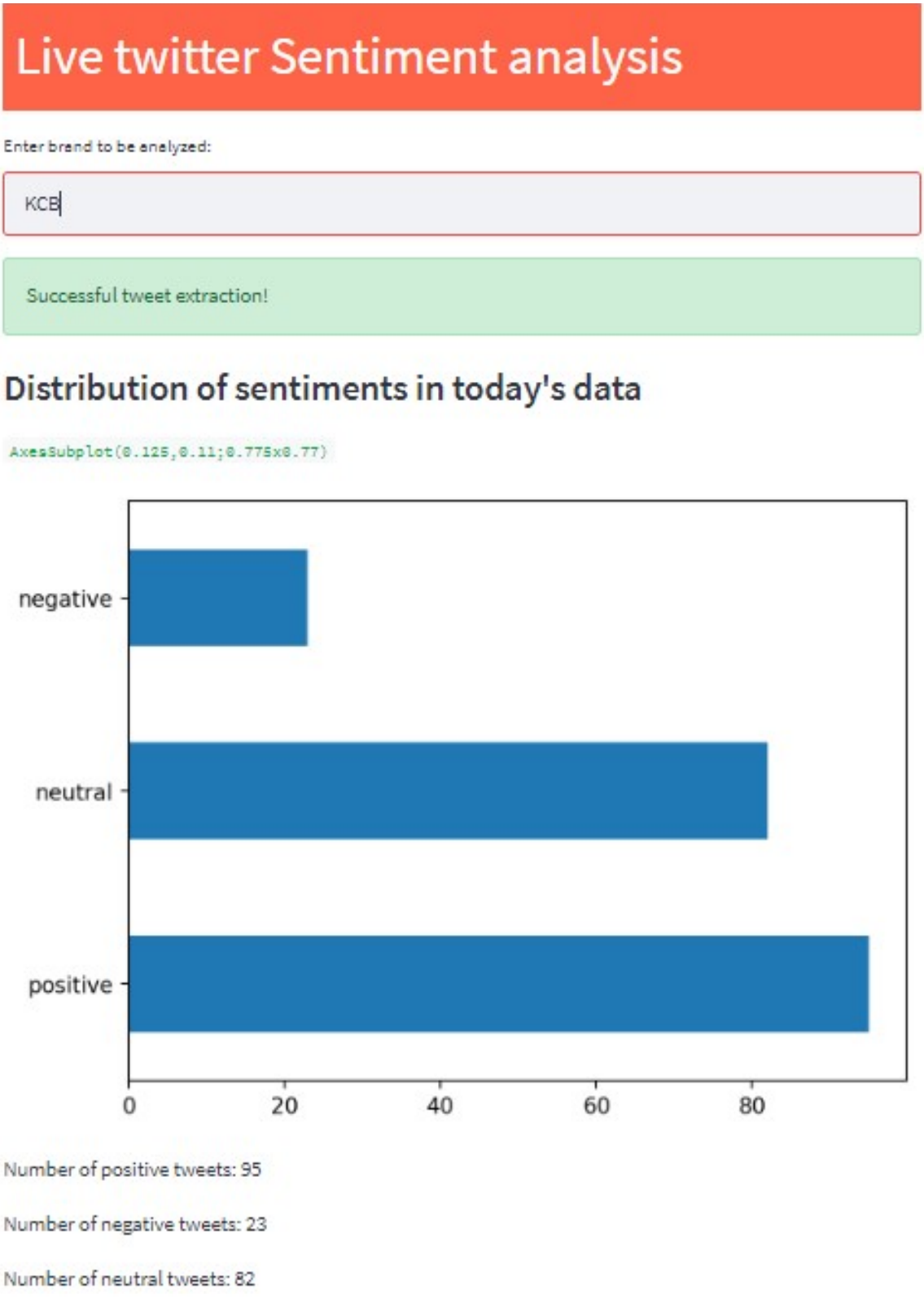

Figure 4. 11 Brand Reputation: KCB

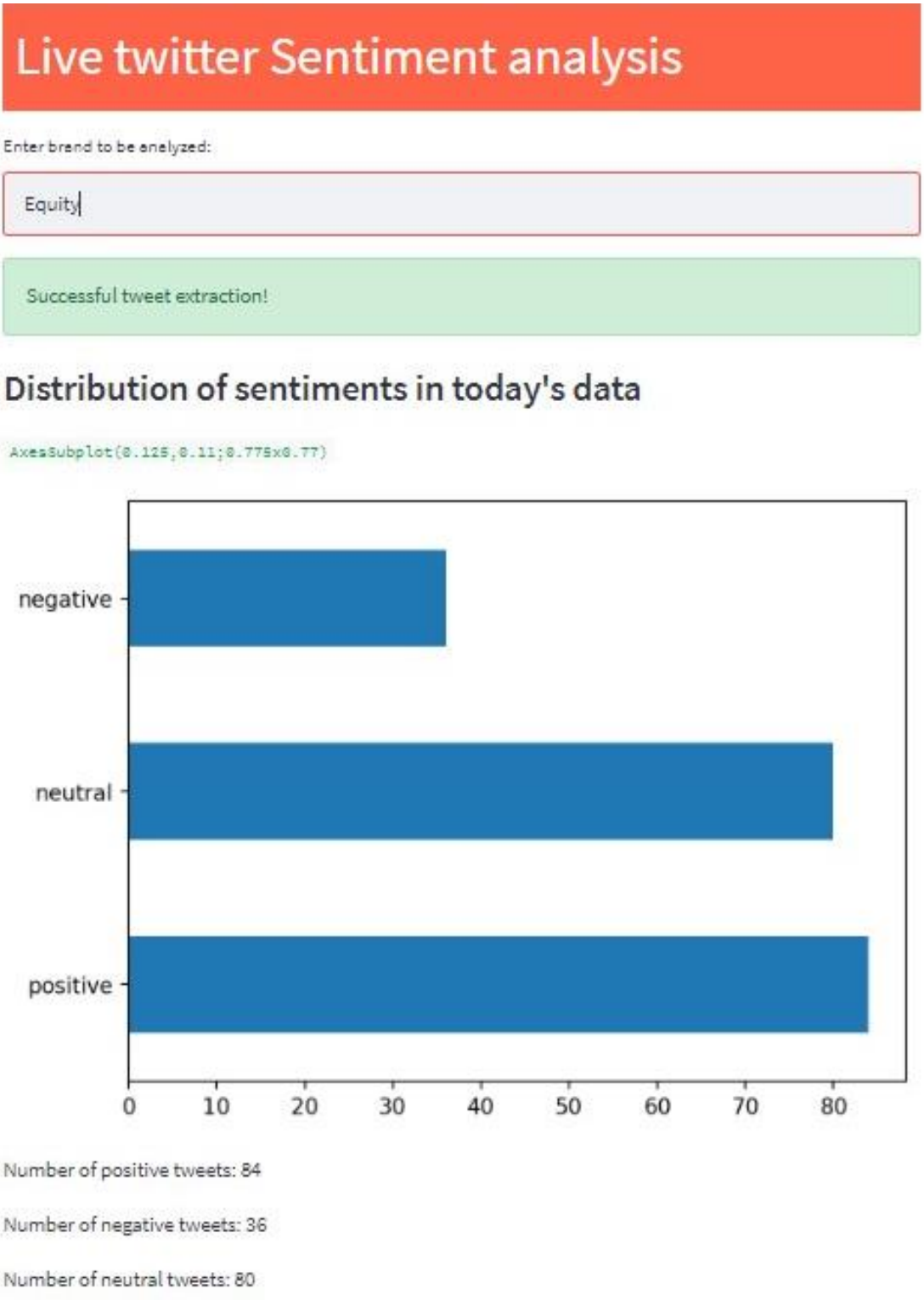

Figure 4. 12 Brand Reputation: Equity Bank

Smartphone Brands

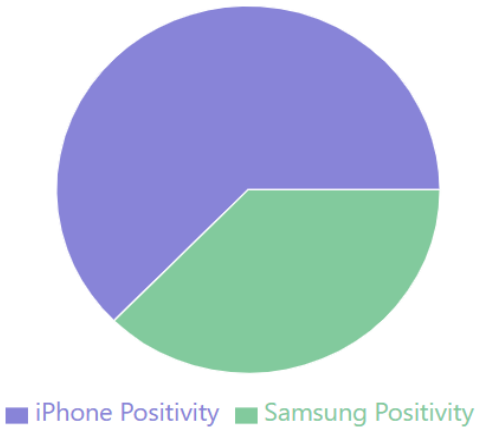

Figure 4. 13 Brand Reputation Comparison: iPhone Vs Samsung

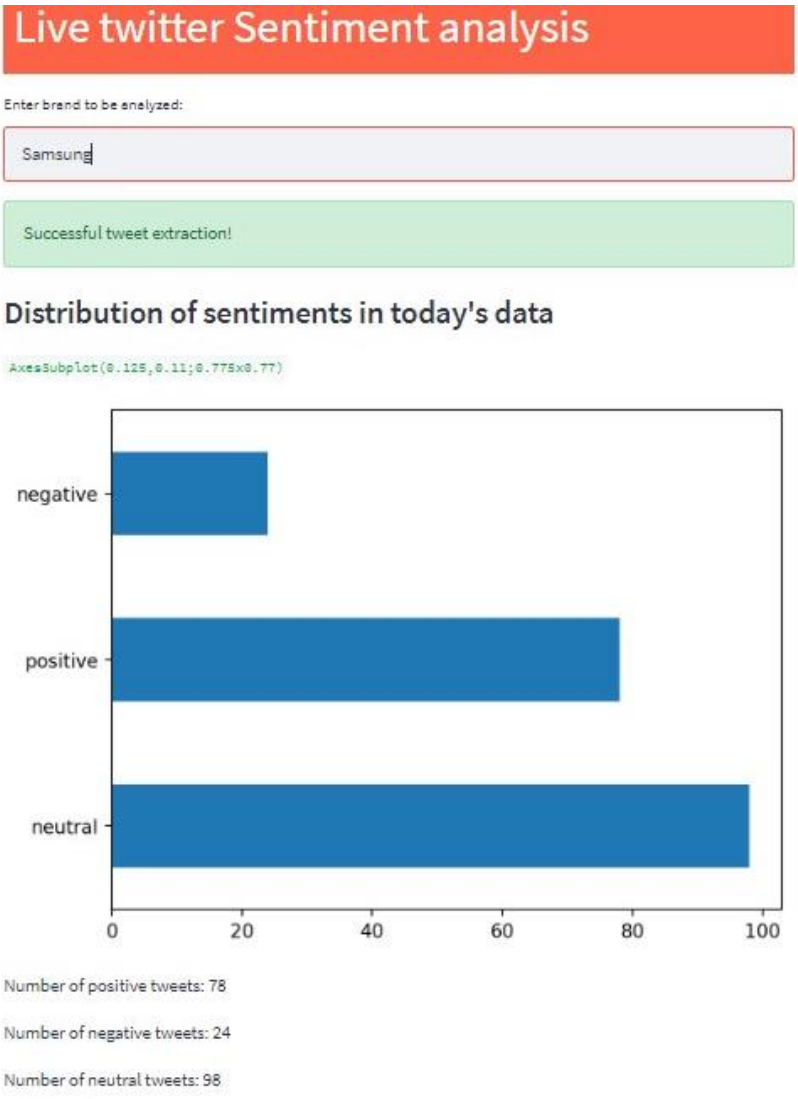

Figure 4. 14 Brand Reputation: Samsung

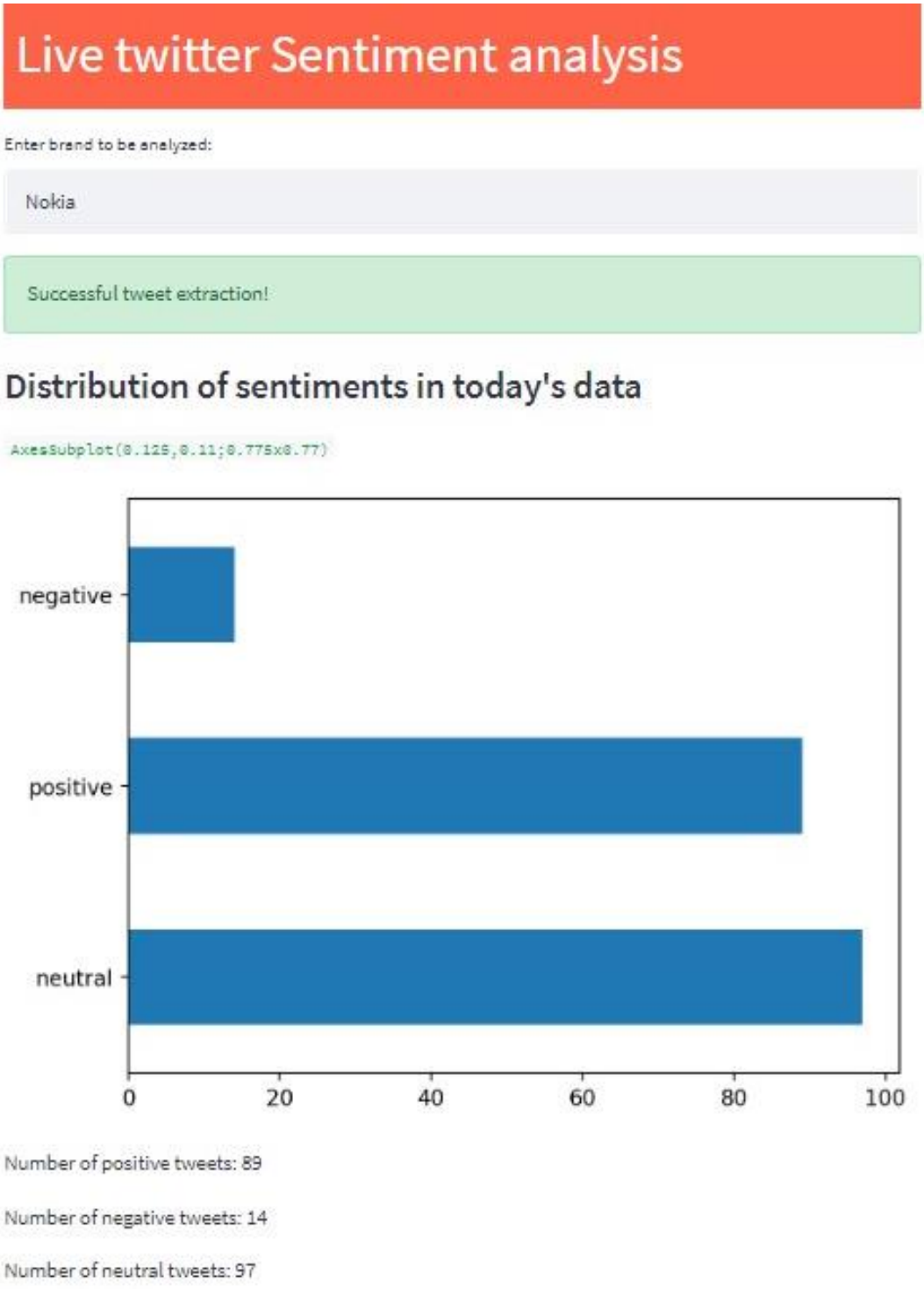

Figure 4. 15 Brand Reputation: Nokia

Higher Education Brands

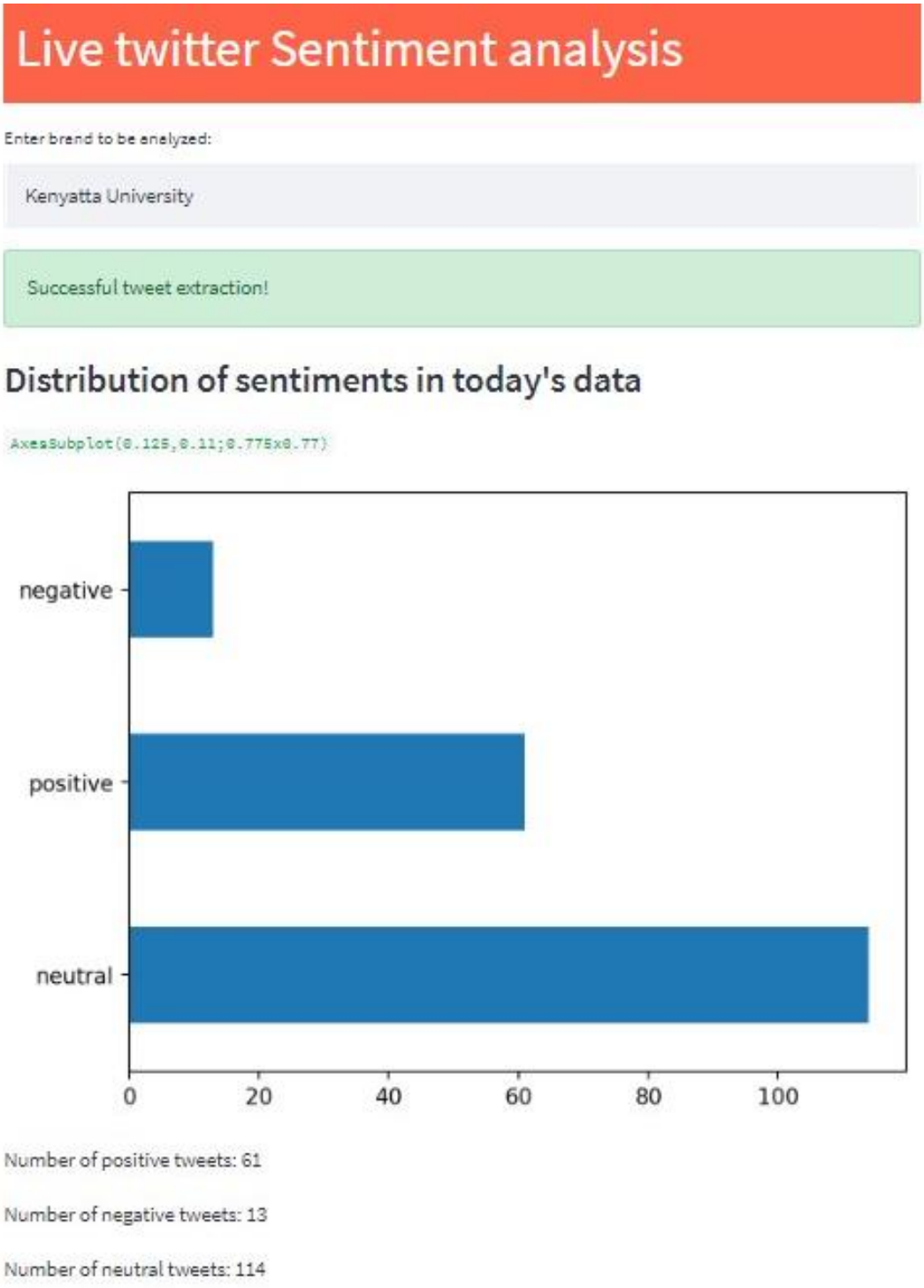

Figure 4. 16 Brand Reputation: Kenyatta University

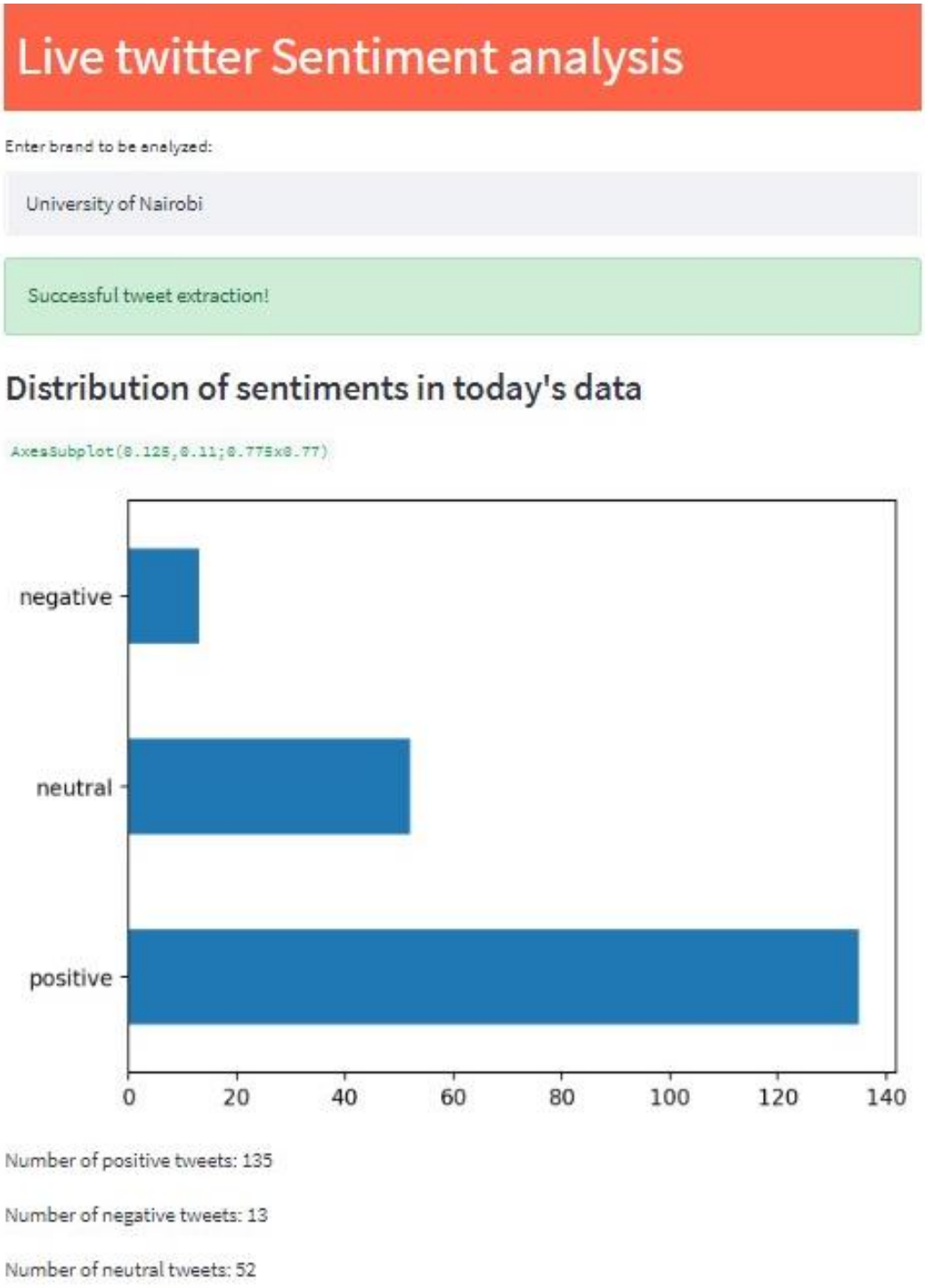

Figure 4. 17 The University of Nairobi

4.2.3.3 Insight into Brand Aspects

Brand aspects are important in gaining competitive advantage. Even though detailed brand aspect mining was out of the scope of this study, the researched tried to perform basic mining to demonstrate the capability of the hybrid model in opinion mining. A sentiment word cloud was used to generate common terms mentioned in the reviews or texts analyzed. Figure 4.20 - 4.22 are screenshots showing some of the raw outputs. The words or terms output by the model are good pointers to what a brand manager could use to identify what most opinion holders were positive or negative about concerning their brand's products, services, or general business. The column on aspects represents potential brand features (aspects) for use in brand reputation monitoring.

Figure 4. 18 Brand Aspects: Telecommunications Domain

Figure 4. 19 Brand Aspects: Banking Domain

Figure 4. 20 *Brand Aspects: Roma vs Porsche*

4.2.3.4 Brand Reputation Trends

Long-term trends in reputation is based on historical data and it helps to capture long-term opinion trends (Das et al., 2018). The reputation of a brand over a period can be traced using the sentiment positivity trend.

Figure 4. 21 Telecommunications Brand Reputation Monitor: Daily Trends

Figure 4. 22 Banking Brand Reputation Monitor: Daily Trends

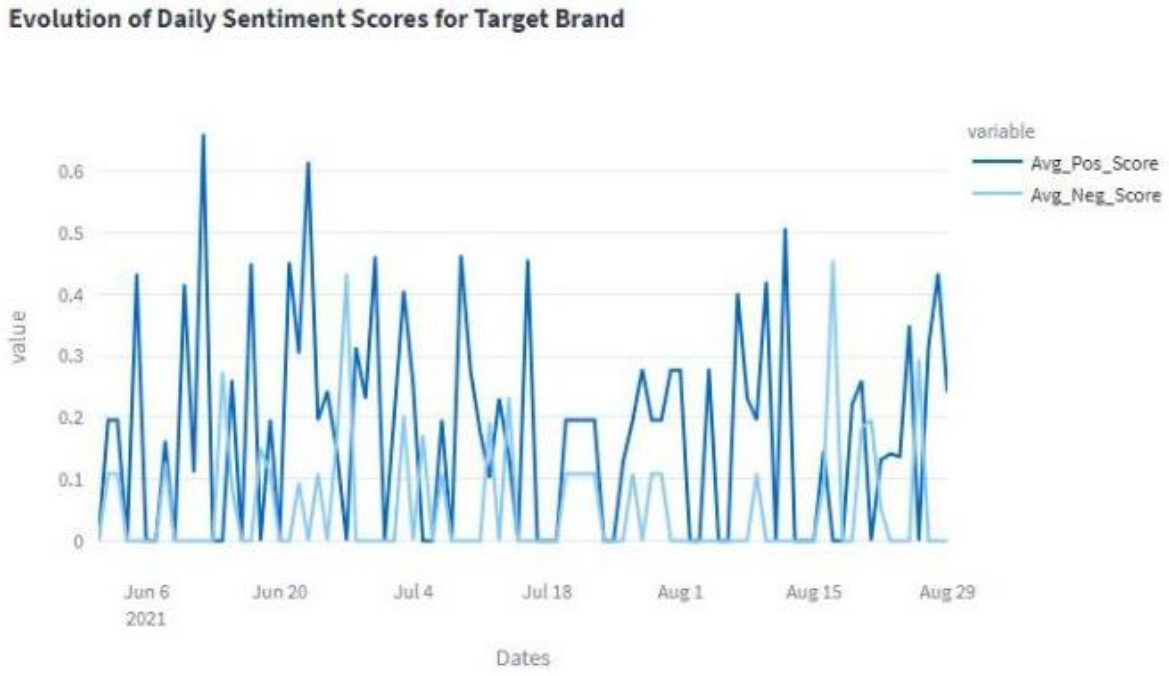

Figure 4. 23 Smartphone Gadgets Brand Reputation Monitor: Daily Trends

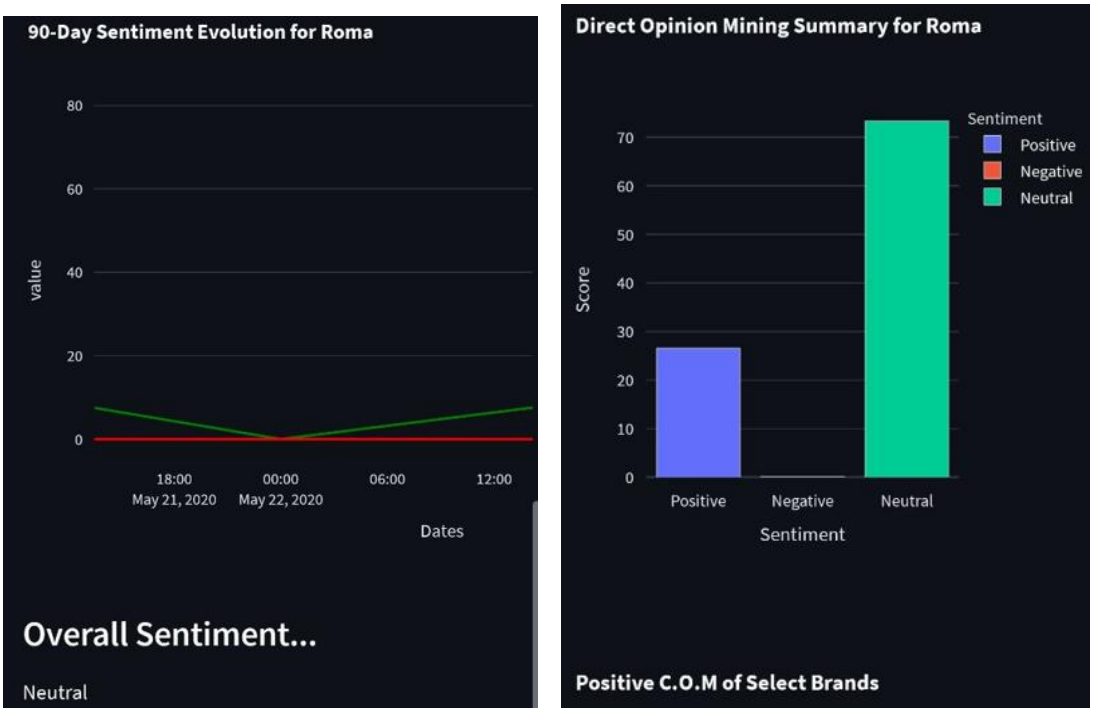

Figure 4. 24 *Overall Brand Reputation for Roma: DOM vs COM*

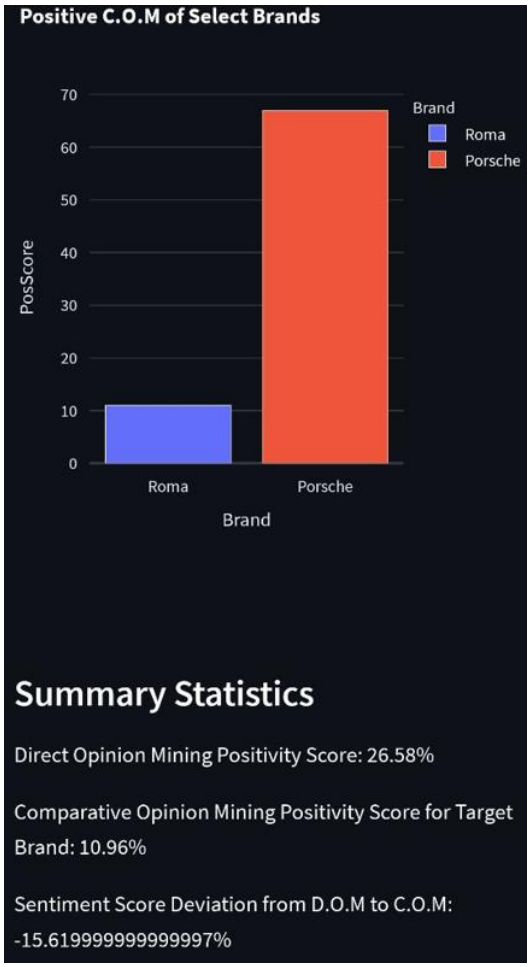

Figure 4. 25 *Brand Reputation for Roma Vs Porsche: DOM vs COM*

4.2.3.5 Ground Truth (Results)

From the 100 reviews classified by human experts in the ground truth experiment, it was observed that the three human classifiers agreed in the correct classification 95 of the 100 reviews. That is, they could not agree on the correct opinion classes for six of the reviews. On the other hand, the hybrid machine-learning model (MLP + RF, in this case) made 93 predictions accurately. The variance between the model's prediction accuracy and the ground truth (human classification) was 2%, an insignificant variance between human classifiers and machine classifiers through the hybrid machine-learning model. This indicated the hybrid model's prediction accuracy is reliable and satisfactory for use. The best hybrid machine-learning models developed in this experiment attained 93% accuracy.

For model training, the researcher used three datasets that had been used in carrying out comparative opinion mining experiments. However, this study focused on using one machine-learning algorithm or model while our study focused on using hybrid machine learning models. The three datasets were published at https://www.kaggle.com/umairyounis/comparative-reviews-datasets. To ascertain the credibility of the datasets, the research had three human experts to annotate the datasets before using them for model training and validation. The Kappa (K) score for this annotation exercise was 0.81, indicating a strong agreement among the three human experts involved in annotating the data. The details of the datasets are in Table 4.32 as was published in a journal article "Hybrid Machine Learning Techniques for Comparative Opinion Mining" (Ondara et al., 2023)

Table 4. 36 Primary Data for Model Validation

Dataset	Reviews	pos_pos	pos_neg	pos_neu	neg_neg	neg_pos	neg_neu	neu_neu	neu_pos	neu_neg
Microsoft vs Google	3011	360	1268	396	62	380	46	321	148	30
Facebook vs Twitter	3000	440	1208	447	54	307	59	310	143	32
Pearl Continental vs Marriott	1012	276	138	46	92	138	46	138	92	46

From a total of 207 records taken from one of the primary datasets collected and annotated by human experts before the same raw dataset was fed into MLP + RF hybrid model, the agreement level between the humans and this hybrid ML model for comparative opinion mining was 64%. This percentage was computed using the first 100 records in the dataset, followed a random sampling strategy to avoid bias (Kothari, 2015). Table 4.37 shows a sample of classification results comparing how the human classifiers and hybrid model classifier labelled 10 comparative reviews. The label "neut_neut" shows that the opinion classes for entity 1 and entity 2 were both neutral. For pos_pos, the opinion labels for both entity 1 and 2 were positive. For neg_neg, the opinion classes for entities 1 and 2 were both negative. Thus, the X_Y format could be used to ascribe the opinion classes, where X represents the opinion class for entity 1 and Y for the opinion class for entity 2.

Table 4. 37 Sample Classification Results (Human Classifier vs. Hybrid Model Classifier)

SN	Human Classifier Label	MLP + RF Hybrid Classifier Label	Agreement
1	neut_neut	neut_neut	Yes
2	pos_pos	pos_pos	Yes
3	neut_pos	neut_pos	Yes
4	neut_pos	neut_neut	No
5	neut_neut	neut_neut	Yes
6	neut_pos	neut_pos	Yes
7	neut_pos	neut_pos	Yes
8	neg_neut	neut_pos	No
9	pos_pos	neut_neut	No
10	neut_pos	neut_pos	Yes

According to Cohen's Kappa statistic, this agreement percentage represents substantial agreement level (McHugh, 2012) between the human classifier and the hybrid machine-learning model. Therefore, the hybrid model was found to be a satisfactory replacement of human classifiers in the task of comparative opinion mining. The benefits of computer automation in opinion mining include achieving high opinion classification efficiency and effectiveness and opinion classification reliability.

With automation, brand managers may get real-time brand reputation reports that would help them make informed decisions regarding the performance of their brands in relation to competitor brands over time. Table 4.28 shows a sample of 10 records classified by a human expert and the MLP + RF hybrid machine-learning model for comparative opinion mining. The classification of the opinion labels is based on two entities for each comparative review or text (tweets and YouTube comments).

For the developed hybrid model, the opinion class indicates a brand reputation using the matrix show in table 4.29. Thus, it was possible to use the model in monitoring the reputation of a target brand, relative to the reputation of competitor brands through comparative opinion mining using available data on platforms like X and YouTube.

Table 4. 38 Mapping of the Hybrid ML Model's COM Results to Brand Reputation

Model Output	*Brand A Reputation*	*Brand B Reputation*
Pos_Pos	Positive	Positive
Pos_Neg	Positive	Negative
Pos_Neut	Positive	Neutral
Neg_Pos	Negative	Positive
Neg_Neg	Negative	Negative
Neg_Neut	Negative	Neutral
Neut_Pos	Neutral	Positive
Neut_Neg	Neutral	Negative
Neut_Neu	Neutral	Neutral

4.2.4 Hypothesis Testing

This section is dedicated to testing the five hypotheses in Section 1.6

Test 1: Differences in the Accuracy of Hybrid ML Models versus Single ML Models

$H0_1$: *There are no statistically significant performance differences between the hybrid ML models and the single ML models in COM.*

HA_1: *There are statistically significant performance differences between the hybrid ML models and the single ML models in COM.*

Table 4.39 shows a p-value of <0.05, implying the presence of statistically significant performance differences between Hybrid ML models and Single ML models, hence we accept the alternative hypothesis while rejecting the null hypothesis. A posthoc analysis on this showed that this significant differences were between KNN single model and the following hybrid models: MLP + DT, MLP + RF, MLP + SGD, MLP+SVM, SGD + MLP, SGD + RF, and SGD+SVM. The posthoc analysis report is in Appendix V.

Table 4. 39 Statistical Differences in the Performance of Single ML Models in COM

ANOVA - Accuracy

	Sum of Squares	df	Mean Square	F	p
ML Model	3385	15	226	2.10	0.018
Residuals	8600	80	107		

Test 2: Differences in the Hybrid ML Models Performance in COM

***$H0_2$**: There are no statistically significant performance differences in COM between the different hybrid ML model.*

***HA_2**: There are statistically significant performance differences in COM between the different hybrid ML models in COM.*

From Table 4.40, it can be seen that the p value is 0.977, implying that there are no statistically significant performance differences between the different Hybrid ML models. There was no need for posthoc analysis in this case as the p-value was above 0.05. This shows that the hybrid ML models have accuracies with no major deviations from each other.

Table 4. 40 Statistical Differences in the Performance of Single Models vs Hybrid Models

ANOVA - Accuracy

	Sum of Squares	df	Mean Square	F	p
ML Model	111	7	15.9	0.225	0.977
Residuals	2823	40	70.6		

Test 3: Differences in the Hybrid Models Performance in COM Based on Dataset Choice

***$H0_3$**: There are no statistically significant performance differences in COM between the Hybrid ML models based on the choice of dataset.*

***HA_3**: There are statistically significant performance differences in COM between the Hybrid ML models based on the choice of dataset.*

Table 4.41 shows a p-value of <0.01, implying that there are statistically significant performance differences in the ML models based on the choice of datasets. A posthoc analysis on this result revealed the significant difference was between datasets 1 (D1) and 3 (D3).

Table 4. 41 Stat. Diff. in Models' Performances Based on Dataset Choice.

ANOVA - Accuracy

	Sum of Squares	df	Mean Square	F	p
Dataset	1971	2	985.3	46.5	< .001
Residuals	932	44	21.2		

Test 4: Differences in the Hybrid Models Performance in COM Based on Feature Extraction Technique

***$H0_4$**: There are no statistically significant performance differences in COM between the Hybrid ML models based on the choice of feature extraction technique.*

***HA_4**: There are statistically significant performance differences in COM between the Hybrid ML models based on the choice of feature extraction technique.*

Table 4.42 shows a p-value of <0.01, implying that there are statistically significant performance differences in the various Hybrid ML models based on the choice of datasets. From the posthoc analysis (Appendix V), the difference is due to statistically significant differences between CV3 and CBOW5, CV3 and SkipGram; and TFIDF3 and CBOW5 and TFIDF3 and SkipGram. This shows that Count Vectorizers and TFIDF models for feature extraction have a significant performance difference. Similarly, the performance of Hybrid ML Models using Count Vectorizers differs significantly from those models using SkipGram feature extraction techniques.

Table 4. 42 Hybrid Models Performances Based on Feature Extraction Techniques .

ANOVA - Accuracy

	Sum of Squares	**df**	**Mean Square**	**F**	**p**
FET	88508	3	29503	221	< .001
Residuals	18678	140	133		

Test 5: Interaction between ML Models and Datasets

This Two-way ANOVA test was carried to test if there was an interaction between ML models and the datasets chosen for COM. The results in Table 4.43 show that there was no significant interaction between ML models and datasets in COM. This means that any ML model could be used with any dataset without significantly affecting the performance of the model in COM.

Table 4. 43 Interaction Between ML Models and Datasets

ANOVA - Accuracy

	Sum of Squares	df	Mean Square	F	p
ML Model	30039	15	2002.6	2.7326	0.002
Dataset	4896	2	2448.1	3.3405	0.040
ML Model ✻ Dataset	1387	30	46.2	0.0631	1.000
Residuals	70355	96	732.9		

Test 6: Interaction between Datasets and Feature Extraction Techniques

This Two-way ANOVA test was carried to test if there was an interaction between datasets and feature extraction techniques in COM. Table 4.44 shows that there was no significant interaction between Feature Extraction techniques and datasets. Thus, any feature extraction technique can be used on any dataset without significantly affecting model performance.

Table 4. 44 Interaction Between ML Models and Datasets

ANOVA - Accuracy

	Sum of Squares	df	Mean Square	F	p
Feature ET	88508	3	29502.6	312.90	< .001
Dataset	3723	2	1861.3	19.74	< .001
Feature ET ✻ Dataset	827	6	137.8	1.46	0.196
Residuals	12446	132	94.3		

4.3 Summary of Key Findings

i. The performance of different ML models in COM varies depending on several factors including the feature extraction techniques and the datasets used.

ii. A tried-and-tested way of developing powerful hybrid machine-learning models is the ensemble learning method. The reason this works powerfully is that it leverages the strengths of each of the individual models that make up the hybrid-learning model.

iii. There are statistically significant differences in the performance of single ML models and hybrid ML models in COM.

iv. There are no statistically significant performance differences in COM between different hybrid ML models in performing COM.

v. There are statistically significant performance differences in COM between different machine learning models based on the choice of datasets.

vi. There is a statistically significant differences in the performance of machine learning models based on the choice of feature extraction techniques.

vii. Hybrid ML models developed in this study outperformed the single ML models from which they were developed. This confirms the fact that hybrid models often outperform single models. Testing this model via a prototype proved that the hybrid ML models for comparative opinion mining are effective in brand reputation monitoring.

viii. The best performing hybrid models (MLP + RF and SGD + RF) both outperform the single models from which they were created, confirming the superior performance of hybrid models over single models.

ix. Deep Learning models performed well. In many cases, they performed better than machine learning models. However, it is more challenging integrating them with machine learning models to develop hybrid architectures. The MLP model is easier to integrate with machine learning models.

x. Dense vectorizers likes CBOW and Skip-Gram do not perform well with small datasets.

CHAPTER FIVE

DISCUSSION

5.1 Introduction

This section presents a discussion of the results, organized according to the three research questions and four hypotheses that guided this study.

RQ1: How do the existing machine learning models for comparative opinion mining compare in their performance?

RQ2: How can a hybrid machine-learning model for comparative opinion mining be developed based on the experimental results?

RQ3: How effective would the hybrid machine-learning model be in performing comparative opinion mining?

5.2 Research Question 1 (RQ1)

How do the objectively selected, existing machine learning models for comparative opinion mining compare in their performance?

The existing models used in comparative opinion mining have varying performances based on their architectural designs, features extracted, and parameters implemented. From this research, most of the existing models for comparative opinion mining consist of single machine learning algorithms. As such, this research found scanty details about the application of hybrid models in performing comparative opinion mining. The single models vary in their performance. For instance, in terms of average accuracy across three datasets, the best performing single model was Stochastic Gradient Descent in the

case of conventional machine learning models. The multilayer perceptron was the best performing deep learning model in this study.

The least performing single model out of these two groups was the K-Nearest Neighbor. This finding is similar to one reported by Younis et al. (2020), who also found K-Nearest Neighbor to have performed the least while Random Forest performed the best in an experiment testing a set of seven single classifiers on average across three datasets. The KNN is a lazy model. By default, it uses the Euclidean distance, which presents a challenge when working with noisy features and high dimensions' data. These two factors often negatively affect the performance of the model. To solve this performance challenge, the value of K needs to be optimized.

In this study, the Random Forest model served as the final estimator in the hybrid architecture. The choice of this model was because it was the second best performing single machine-learning model in the class of machine learning algorithms in this study. The model is easy to use and probably the most flexible algorithm with a high performance on classification tasks because it is an ensemble model consisting of multiple decision trees resulting in improved model generalization (Younis et al., 2020). In Section 2.11, three of the six existing hybrid machine-learning models described also used the Random Forest models as the top-level model. The analysis of the findings from previous studies on hybrid models as well as the findings from this study reveal the power of the Random Forest classifier, making it a good choice in developing hybrid models for classification tasks besides other machine learning tasks where it has been applied in various studies.

Moreover, the finding that the Random Forest model is a powerful classifier is in agreement with that by Younis et al. (2020), who implemented the same algorithm for comparative opinion mining, getting an average accuracy of 92% while in our study, the model had an average accuracy of 84.8%. However, their experiment involved one dataset with which the performance was 100%, an indication of model overfitting. High data dimensionality, sparse and noisy features, and a small amount of labelled data may cause model overfitting. Contrary to our study, the researcher did not include results associated with model overfitting. Discounting the dataset where model overfitting was suspect, their average accuracy using the Random Forest model was 88.3%, which is close to the empirical results reported in this study. This study finds the random forest model a good choice for top-level models in hybrid models. The problem of model overfitting is addressed using various techniques. Specifically, this study employed feature selection and ensemble learning. The results show that there was no statistically significant difference in the performance of single machine learning models. This is because every machine learning or deep learning algorithm has its strengths and weaknesses for solving particular problems such text classification. Their performance may not be optimal, lacking the benefits of hybridization (Banihashemi et al., 2017).

5.3 Research Question 2 (RQ2)

How can a hybrid machine-learning model for comparative opinion mining be developed?

From many publications in opinion mining, a conventional hybrid algorithm would consist of a machine-learning model combined with a lexical model. The lexical model would provide features that the machine-learning model would use to perform classification. Recent studies have witnessed a rise in developing hybrid machine-

learning models for application in different areas including general text classification. The researcher, however, found limited publications on the development and application of a hybrid machine learning based model in comparative opinion mining. Borrowing from existing research in support of developing ensemble learning hybrid models, this study was able to come up with a hybrid machine-learning model that employs two machine learning models in one case and a machine learning – deep learning model in another case. Both cases performed satisfactorily in terms of classification accuracy in comparative opinion mining, revealing the potential of hybrid models in comparative opinion mining. This approach for developing hybrid machine-learning models is in line with the approach in a study by Al Amrani et al. (2018) who developed a hybrid model consisting of SVM and RF using ensemble learning method. In both this studies, one model was used to extract features while the other model was used to perform opinion classification.

The agile development methodology as documented in Section 3.10.1 was followed in implementing the hybrid model as a prototype, which was used as a proof of concept in brand reputation monitoring. The researcher found this software development methodology is suitable in machine learning projects (Elman & Turk, 2017). However, there may be other software development methodologies suitable for this. In terms of applying the ensemble learning method during hybrid model development, this study, supported by evidence from other studies such as Jain et al. (2021), and Kim (2014). For the best performing hybrid model (MLP + RF), 32 hidden layers were defined for the multilayer perceptron. Further to this, the MLP performed well with a maximum iteration of 50 and a learning rate of 0.001.

5.4 Research Question 3 (RQ3)

How effective would the hybrid machine-learning model be for comparative opinion mining?

Research provides several reasons supporting our finding from this study that hybrid machine-learning models outperform single machine learning models. This is because a hybrid model leverages the strengths while diminishing the weaknesses of each single model (Bergstra & Bengio, 2012). Theoretically, a single machine-learning model A may have powerful feature extraction ability but less powerful classification accuracy. A different single machine-learning model B may have weak feature extraction ability but is powerful at classification. Creating a hybrid model consisting of A and B, with model A being used for feature extraction and model B used for classification should, theoretically, result in performance gains.

Since deep learning models are better than machine learning models in feature extraction, choosing a deep learning model as a base model in the hybrid architecture would benefit a hybrid model especially where huge datasets are used. Nonetheless, the difference in the performance of the pure hybrid machine learning models (e.g. SGD-RF) and hybrid machine learning - deep learning models (e.g. MLP-RF) was insignificant. Thus, one can use either hybrid approaches. The key finding here points to increased accuracy of comparative opinions classification when using hybrid models compared to when using single classification models. A study by Ondara and others (2023) confirm the suitability and effectiveness of hybrid machine learning models for use in comparative opinion mining, indicating the suitability of the models in brand reputation monitoring based on comparative opinion mining.

This finding agrees with the finding by Al Amrani and others (2018), that their SVM-RF hybrid model outperformed their single models (SVM and RF) by 3% in terms of classification accuracy. The least accuracy improvement the researcher observed in this study was 5.7% while the highest accuracy improvement due to hybridization was 7.3%, which is slightly higher than the 3% performance gain the study by Al Amrani et al. (2018). Other studies like those by Jain and others (2021), and Kim (2014), have also demonstrated the superior performance of hybrid models over single models in classification. Their results agree with those the researcher observed while carrying out this study. This show the generalizability of the hybrid machine-learning model for comparative opinion mining developed in this study.

For results presentation, web-based dashboards have been used in many past studies to capture brand mentions in a bid to help brands gauge their reputation online. However, our dashboard was developed purposely to handle comparative opinions, unlike most of the existing dashboards that handle direct opinions. The researcher supports the finding that comparative opinions communicate much more precise information regarding a target brand based on the aspects upon which the multiple entities are compared (Varathan et al., 2017). The visual aspects of the dashboard are an effective way of presenting results from a machine-learning based opinion-mining model. The display of brand mention statistics offers a quick view on the reputation of the brand based on the relationship between positive mentions and negative mentions. The reputation trend charted on the dashboard quickly reveals the performance of the brand's reputation over a specific period thus making it easier for brand managers and interested users to establish the trend of their brand's reputation for strategic decision-making.

Statistical tests (Table 4.43) showed that there was a significant statistical difference between the performance of the hybrid machine learning models and the single machine learning models. The hybrid models had, on average, higher accuracy and f1-score values. This confirms that the hybrid models are superior in performance compared to single models. This is because hybrid models leverage the strengths and reduce the limitations of the single models, resulting in improved classification accuracy and overall model performance (Tan et al., 2023). The researcher observed, however, that among the hybrid models, there was no statistically significant difference in performance, since their performance is relatively similar. Overall, for application of comparative opinion mining to brand reputation monitoring, hybrid ML models were found to be more suitable compared to single ML models.

5.5 Research Hypotheses

$H0_1$: There are no statistically significant performance differences between the hybrid ML models and the single ML models in COM.

HA_1: There are statistically significant performance differences between the hybrid ML models and the single ML models in COM.

The hybrid machine-learning models outperformed the single machine learning models. This reveals improved classification accuracy. Moreover, there was a statistically significant difference between the performance of the hybrid machine learning models and the single machine learning models studied in this work. Thus, in applications where classification accuracy is not the primary performance indicator, one could use any of the machine-learning models for carrying out comparative opinion mining. However, tt would be best to use a hybrid model where feature extraction is performed by a deep learning model while actual classification is performed by a machine-learning

model. For example, the MLP+RF hybrid model would fit this use case. However, for relatively small datasets, a hybrid of machine learning models such as SGD-RF would yield satisfactory results at no extra computational costs that deep learning models would demand. These findings are similar to those of study by Al Amrani et al. (2018), the statistical difference between their hybrid models and the single models was not significant even though accuracy improved in the case of their hybrid models.

In concluding this chapter, the results show that the hybrid machine-learning model developed in this study are generalizable to applications in comparative opinion mining. This is because, on average, across multiple datasets, the hybrid ML models showed significant performance gains over the respective single models. The researcher noted lack of advancement in exploring the application of hybrid ML or hybrid DL models in COM. To the best of the researcher's findings, there was no hybrid ML model for COM. This makes this study the first one in this area. Thus, future researchers should focus more on the implementation of hybrid ML models for comparative opinion mining to tap into the performance gains demonstrated in this study. As maintained in a related study, feature extraction techniques affect model performance because quality features in model training result in quality ML models (Trupthi et al., 2016).

$H0_2$: There are no statistically significant performance differences in COM between the different hybrid ML model.

HA_2: There are statistically significant performance differences in COM between the different hybrid ML models in COM.

There findings of this study showed that there are no statistically significant differences in the performance of hybrid ML models in COM. This is because hybrid ML models developed in this study leveraged the strengths of the two single models while limiting their weaknesses thus producing much more consistent predictions in COM. A study by Wankhade et al. (2022) showed an improvement in the predictive accuracy of hybrid models. Another study by Sagi and Rokach (2018) showed that hybrid models boosted the performance of single models in classification tasks. However, there was limited information on the comparative analysis of different hybrid ML models with regards to their performance for benchmarking.

$H0_3$: There are no statistically significant performance differences in COM between the Hybrid ML models based on the choice of dataset choice.

HA_3: There are statistically significant performance differences in COM between the Hybrid ML models based on the choice of dataset.

The study found that there is a statistically significant difference in the performance of machine learning models based on the choice of datasets. The performance of any particular algorithm varied across the three datasets used, confirming that the size and nature of the dataset impacted the performance of algorithms. Similar findings have been reported in studies by Khanvilkar and Vora (2019) where a huge dataset negatively affected the classification of the Random Forest algorithm due to increase feature dimensionality. Huge datasets have high dimensionality compared to small datasets hence can influence classification accuracies.

$H0_4$: There are no statistically significant performance differences in COM between the Hybrid ML models based on the choice of feature extraction technique.

HA_4: There are statistically significant performance differences in COM between the Hybrid ML models based on the choice of feature extraction technique.

This study found that there is a statistically significant difference in the performance of machine learning models based on the choice of feature extraction techniques. From testing different feature extraction models such as Count Vectorizer, TFIDF, CBOW and Skip-Gram, it became clear that ML models achieve varying accuracies when different feature extraction models are adopted. This is because these models handle word contexts differently yet the context of words is important in identifying the relationship between entities to establish the correct opinions attributed to each entity. This finding agrees with those by Devlin et al. (2019) and Mikolov et al. (2013) who also noted that one needs to select specific feature extraction models based on needs such as contextualization of words because this has an effect on how the model performs.

CHAPTER SIX

CONCLUSIONS

This chapter presents a final summary of this research, conclusions, recommendations, research contributions, and limitations and future works emanating from this study. Section 6.1 gives the summary of this study. Section 6.2 presents the conclusions in a manner that corresponds to the research objectives in this work. Section 6.3 gives the recommendations of this work followed by section 6.4 lists down the contributions of this research. Finally, section 6.5 presents the limitations and future work associated with this study.

6.1 Summary

The primary purpose of this study was to develop a hybrid machine-learning model for performing comparative opinion mining. The developed hybrid machine learning had a higher accuracy than the independent machine learning or deep learning models used to create the hybrid model. The hybrid model was then implemented as a software prototype as a proof of concept in brand reputation monitoring. The prototype, through its web interface (dashboard), provided useful information on the performance of a specific target brand based on comparative opinion data that it collected from X and YouTube. The three human experts involved in the ground truth experiment reported that the prototype was a very important tool in helping them automatically gauge how their brands performed based on the brand's online mentions on specific data sources like X. This confirmed the reliability and usefulness of the prototype and therefore the hybrid model itself when applied to brand reputation monitoring through comparative opinion mining. Through hypothesis testing, it was evident that the hybrid model outperformed the single models in performing comparative opinion mining. The

difference was statistically significant, confirming that hybrid models were superior to single models in performing comparative opinion mining and hence, for application in brand reputation monitoring, which leverages the opinion mining outputs (opinion classes).

6.2 Conclusions

The following are the conclusions from this study, organized by the research objectives.

This work concludes that for comparative opinion mining, the accurate detection of the following elements that characterize comparative opinions and hence are critical in the development of reliable and accurate comparative opinion-mining models: comparative sentence detection, entity detection, relation detection, and feature detection. This quintuple, in the researcher's opinion, is very essential in comparative opinion mining regardless of the domain where the resulting model would be applied. The study identified various techniques that have been used in previous studies to detect each of the elements. However, the researcher conclude that the machine learning approach is the more efficient and effective way to detect each of these elements because with the increasing volume of data, other approaches would be difficult to apply. At the very basic level, entity detection is the more critical element because without a good technique for detecting entities correctly, the detection of the other three elements would not be meaningful. Therefore, more efforts should be aimed at achieving high levels of effectiveness and efficiency in entity detection.

This work recommends the creation of hybrid machine-learning models for performing comparative opinion mining from a selection of the following models: Multilayer Perceptron, Stochastic Gradient Descent, and Random Forest. This is because these

three models outperformed the other seven models that were independently experimented with on the same datasets and features. The fact that the hybrid models created from these three single models outperformed the single models and, indeed, every other single model confirms that these three are the most optimal ones for use in comparative opinion mining.

While there are different techniques for creating hybrid machine-learning models, a popular and effective method is through ensemble learning. This technique has been used in numerous studies and demonstrated its power because of the manner in which the combined models leverage the strengths of each other, resulting in much more accurate models for use in different tasks including classification tasks like in opinion mining.

The researcher conclude that hybrid machine-learning models should be validated using common methods that are used in testing the single models from which the hybrid models are created. This is to make comparative model performance analysis easier. For this reason, the researcher conclude that accuracy metric should be used especially where the datasets are well balanced. However, in the case of imbalanced datasets, the f-measure would suffice.

Hybrid machine-learning models are more accurate in classifying comparative opinions as compared to single ML models. As evidenced in the statistical hypothesis test results; there is s statistically significant difference in the accuracy between the hybrid models and the single models. Thus, where accuracy is of significant value, brands should use the hybrid models. However, if the difference in accuracy does not matter to a brand, one can use the single models. Results from testing the models using the ground truth

after model training gave more confidence in the reliability of the hybrid model since the results of the human classifiers versus those of the hybrid models were relatively similar, improving confidence in the hybrid models.

Feature extraction techniques have a significant effect on how machine-learning models perform comparative opinion mining. This was evidenced in the statistically significant difference in the accuracy of the developed models when using different feature extraction techniques like count Vectorizer and TFIDF Vectorizer.

The choice of datasets does not have a statistically significant effect on the performance of machine learning models in comparative opinion mining. There was consistent performance

Web-based dashboards are an effective way of monitoring different products and/or services online. With the visualizations such as histograms and line graphs, coupled with summary statistics, a brand manager would instantly be able to understand the current position of a target brand regarding its online reputation. This has the effect of making strategic-decision making easier and faster in a bid to enhance competitive advantage by making use of the competitive intelligence the dashboard avails.

6.3 Recommendations

i. The researcher recommends improvements in named entity identification through machine learning approaches especially because of language nuances that make the use of natural language processing and rule mining approaches unfeasible for large datasets.

ii. This study recommends the development of automatic tools for testing single machine learning or deep learning models on the same dataset to help automatically determine the most optimal models that can be combined into a hybrid model for opinion mining.

iii. To help in improving model performance evaluation, this study recommends the creation of comprehensive, balanced, and huge comparative opinion datasets. This will help in building a strong ground truth for model training and testing.

iv. The researcher recommends the development of library (e.g. for Python) for automatically mixing of one machine learning model with another machine learning or deep learning model in creating a hybrid model with automatic substitution of base model with a final estimator model and vice versa. In this case, the base model extracts features while the final model carries out the classification task.

v. This study recommends further exploration on performing comparative opinion mining involving at least three entities. The current study faced the challenge of complexities associated with multiple classes arising from the multiple entities

present in each text under analysis hence resolving to use two entities to reduce the complexities involved.

vi. Machine learning and deep learning algorithms learn better from huge amounts of training data; it would be interesting for future research to show how hybrid machine-learning models would perform on huge datasets with tens of thousands of reviews.

vii. Dense vectors underperformed in the experiments in this study, probably because of the relatively small datasets the researcher used. It would be interesting to find out how dense vectors would perform in comparative opinion mining if huge datasets were used.

6.4 Research Contributions

i. Developed a Hybrid ML Model and the associated algorithms for use in comparative opinion mining.

ii. Developed and applied hybrid ML model on diverse datasets and using different feature extraction techniques for performing COM, which will help future researchers in opinion mining to advance the exploration on the application of complex hybrid ML models to perform comparative opinion mining to achieve, possibly, more improved opinion classification results. To the best of the researcher's knowledge, this is the first study to develop a hybrid machine learning based approach to performing COM and subsequently applying the model to performing brand reputation monitoring.

iii. Brand managers in charge of monitoring the reputation of their brands will benefit from this study by leveraging the power of hybrid machine-learning models developed and tested in this study, showing improved classification accuracy. The higher accuracy of hybrid models translates to higher accuracy in measuring a brand's reputation.

iv. By leveraging the performance evaluation results of both single ML algorithms and the developed hybrid ML algorithms tested in this study, developers of opinion mining tools can benefit from choosing the best performing single or hybrid classification algorithms for use in COM applications like brand reputation monitoring.

v. A systematic literature review on the state-state-of-the-art in ML algorithms / techniques, datasets, features, and algorithm evaluation metrics suitable in opinion mining was published during this research. This contributes to the advancement of knowledge in NLP (a field within computer science) as applied to COM.

vi. The study led to the development of a hybrid opinion-mining model that can perform direct opinion mining, comparative opinion mining, or both. This was important because, currently, direct opinion mining is the predominant approach to brand reputation management. The inclusion of comparative opinion mining will help take care of comparative opinions that make up 10% of user-generated content, producing more precise opinion results for brands to improve their competitive intelligence from monitoring their reputation with respect to their competitor brands.

6.5 Limitations and Future Work

i. Limited datasets containing comparative text: there was a challenge in accessing datasets containing comparative opinions. Comparative opinion datasets are not readily available. Some comparative datasets have imbalanced classes hence prone to model overfitting. Further works in creating and using bigger comparative opinion datasets for comparative opinion mining may help address this limitation.

ii. Computational resource complexities: the use of deep learning models in some of the hybrid machine learning models, while powerful, faced the challenge of complex computational resource requirements. Training neural network-based models like the multilayer perceptron takes a lot of time and computing resources including high performance CPUs and GPUs. These resources are costly. The duration taken to train the models is longer. Future studies in this area may help reveal optimal computational resource requirements for deep learning models in comparative opinion mining.

iii. Multi-class complexities: handling more than two entities in comparative opinion mining introduces complexities in developing, training, and testing a hybrid machine-learning model. For this reason, this study worked with two entities only. Future research may focus on handling more than three comparable entities in opinion reviews.

REFERENCES

Ahmad, M., Aftab, S., Ali, I., & Hameed, N. (2017). Hybrid Tools and Techniques for Sentiment Analysis: A Review. *International Journal of Multidisciplinary Sciences and Engineering*, *8*(4), 28–33. https://www.researchgate.net/publication/318351105_Hybrid_Tools_and_Techniques_for_Sentiment_Analysis_A_Review

Ahmad, M., Aftab, S., Bashir, M. S., & Hameed, N. (2018). Sentiment analysis using SVM: A systematic literature review. *International Journal of Advanced Computer Science and Applications*, *9*(2), 182–188. https://doi.org/10.14569/IJACSA.2018.090226

Ahmad, R., Okechukwu, W., & Shaari, H. (2022). The Effect of Rebranding on Brand Loyalty: Brand Reputation As Mediator. *International Journal of Academic Research in Business and Social Sciences*, *12*(11). https://doi.org/10.6007/IJARBSS/v12-i11/14581

Al-Dmour, R., Alkhatib, O. H., Al-Dmour, H., & Basheer Amin, E. (2023). The Influence of Social Marketing Drives on Brand Loyalty via the Customer Satisfaction as a Mediating Factor in Travel and Tourism Offices. *SAGE Open*, *13*(2). https://doi.org/10.1177/21582440231181433

Al Amrani, Y., Lazaar, M., & El Kadirp, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*, *127*, 511–520. https://doi.org/10.1016/j.procs.2018.01.150

Alkharabsheh, K., Alawadi, S., Kebande, V. R., Crespo, Y., Fernández-Delgado, M., & Taboada, J. A. (2022). A comparison of machine learning algorithms on design smell detection using balanced and imbalanced dataset: A study of God class. *Information and Software Technology*, *143*, 106736.

https://doi.org/10.1016/j.infsof.2021.106736

Almatarneh, S., & Gamallo, P. (2018). Linguistic Features to Identify Extreme Opinions: An Empirical Study. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11314 LNCS*, 215–223. https://doi.org/10.1007/978-3-030-03493-1_23

Alsaeedi, A., & Khan, M. Z. (2019). A study on sentiment analysis techniques of Twitter data. *International Journal of Advanced Computer Science and Applications*, *10*(2), 361–374.

Anjaria, M., & Guddeti, R. M. R. (2014). Influence factor based opinion mining of Twitter data using supervised learning. *2014 6th International Conference on Communication Systems and Networks, COMSNETS 2014*. https://doi.org/10.1109/COMSNETS.2014.6734907

Ankit, & Saleena, N. (2018). An Ensemble Classification System for Twitter Sentiment Analysis. *Procedia Computer Science*, *132*(Iccids), 937–946. https://doi.org/10.1016/j.procs.2018.05.109

Anuradha, K., Krishna, M. V., & Mallik, B. (2023). Opinion Mining Using Normal Discriminant Piecewise Regressive (NDPR) Sentiment Classification Technique. *Journal of Uncertain Systems*, *16*(03). https://doi.org/10.1142/S1752890922500131

Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, *77*, 236–246. https://doi.org/10.1016/j.eswa.2017.02.002

Arboleda, F. J. M., Velásquez, G. A. A., & Jiménez, G. L. P. (2017). A Proposal for

Brand Analysis with Opinion Mining. *E-Business - State of the Art of ICT Based Challenges and Solutions*. https://doi.org/10.5772/66567

Arora, H., & Bansal, M. (2020). Developing a Model for Sentiment Analysis Technique in the field of Tourism using Deep Learning. *International Journal of Recent Technology and Engineering*, *8*(6), 456–462. https://doi.org/10.35940/ijrte.f7390.038620

Arora, P., Bakliwal, A., & Varma, V. (2012). Hindi Subjective Lexicon Generation using WordNet Graph Traversal. *International Journal of Computational Linguistics and Applications*, *3*(Jan-Jun 2012), 25–29.

Asghar, M. Z., Khan, A., Ahmad, S., & Kundi, F. M. (2014). A Asghar, M. Z., Khan, A., Ahmad, S., & Kundi, F. M. (2014). A Review of Feature Extraction in Sentiment Analysis. J. Basic. Appl. Sci. Res, 4(3), 181–186.Review of Feature Extraction in Sentiment Analysis. *J. Basic. Appl. Sci. Res*, *4*(3), 181–186.

Banihashemi, S., Ding, G., & Wang, J. (2017). Developing a Hybrid Model of Prediction and Classification Algorithms for Building Energy Consumption. *Energy Procedia*, *110*, 371–376. https://doi.org/10.1016/j.egypro.2017.03.155

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*, 281–305.

Bjurstrom, S., & Plachkinova, M. (2015). Sentiment Analysis Methodolog y for Social Web Intelligence. *2015 Americas Conference on Information Systems, AMCIS 2015*, 1–12.

Bland, J. M., & Altman, D. J. (2019). Statistics notes: The use of transformation in

statistical analysis. BMJ, 318(7190), 1482.

Borele, **P.**, & Borikar, D. A. (2016). An Approach to Sentiment Analysis using Artificial Neural Network with Comparative Analysis of Different Techniques. *IOSR Journal of Computer Engineering*, *18*(2), 2278–2661. https://doi.org/10.9790/0661-1802056469

Bottou, **L**., Curtis, F. E., & Nocedal, J. (2018). Optimization Methods for Large-Scale Machine Learning. *SIAM Rev.*, *60*(2), 223–311. https://doi.org/10.1137/16M1080173

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(2), 5–32. https://doi.org/.1023/A:1010950718922

Cambria, E., & Hussain, A. (2012). *Sentic Computing* (Vol. 2). Springer Netherlands. https://doi.org/10.1007/978-94-007-5070-8

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2), 15-21.

Carrillo-de-Albornoz, J., Vidal, J. R., & Plaza, L. (2018). Feature engineering for sentiment analysis in e-health forums. *PLoS ONE*, *13*(11), 1–25. https://doi.org/10.1371/journal.pone.0207996

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. https://doi.org/10.1145/2783258.2788613

Castillo, C., Donato, D., Gionis, A., Murdock, V., & Silvestri, F. (2007). Know your neighbors. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 423–430. https://doi.org/10.1145/1277741.1277814

Chao, **W.-L.** (2011). Machine learning tutorial. *Aec-Apc*. https://sites.google.com/site/jeromelacaille/semiconducteurs/2006-03---arc-apc-aix-en-provence

Chaturvedi, **I.**, Poria, S., & Cambria, E. (2018). Sentiment Analysis, Basic Tasks of. *Encyclopedia of Social Network Analysis and Mining*, 2434–2454. https://doi.org/10.1007/978-1-4939-7131-2_110159

Chauhan, P., & Singh, A. J. (2017). Sentiment Analysis: A Comparative Study of Supervised Machine Learning Algorithms Using Rapid miner. *International Journal for Research in Applied Science and Engineering Technology*, *V*(XI), 80–89. https://doi.org/10.22214/ijraset.2017.11011

Chen, D., & Manning, C. (2014). A fast and accurate dependency parser using neural networks. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, *i*, 740–750.

Chen, T., Xu, R., He, Y., & Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*, *72*, 221–230. https://doi.org/10.1016/j.eswa.2016.10.065

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

Cucerzan, S. (2007). Large-Scale Named Entity Disambiguation Based on Wikipedia Data. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 708–716. https://aclanthology.org/D07-1074

Cumming, J. (2015). *An Investigation into the Use of Reinforcement Learning Techniques within the Algorithmic Trading Domain*. 79.

Das, S., Behera, R. K., Kumar, M., & Rath, S. K. (2018). Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction. *Procedia Computer Science*, *132*(Iccids), 956–964. https://doi.org/10.1016/j.procs.2018.05.111

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot. *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, 273–274. https://doi.org/10.1145/2872518.2889302

Davis, J., Brown, L., & Miller, A. (2020). Feature selection using ANOVA in machine learning: A practical approach. Journal of Machine Learning Research, 21(1), 345-367.

de Albornoz, J. C., Plaza, L., & Gervás, P. (2010). A hybrid approach to emotional sentence polarity and intensity classification. *Conference on Computational Natural Language Learning, Proceedings of the Conference*, *July*, 153–161.

de Vries, J., McGrath, A., & Vaidis, D. (2023). Teaching cognitive dissonance theory: Practical advice for the classroom. *Scholarship of Teaching and Learning in Psychology*. https://doi.org/10.1037/stl0000346

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dey, A. (2016). Machine Learning Algorithms: A Review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174–1179. www.ijcsit.com

Diamantini, C., Mircoli, A., Potena, D., & Storti, E. (2019). Social information discovery enhanced by sentiment analysis techniques. *Future Generation Computer Systems*, *95*(July), 816–828. https://doi.org/10.1016/j.future.2018.01.051

Dixon, B. S. (2020). *Dewey's Pragmatism as a Philosophy for Practice in Design Research Practice* (pp. 175–200). https://doi.org/10.1007/978-3-030-47471-3_7

Doni Abdul Fatah, Eka Mala Sari Rochman, Fajrul Ihsan Kamil, & Ahmad Su'ud. (2023). Sentiment Analysis of Madura Tourism Opinion Using Support Vector Machine (SVM). *Technium: Romanian Journal of Applied Sciences and Technology*, *16*, 243–249. https://doi.org/10.47577/technium.v16i.9988

Ejaz, A., Turabee, Z., Rahim, M., & Khoja, S. (2017). Opinion mining approaches on Amazon product reviews: A comparative study. *2017 International Conference on Information and Communication Technologies (ICICT)*, 173–179. https://doi.org/10.1109/ICICT.2017.8320185

El-Mawass, N., Honeine, P., & Vercouter, L. (2020). SimilCatch: Enhanced social spammers detection on Twitter using Markov Random Fields. *Information Processing & Management*, *57*(6), 102317. https://doi.org/10.1016/j.ipm.2020.102317

El Haddaoui, B., Chiheb, R., Faizi, R., & Afia, A. El. (2018). Toward a sentiment analysis framework for social media. *ACM International Conference Proceeding Series*, *May*. https://doi.org/10.1145/3230905.3230919

Fernandez-Delgado, M., Cernadas, E., Barro, S. (2014). Do we Need Hundreds of Classi ers to Solve Real World Classi cation Problems? *Journal of Machine Learning Research*, *15*, 133–3181.

Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics. Sage Publications.

Fisher, R. A. (1992). *Statistical Methods for Research Workers* (pp. 66–70). https://doi.org/10.1007/978-1-4612-4380-9_6

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Fronzetti Colladon, A. (2019). Brand Intelligence Analytics. *Paper Presented at Text 4 Business. A Workshop on Industrial Applications of Text Mining and NLP Techniques*, 1–14.

Ganapathibhotla, M., Street, S. M., Liu, B., & Street, S. M. (n.d.). *Coling2008-Mining Opinions in Comparative Sentences.pdf*.

Gao, H., Hu, J., Wilson, C., Wang, Z., Zhao, B. Y., & Dai, Y. (2010). Detecting and Characterizing Social Spam Campaigns. *Proceedings of The 10th ACM SIGCOMM Internet Measurement Conference (IMC 2010)*.

Gao, S., Tang, O., Wang, H., & Yin, P. (2018). Identifying competitors through comparative relation mining of online reviews in the restaurant industry. *International Journal of Hospitality Management*, *71*, 19–32. https://doi.org/10.1016/j.ijhm.2017.09.004

Ghag, K. V., & Shah, K. (2018). Conceptual sentiment analysis model. *International Journal of Electrical and Computer Engineering*, *8*(4), 2358–2366. https://doi.org/10.11591/ijece.v8i4.pp2358-2366

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering*, *23*(10), 1498–1512. https://doi.org/10.1109/TKDE.2010.188

González-Gonzalo, C., Thee, E. F., Klaver, C. C. W., Lee, A. Y., Schlingemann, R. O., Tufail, A., Verbraak, F., & Sánchez, C. I. (2022). Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice. *Progress in Retinal and Eye Research*, *90*, 101034. https://doi.org/10.1016/j.preteyeres.2021.101034

Gonzalez, J. C., Martinez, A., & Li, F. (2023). Benchmarking machine learning models

using Kaggle datasets. International Journal of Data Mining, 14(2), 95-112.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning Ian. In *Foreign Affairs* (Vol. 91, Issue 5). MIT Press.

Gu, Y. H., & Yoo, S. J. (2009). Rules for Mining Comparative Online Opinions. *2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*, 1294–1299. https://doi.org/10.1109/ICCIT.2009.16

Guest, G., Bunce, A., & Johnson, L. (2006). How Many Interviews Are Enough? *Field Methods*, *18*(1), 59–82. https://doi.org/10.1177/1525822X05279903

Gulli, A., & Pal, S. (2017). *Deep Learning with Keras: Implement neural networks with Keras on Theano and TensorFlow* (First). Packt Publishing.

Gürsoy, U. T., Bulut, D., Yiğit, C., & Co, S. (2017). Social Media Mining and Sentiment Analysis for Brand Management. *GJETeMCP) An Online International Research Journal*, *3*, 1. www.globalbizresearch.org

Hamdard, A. P. M. S., & Lodin, A. P. H. (2023). Effect of Feature Selection on the Accuracy of Machine Learning Model. *INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND ANALYSIS*, *06*(09). https://doi.org/10.47191/ijmra/v6-i9-66

Harfoushi, O., Hasan, D., & Obiedat, R. (2018). Sentiment Analysis Algorithms through Azure Machine Learning: Analysis and Comparison. *Modern Applied Science*, *12*(7), 49. https://doi.org/10.5539/mas.v12n7p49

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328–339.

https://doi.org/10.18653/v1/P18-1031

Huda, K. (2017). *Classification Technique for Sentiment Analysis of Twitter Data*. *8*(5), 2551–2555.

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM*.

Iqbal, F., Hashmi, J. M., Fung, B. C. M., Batool, R., Khattak, A. M., Aleem, S., & Hung, P. C. K. (2019). A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction. *IEEE Access*, *7*, 14637–14652. https://doi.org/10.1109/ACCESS.2019.2892852

Isah, H., Trundle, P., & Neagu, D. (2014). Social media analysis for product safety using text mining and sentiment analysis. *2014 14th UK Workshop on Computational Intelligence, UKCI 2014 - Proceedings*. https://doi.org/10.1109/UKCI.2014.6930158

Istrati, L., & Ciobotaru, A. (2022). *Automatic Monitoring and Analysis of Brands Using Data Extracted from Twitter in Romanian* (pp. 55–75). https://doi.org/10.1007/978-3-030-82199-9_5

Jahanbin, K., Rahmanian, F., Rahmanian, V., Shakeri, M., Shakeri, H., Rahmanian, Z., & Jahromi, A. S. (2019). A Perspective on Text Classification, Clustering, and Named-entity Recognition in Social Media. *Ambient Science*, *06* & *06h*(2), 2016–2017. https://doi.org/10.21276/ambi.2019.06.1.ga01

Jain, P. K., Saravanan, V., & Pamula, R. (2021). A Hybrid CNN-LSTM: A Deep Learning Approach for Consumer Sentiment Analysis Using Qualitative User-Generated Contents. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *20*(5), 1–15. https://doi.org/10.1145/3457206

Jindal, N., Liu, B., & Street, S. M. (2006). *SIGIR06-Identifying Comparative Sentences in Text Documents-sigir06.pdf*.

Joachims, T. (1998). *Text categorization with Support Vector Machines: Learning with many relevant features* (pp. 137–142). https://doi.org/10.1007/BFb0026683

Kaggle. (2021). Datasets and competitions on Kaggle. Retrieved from Kaggle.

Kalaivani, A., & Thenmozhi, D. (2019). Sentimental analysis using deep learning techniques. *International Journal of Recent Technology and Engineering*, *7*(6), 600–606.

Kane, S. N., Mishra, A., & Dutta, A. K. (2016). A Framework for Sentiment Analysis Implementation of Indonesian Language Tweet on Twitter. *Journal of Physics: Conference Series*, *755*(1), 1–6. https://doi.org/10.1088/1742-6596/755/1/011001

Karijadi, I., & Chou, S.-Y. (2022). A hybrid RF-LSTM based on CEEMDAN for improving the accuracy of building energy consumption prediction. *Energy and Buildings*, *259*, 111908. https://doi.org/10.1016/j.enbuild.2022.111908

Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R., & Mora, H. (2019). Managing marketing decision-making with sentiment analysis: An evaluation of the main product features using text data mining. *Sustainability (Switzerland)*, *11*(15), 1–19. https://doi.org/10.3390/su11154235

Kaur, B., & Kumari, N. (2016). *A Hybrid Approach to Sentiment Analysis of Technical Article Reviews*. *November*, 1–11. https://doi.org/10.5815/ijeme.2016.06.01

Kaur, J. (2016). A Review Paper on Twitter Sentiment Analysis Techniques. *International Journal for Research in Applied Science & Engineering Technology*, *4*(X), 61–70.

Kessler, W., & Kuhn, J. (2013). Detection of product comparisons - How far does an out-of-the-box semantic role labeling system take you? *EMNLP 2013 - 2013*

Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, *July*, 1892–1897. https://aclanthology.org/D13-1194

Khairi, N. I., Mohamed, A., & Yusof, N. N. (2020). Feature Selection Methods in Sentiment Analysis: A Review. *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, 1–7. https://doi.org/10.1145/3386723.3387840

Khairnar, J., & Kinikar, M. (2013). Machine Learning Algorithms for Opinion Mining and Sentiment Classification. *International Journal of Scientific and Research Publications*, *3*(6), 1–6. www.ijsrp.org

Khanvilkar, G., & Vora, D. (2019). Product recommendation using sentiment analysis of reviews: A random forest approach. *International Journal of Engineering and Advanced Technology*, *8*(2), 146–152.

Kharde, V., & Sonawane, S. S. (2016). Sentiment analysis of Twitter data: A hybrid approach using machine learning. International Journal of Computer Applications, 139(11), 45-48.

Khomsah, S., Cahyana, N. H., & Aribowo, A. S. (2023). Hyperparameter Tuning of Semi-Supervised Learning for Indonesian Text Annotation. *International Journal of Advanced Computer Science and Applications*, *14*(9). https://doi.org/10.14569/IJACSA.2023.0140927

Khuri, A. I. (2013). Introduction to Linear Regression Analysis, Fifth Edition by Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining. *International Statistical Review*, *81*(2), 318–319. https://doi.org/10.1111/insr.12020_10

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. https://doi.org/10.3115/v1/D14-1181

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment Analysis of Short Informal Texts. *Journal of Artificial Intelligence Research*, *50*, 723–762. https://doi.org/10.1613/jair.4272

Kothari, C. R. (2015). Research methodology: Methods and Techniques. In *Syria Studies* (2nd ed., Vol. 7, Issue 1). New Age International (P) Limited, Publishers. https://www.researchgate.net/publication/269107473_What_is_governance/link/548173090cf22525dcb61443/download%0Ahttp://www.econ.upf.edu/~reynal/Civil wars_12December2010.pdf%0Ahttps://think-asia.org/handle/11540/8282%0Ahttps://www.jstor.org/stable/41857625

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, *31*, 249–268. https://doi.org/10.1007/s10751-016-1232-6

Kourentzes, N., et al. (2020). Time series forecasting with machine learning: A review. Journal of Business Research, 107, 1-16.

Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of Emojis. *PLOS ONE*, *10*(12), e0144296. https://doi.org/10.1371/journal.pone.0144296

Kumar, A., & Sharma, A. (2017). Systematic literature review on opinion mining of big data for government intelligence. *Webology*, *14*(2), 6–47.

Kumar, V., Aksoy, L., Donkers, B., Venkatesan, R., Wiesel, T., & Tillmanns, S. (2010). Undervalued or Overvalued Customers: Capturing Total Customer Engagement Value. *Journal of Service Research*, *13*(3), 297–310. https://doi.org/10.1177/1094670510375602

Kurashima, T., Bessho, K., Toda, H., Uchiyama, T., & Kataoka, R. (2008). Ranking Entities Using Comparative Relations. In S. S. Bhowmick, J. Küng, & R. Wagner (Eds.), *Database and Expert Systems Applications* (Vol. 5181, pp. 124–133).

Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-85654-2_15

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, S.-H., & Jeong, G.-Y. (2022). The Effect of Corporate Social Responsibility Compatibility and Authenticity on Brand Trust and Corporate Sustainability Management: For Korean Cosmetics Companies. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.895823

Levy, O., & Goldberg, Y. (2014). Neural Word Embedding as a Substitution Language Model. Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), 2148-2156.

Li, S., Zha, Z.-J., Ming, Z., Wang, M., Chua, T.-S., Guo, J., & Xu, W. (2011). Product comparison using comparative relations. *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1151–1152. https://doi.org/10.1145/2009916.2010094

Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: a tertiary study. *Artif Intell Rev*, *54*(7), 4997–5053. https://doi.org/10.1007/s10462-021-09973-3

Liu, B. (2012a). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–184. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, B. (2012b). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, B. (2015a). Sentiment analysis: Mining opinions, sentiments, and emotions. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, *May*, 1–367.

https://doi.org/10.1017/CBO9781139084789

Liu, B. (2015b). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. In *Cambridge University Press*. https://doi.org/10.14569/ijacsa.2018.090981

Liu, H., Gao, Y., Lv, P., Li, M. M., Geng, S., Li, M. M., & Wang, H. (2017). Using Argument-based features to predict and analyse review helpfulness. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 1358–1363. https://doi.org/10.18653/v1/d17-1142

Liu, Q., Huang, H., Zhang, C., Chen, Z., & Chen, J. (2013). Chinese Comparative Sentence Identification Based on the Combination of Rules and Statistics. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, H. Motoda, Z. Wu, L. Cao, O. Zaiane, … W. Wang (Eds.), *Advanced Data Mining and Applications* (Vol. 8347, pp. 300–310). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-53917-6_27

Liu, S., & Forss, T. (2014). Combining N-gram based Similarity Analysis with Sentiment Analysis in Web Content Classification: *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, 530–537. https://doi.org/10.5220/0005170305300537

Liu, X., Burns, A. C., & Hou, Y. (2017). An Investigation of Brand-Related User-Generated Content on Twitter. *Journal of Advertising*, *46*(2), 236–247. https://doi.org/10.1080/00913367.2017.1297273

Liu, Z., Xia, R., & Yu, J. (2021). Comparative Opinion Quintuple Extraction from Product Reviews. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 955–3965. https://doi.org/https://doi.org/10.48448/bxgr-xw55

Lu, H., Ehwerhemuepha, L., & Rakovski, C. (2022). A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Med Res Methodol*, *22*(1), 181. https://doi.org/10.1186/s12874-022-01665-y

Lusch, R. F., & Vargo, S. L. (2014). *The Service-Dominant Logic of Marketing*. Routledge. https://doi.org/10.4324/9781315699035

Maas, A. L., Daly, R. E., Pham, T. P, Ng, A. Y., Potts, C. (2011). Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. In R. M. Dekang Lin, Yuji Matsumoto (Ed.), *Association for Computational Linguistics* (pp. 142–150). Association for Computational Linguistics. https://aclanthology.org/P11-1015

Madhoushi, Z., Hamdan, A. R., & Zainudin, S. (2015). Sentiment analysis techniques in recent works. *Proceedings of the 2015 Science and Information Conference, SAI 2015*, 288–291. https://doi.org/10.1109/SAI.2015.7237157

Malik, V., & Kumar, A. (2018). Analysis of Twitter Data Using Deep Learning Approach: LSTM. *International Journal on Recent and Innovation Trends in Computing and Communication*, *6*(4), 144–149.

Malmqvist, J., Hellberg, K., Möllås, G., Rose, R., & Shevlin, M. (2019). Conducting the Pilot Study: A Neglected Part of the Research Process? Methodological Findings Supporting the Importance of Piloting in Qualitative Research Studies. *International Journal of Qualitative Methods*, *18*, 160940691987834. https://doi.org/10.1177/1609406919878341

Manning, C. D., & Schütze, H. (2000). *Foundations of statistical natural language processing.* MIT Press.

Manouselis, N., Drachsler, H., Verbert, K., Santos, O. C., & Konstan, J. A. (2014).

Recommender systems for technology enhanced learning: Research trends and applications. *Recommender Systems for Technology Enhanced Learning: Research Trends and Applications*, *July*, 1–306. https://doi.org/10.1007/978-1-4939-0530-0

Marneffe, M. De, & Manning, C. D. (2012). *Stanford typed dependencies manual*. *November*, 1–26.

Martins, F. S., Cunha, J. A. C. da, & Serra, F. A. R. (2018). Secondary Data in Research – Uses and Opportunities. *Revista Ibero-Americana de Estratégia*, *17*(04), 01–04. https://doi.org/10.5585/ijsm.v17i4.2723

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem Med*, 276–282. https://doi.org/10.11613/BM.2012.031

McAuley Lab (2023) - URL: https://amazon-reviews-2023.github.io/

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Meusel, R., Mika, P., & Blanco, R. (2014). Focused Crawling for Structured Data. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 1039–1048. https://doi.org/10.1145/2661829.2661902

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Efficient Estimation of Word Representations in Vector Space*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations ofwords and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 1–9.

Money, V. (2023). Demonstrating Anticipatory Deflection and a Preemptive Measure to Manage It: An Extension of Affect Control Theory. *Social Psychology Quarterly*, *86*(2), 151–169. https://doi.org/10.1177/01902725221132508

Montgomery, D. C. (2017). Design and Analysis of Experiments. Wiley.

Morse, J. M. (2015). Critical Analysis of Strategies for Determining Rigor in Qualitative Inquiry. *Qualitative Health Research*, *25*(9), 1212–1222. https://doi.org/10.1177/1049732315588501

Muhaise, H., Ejiri, A. H., Muwanga-zake, J. W. F., & Kareyo, M. (2020). The Research Philosophy Dilemma for Postgraduate Student Researchers. *International Journal of Research and Scientific Innovation (IJRSI)*, *VII*(Iv), 201–204. https://doi.org/2321–2705

Mukwazvure, A., & Supreethi, K. P. (2015). A hybrid approach to sentiment analysis of news comments. *2015 4th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO 2015*, 1–6. https://doi.org/10.1109/ICRITO.2015.7359282

Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., & By, T. (2012). Sentiment analysis on social media. *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012, August*, 919–926. https://doi.org/10.1109/ASONAM.2012.164

Nimbhore, A. P., & Siledar, S. B. (2014). Identifying Opinion Features Using Intrinsic and Extrinsic Domain Relevance. Int. J. Tech. Res. Appl, 4(1), 2320-8163.

Novalita, N., Herdiani, A., Lukmana, I., & Puspandari, D. (2019). Cyberbullying identification on twitter using random forest classifier. *Journal of Physics: Conference Series*, *1192*(1). https://doi.org/10.1088/1742-6596/1192/1/012029

O'Dea, R. E., Lagisz, M., Jennions, M. D., Koricheva, J., Noble, D. W. A., Parker, T.

H., Gurevitch, J., Page, M. J., Stewart, G., Moher, D., & Nakagawa, S. (2021). Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology: a <scp>PRISMA</scp> extension. *Biological Reviews*, *96*(5), 1695–1722. https://doi.org/10.1111/brv.12721

Ojo, O. E., Gelbukh, A., Calvo, H., & Adebanji, O. O. (2021). Performance Study of N-grams in the Analysis of Sentiments. *J. Nig. Soc. Phys. Sci.*, 477–483. https://doi.org/10.46481/jnsps.2021.201

Omar, N., Albared, M., Al-Moslmi, T., & Al-Shabi, A. (2014). *A Comparative Study of Feature Selection and Machine Learning Algorithms for Arabic Sentiment Classification* (pp. 429–443). https://doi.org/10.1007/978-3-319-12844-3_37

Ondara, B., Waithaka, S., Kandiri, J., & Muchemi, L. (2022). Machine Learning Techniques, Features, Datasets, and Algorithm Performance Parameters for Sentiment Analysis: A Systematic Review. *Open Journal for Information Technology*, *5*(1), 1–16. https://doi.org/10.32591/coas.ojit.0501.01001o

Ondara, B., Waithaka, S., Kandiri, J., & Muchemi, L. (2023). Hybrid Machine Learning Techniques for Comparative Opinion Mining. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, *6*(2), 131 – 143.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *FNT in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Pang, B., Lee, L., & Vaithyanathan, S. (2012). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, *57*(July), 79–86. https://doi.org/10.1515/9783110239171.151

Pantano, E., Giglio, S., & Dennis, C. (2019). Making sense of consumers' tweets: Sentiment outcomes for fast fashion retailers through Big Data analytics.

International Journal of Retail and Distribution Management, *47*(9), 915–927. https://doi.org/10.1108/IJRDM-07-2018-0127

Passon, M., Lippi, M., Serra, G., & Tasso, C. (2019). *Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining*. 35–39. https://doi.org/10.18653/v1/w18-5205

Perakakis, E., Mastorakis, G., & Kopanakis, I. (2019). Social Media Monitoring: An Innovative Intelligent Approach. *Designs*, *3*(2), 24. https://doi.org/10.3390/designs3020024

Peters, J., Duma, C., & Scherer, R. (2020). Data quality assessment for machine learning applications. Data Science & Engineering, 5(1), 12-25.

Ploder, A., & Eder, A. (2015). Semantic Differential. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* (Second Edi, Vol. 21). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.03231-1

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A Stylometric Inquiry into Hyperpartisan and Fake News. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 231–240. https://doi.org/10.18653/v1/P18-1022

Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. https://doi.org/10.9735/2229-3981

Raschka, S., & Mirjalili, V. (2019). Python Machine Learning. Packt Publishing.

Rathor, A. S., Agarwal, A., & Dimri, P. (2018). Comparative Study of Machine Learning Approaches for Amazon Reviews. *Procedia Computer Science*, *132*, 1552–1561. https://doi.org/10.1016/j.procs.2018.05.119

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis:

Tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14–46. https://doi.org/10.1016/j.knosys.2015.06.015

Reel, S., Wong, P., Wu, B., Mostefaoui, S. K., & Liu, H. (2019). *Identifying tweets from Syria refugees using a Random Forest classifier*. *January*.

Reynolds, K. J., Turner, J. C., Haslam, S. A., Reynolds, K. J., Turner, J. C., & Haslam, S. A. (2015). Identity Issues in Groups SELF-CATEGORIZATION THEORIES ' CONTRIBUTION TO UNDERSTANDING IDENTIFICATION , SALIENCE AND DIVERSITY IN TEAMS AND ORGANIZATIONS. *Identity Issues in Groups*.

Ruder, S. (2016). *An overview of gradient descent optimization algorithms*. http://arxiv.org/abs/1609.04747

Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer Learning in Natural Language Processing. *Proceedings of the 2019 Conference of the North*, 15–18. https://doi.org/10.18653/v1/N19-5004

Rushing, B. (2022). No free theory choice from machine learning. *Synthese*, *200*(5), 414. https://doi.org/10.1007/s11229-022-03901-w

Saberi, B., & Saad, S. (2017). Sentiment analysis or opinion mining: A review. *International Journal on Advanced Science, Engineering and Information Technology*, *7*(5), 1660–1666. https://doi.org/10.18517/ijaseit.7.5.2137

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4). https://doi.org/10.1002/widm.1249

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Saranya, N., Phil, M., & Gunavathi, R. (2016). A Study on Various Classification

Techniques for Sentiment Analysis on Social Networks. *International Research Journal of Engineering and Technology*, *3*(8), 1332–1337.

Schrauwen, S. (2010). Ctrs-001. *Computational Linguistics and Psycholinguistics Technical Report Series*.

Sebastiani, F., Day, M., Extremes, M., Liu, B., Pang, B., Lee, L., Wiegand, M., Remus, R., Gindl, S., Liu, B., Khan, K., Khan, W., Ur Rahman, A., Khan, A. U. A. A., Khan, A. U. A. A., Khan, A. U. A. A., Saqia, B., Jindal, N., Liu, B., … Song, Y. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *1*(1), 1–184. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Setiawan, A., & Sutarso, Y. (2023). THE INFLUENCE OF BRAND SIGNATURE, AWARENESS, ATTITUDE, AND REPUTATION ON PRIMEBIZ HOTEL SURABAYA'S BRAND PERFORMANCE. *Business and Finance Journal*, *8*(2), 126–140. https://doi.org/10.33086/bfj.v8i2.5222

Shalev-Shwartz, S., & Ben-David, S. (2013). Understanding machine learning: From theory to algorithms. In *Understanding Machine Learning: From Theory to Algorithms* (Vol. 9781107057). https://doi.org/10.1017/CBO9781107298019

Shang, E., Vignali, G., & Henninger, C. (2023). *The Influence of Sensory Marketing on Consumers with Different Characteristics Regarding Physical Store Shopping* (pp. 209–240). https://doi.org/10.1007/978-3-031-33302-6_12

Sharma, A., & Dey, S. (2012). A hybrid approach to sentiment analysis of Twitter data. Proceedings of the 2012 International Conference on Cloud Computing and Social Networking, 1-7.

Sharma, R., Alavi, S., & Ahuja, V. (2017). Generating trust using Facebook-A study of 5 online apparel brands. *Procedia Computer Science*, *122*, 42–49.

https://doi.org/10.1016/j.procs.2017.11.339

Smith, A. N., Fischer, E., & Yongjian, C. (2012). How Does Brand-related User-generated Content Differ across YouTube, Facebook, and Twitter? *Journal of Interactive Marketing*, *26*(2), 102–113. https://doi.org/10.1016/j.intmar.2012.01.002

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Soscia, I., Girolamo, S., Busacca, B., Journal, S., March, N., Soscia, I., & Girolamo, S. (2017). *The Effect of Comparative Advertising on Consumer Perceptions : Similarity or Differentiation ? Stable URL : http://www.jstor.org/stable/40605749 REFERENCES Linked references are available on JSTOR for this article : The Effect of Comparative Advertising* . *25*(1), 109–118.

Sun, J., Long, C., Zhu, X., & Huang, M. (2009). Mining Reviews for Product Comparison and Recommendation. *Polibits*, *39*, 33–40. https://doi.org/10.17562/PB-39-5

Sundaram, J., Gowri, K., Devaraju, S., Gokuldev, S., Jayaprakash, S., Anandaram, H., Manivasagan, C., & Thenmozhi, M. (2023). *An Exploration of Python Libraries in Machine Learning Models for Data Science* (pp. 1–31). https://doi.org/10.4018/978-1-6684-8696-2.ch001

Sureka, A., Mirajkar, P. P., & Varma, K. I. (2009). A rapid application development framework for rule-based named-entity extraction. *Proceedings of the 2nd Bangalore Annual Compute Conference*, 1–4. https://doi.org/10.1145/1517303.1517330

Suresh, A., & Bharathi, C. R. (2016). Sentiment Classifi cation using Decision Tree

Based Feature Selection Sentiment Classi fi cation using Decision Tree Based Feature Selection. *International Journal of Control Theory and Applications*, *9*(36), 419–425.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

Taboada, M., Brooke, J., & Voll, K. (2011). *Lexicon-Based Methods for Sentiment Analysis*. *September 2010*.

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis. *Applied Sciences*, *13*(6), 3915. https://doi.org/10.3390/app13063915

Thomas, K., Grier, C., Ma, J., Paxson, V., & Song, D. (2011). Design and Evaluation of a Real-Time URL Spam Filtering Service. *2011 IEEE Symposium on Security and Privacy*, 447–462. https://doi.org/10.1109/SP.2011.25

Tkachenko, M., & Lauw, H. W. (2014). Generative Modeling of Entity Comparisons in Text. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 859–868. https://doi.org/10.1145/2661829.2662016

Trepte, S., & Loy, L. S. (2017). Social Identity Theory and Self-Categorization Theory. *The International Encyclopedia of Media Effects*, 1–13. https://doi.org/10.1002/9781118783764.wbieme0088

Tripathy, A., Anand, A., & Rath, S. K. (2017). Document-level sentiment classification using hybrid machine learning approach. *Knowledge and Information Systems*, *53*(3), 805–831. https://doi.org/10.1007/s10115-017-1055-z

Tripopsakul, S., & Puriwat, W. (2023). Exploring the relationship between ESG, trust, brand reputation, and brand equity. *International Journal of ADVANCED AND*

APPLIED SCIENCES, *10*(10), 71–77. https://doi.org/10.21833/ijaas.2023.10.008

Trupthi, M., Pabboju, S., & Narasimha, G. (2016). Improved feature extraction and classification - Sentiment analysis. *2016 International Conference on Advances in Human Machine Interaction, HMI 2016*, 117–122. https://doi.org/10.1109/HMI.2016.7449189

Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 417–424. https://doi.org/10.3115/1073083.1073153

van Doorn, J., Lemon, K. N., Mittal, V., Nass, S., Pick, D., Pirner, P., & Verhoef, P. C. (2010). Customer Engagement Behavior: Theoretical Foundations and Research Directions. *Journal of Service Research*, *13*(3), 253–266. https://doi.org/10.1177/1094670510375599

Varathan, K. D., Giachanou, A., & Crestani, F. (2017). Comparative opinion mining: A review. *Journal of the Association for Information Science and Technology*, *68*(4), 811–829. https://doi.org/10.1002/asi.23716

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Illia Polosukhin. (2017). Attention Is All You Need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 1–15.

Vidya, N. A., Fanany, M. I., & Budi, I. (2015). Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers. *Procedia Computer Science*, *72*(December), 519–526. https://doi.org/10.1016/j.procs.2015.12.159

Wang, W., Zhao, T. J., Xin, G. D., & Xu, Y. D. (2015). Exploiting Machine Learning for Comparative Sentences Extraction. *IJHIT*, *8*(3), 347–354. https://doi.org/10.14257/ijhit.2015.8.3.31

Wang, Wu, & Dong. (2015). Exploring the impacts of social networking on brand image and purchase intention in cyberspace. *Journal of Universal Computer Science*, *21*(11), 1425–1438.

Wang, Y., & Li, B. (2016). Sentiment Analysis for Social Media Images. *Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015*, *November*, 1584–1591. https://doi.org/10.1109/ICDMW.2015.142

Wang, W., Lu, L., & Zhang, M. (2020). Sentiment analysis with reinforcement learning: A comprehensive survey. IEEE Transactions on Neural Networks and Learning Systems, 31(4), 1410-1425.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artif Intell Rev*, *55*(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Weitzl, W. J. (2019). Webcare's effect on constructive and vindictive complainants. *Journal of Product and Brand Management*, *28*(3), 330–347. https://doi.org/10.1108/JPBM-04-2018-1843

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, *39*(2–3), 165–210. https://doi.org/10.1007/s10579-005-7880-9

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, 347–354. https://doi.org/10.3115/1220575.1220619

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Xu, K., Liao, S. S., Li, J., & Song, Y. (2011). Mining comparative opinions from customer reviews for Competitive Intelligence. *Decision Support Systems*, *50*(4), 743–754. https://doi.org/10.1016/j.dss.2010.08.021

Yan, J. (2022). Text Mining with R: A Tidy Approach, by Julia Silge and David Robinson. Sebastopol, CA: O'Reilly Media, 2017. ISBN 978-1-491-98165-8. XI + 184 pages. *Nat. Lang. Eng.*, *28*(1), 137–139. https://doi.org/10.1017/S1351324920000649

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical Attention Networks for Document Classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489. https://doi.org/10.18653/v1/N16-1174

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, *27*(NIPS '14)), **3320–3328.**

Younis, E. M. G. (2015). Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study. *International Journal of Computer Applications*, *112*(5), 44–48.

Younis, U., Asghar, M. Z., Khan, A., Khan, A., Iqbal, J., & Jillani, N. (2020). Applying Machine Learning Techniques for Performing Comparative Opinion Mining. *Open Computer Science*, *10*(1), 461–477. https://doi.org/10.1515/comp-2020-0148

Yu, S., & Kak, S. (2012). *A Survey of Prediction Using Social Media*. 1–20.

http://arxiv.org/abs/1203.1647

Yueyang, L., & Wang, Y. Z. (2019). Detecting Opinion Polarities Using Ensemble of Classification Algorithms. *Journal of Physics: Conference Series*, *1229*, 012065. https://doi.org/10.1088/1742-6596/1229/1/012065

Zembik, M. (2015). Brand image in social media – An outline of the research related issues. *Polish Journal of Management Studies*, *11*(2), 203–212.

Zhang, H., et al. (2019). A survey of machine learning techniques in wireless sensor networks. Journal of Network and Computer Applications, 127, 94-108.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107–115. https://doi.org/10.1145/3446776

APPENDICES

Appendix I - Work Plan

Duration / ***Tasks***	***May-Dec 2019***	***Jan-Jun 2020***	***Nov 2020***	***Dec 2020***	***Jan 2021***	***Feb-Mar 2021***	***Apr-May 2021***	***May-Jun 2023***	***Feb-Mar 2024***
Research Concept Development									
Proposal Writing									
Proposal Defense									
Data Collection									
Model Design									
Model Development									
Model Validation									
Analysis of Results									
Thesis Writing									
Research Publications									
Thesis Defense									

Appendix II – Budget

S.N	Item Description	Cost Per Unit	Quantity	Total Cost (KES)
1	Research Stationeries	N/A	N/A	15,000
2	Data Collection	25,000	Twice	50,000
3	Experimentation	75,000	Once	75,000
4	Validation	75,000	Once	75,000
5	Conference Fees	25,000	4 times	100,000
6	Journal Registration	10,000	4 times	40,000
7	Workshops / Seminars	20,000	3	60,000
8	Internet Bundles	50,000	-	50,000
9	Tool development & integration costs	250,000	1	150,000
	Total			**615,000**

Appendix III – Primary Data

Nissan Vs Toyota Land Cruiser		
ID	**Text**	**Opinion Label**
1	Both cars are rlly perfect depends on your brand you like	pos_pos
2	Nissan is better than Toyota in general!	pos_neg
3	Toyota is dated and boring, will take the Nissan	pos_neg
4	your wrong about the fuel---my 200 series used 25 to 27 litres per 100 towing a 2700kgs van, my y62 uses 20l to 24 per 100 towing a 3300kg van, i would doubt the v6 is any better	net_neg
5	The Landcruiser was and is a great SUV. But the Patrol is clearly the better car. Saying anything else would be untrue.	pos_neut
6	If you don't have 130,000 dollar buy a Nissan ðŸ˜‚ðŸ˜‚	pos_neut
7	Offroad land cruiser Good rev patrol	neut_neut
8	This Man driver looks like he driving a toy car, look at his hands on the steering wheel and the seat space.	neut_neut
9	Day by day Toyota land cruiser build quality is Down. But Nissan Patrol is still strong.	pos_neg
10	I likes Nissan best cars	pos_neut
11	I would go with the Patrol for stability and better interior	pos_neut
12	Did you guys notice grs lc300	neut_neut
13	The new lc 300 is beautiful and i would recommend it over nissan patrol if you have the money	neg_pos
14	I drive both of them in the Philippines but I patrol much better.I experience I drive the patrol flood area the water on my seat already almost 1 hour in flood area but the patrol I go home safe and the patrol still strong	pos_neut
15	Why fix it if it ain't broken keeps it affordable	neut_neut
16	Nissan gives Australia the second shit box interior whilst the rest of the world gets the upgraded interior, that is a slap in the face and does me out of the Nissan.	neg_neut
17	I heard a new Patrol is slated to come out in 2025 with a turbo 6	neut_neut
18	First time i will choose a nissan over a toyota !	pos_neg
19	The Lc 300 has the ugliest rear end in the automotive industry	neut_neg
20	west or east land cruiser is the beast	neut_pos

HP Vs Dell		
ID	**Text**	**Opinion Label**
1	How can this guy say that upgrades are not important?	nuet_neut
2	dell for the win..perfect portable size for me ..regarding the ports, a usb c hub that consist all other essential ports solves the issue..	neut_pos
3	I have used the Dell xps 13 plus for about a week now and I would have to say its the best small laptop that I have ever owned. Very easy & comfortable to use and so light & small. For me its almost perfect, I'll give it a 9.9 out of 10 bcuz of the lack of earphone jack.	neut_pos
4	Hp spectre is the winner. We can't compare 360 2 in 1 spectra with normal 180 laptop. Obviously 2in 1 has to be thicker. Hp spectre is the winner ???	pos_neg
5	So Dell it is for the categories i care about..appreciate it	neut_pos
6	Everything from HP is ugly.	neg_neut
7	I was a HP laptop user in the past, and now I am a Dell laptop user. The main reason I switched is because of the lack of services HP provides. But, after looking at the Spectre x360 13 at best buy, I wanted to give it a try, and when I compared the Spectre x360 13 to the Dell XPS 13 plus, I liked the Dell more (my personal choice) because I barely us external thumb drives or conventional USB connectors (I store everything in the cloud). When I decided to return the HP Laptop, they charged me 10% (~$125) as a restocking fee, which is not at all fair for such a top brand company, and if I wanted to return a Dell laptop it is hassle free, and their services are top notch, and one example I can say is the shipping time, for HP to ship and deliver it took minimum 5 business days, and dell delivered the laptop the very next day I placed the order, and I had to replace the Dell laptop that I first bought because of the miss alignment of one of the buttons on the keyboard, and the replacement one arrived on the 3rd business day from the day I placed the order. Finally the services between Dell and HP are like Heaven and Hell, and I wouldn't recommend HP over Dell, because their customer service sucked. So, if you don't care about the customer service, and don't mind HP breaking your bank by charging a hefty restocking fee from your wallet, then buy an HP laptop, and if you want top notch services then go for dell.	neut_neut
8	I bought the Spectre and I love it!	pos_neut
9	Tried the XPS 13 plus keyboard, hated it. You can't tell where the tips of your fingers are at times.	neut_neg
10	I've had two spectre x360 , both got about 3-4 hours battery on full brightness watching youtube so there is that	pos_neut
11	Does the HP Spectre actually come with the docking hub in the box? I purchased one with the exact specs but it did not come with the hub. HP says it does not include the hub in the box it is separate only as a bundle deal.	neut_neut
12	Dell build quality, historically, has always exceeded that of HP.	neg_pos
13	After 6 months of usage, the most disadvantage is the battery running away so fast maybe last for maximum 6hr working from 100% to 20%. I am really very upset.	neg_neg
14	Whoever design the XPS 13 Plus is definitely someone who goes after bimbo in real life. (Aesthetic > Functionality)	neut_neg
15	Obviously HP cannot keepup in laptop race with their big border screen. Gladly, they're still have some hardcore fanbase.	neg_neut

iPhone vs Samsung		
	Text	Opinion Label
1	Years ago I had 4s then the 6 and for some time the 7. I loved the 6. Then I tried Galaxy S8 then the GS9. Loved them. Then I missed the iOS eco system so I got the 11. Love it. Got myself the 14 before I traed back to the 11. Thinking about switching to the Pixel line up. For now itâ€™s iPhone all the way. Because the phones last long and I no longer need a lot of bells and whistles like I use to.	pos_pos
2	.....samsung user since at my 10 grade....then went for J7 series..currently using s22 plus (great devise)and wanna sticking with prior choice in following yrs... Samsung=ðŸ'¥â¤....	neut_pos
3	Samsung s23 I like the decisions I can make on my phone. I'm 65 love being independent with what I want.	neut_pos
4	I love my galaxy, I play around with font, and themes, etc.. and I'm 66 years old. I have been thinking about going to iPhone, but really can't make up my mind. Just for a change, you know!	pos_pos
5	How about buying both of them ðŸ¤" ðŸ˜ so you can end the problem of which is better .	pos_pos
6	I've been a loyal Samsung cellphones customer every since i had first job in 2016 when i was 17 years old when i bought my Samsung Galaxy J3. Fast forward few years later, I still use a Samsung Galaxy phone.	neut_pos
7	The only thing that keeps me away from samsung is the luck of privacy. I use iPhone but iâ€™ve tried some galaxyâ€™s that I had at home and I was really shocked by the fact that the Samsung came with Facebook and other Google apps that I couldnâ€™t really uninstall. Although Apple isnâ€™t that privacy friendly either, I still feel that itâ€™s better than Samsung in that.	pos_neg
8	iPhone â¤	pos_neut
9	I'm going to samsung cause they let you Customize your phone how you want it	neut_pos
10	I feel like apple systems are a lot more simple, and less complex. Which is better for me.	pos_pos
11	I love Samsung, but man its incredibly annoying when Im sent photos and videos from iPhone users. It always makes me wanna switch back, but I just dont want to	pos_pos
12	The whole Samsung versus Apple comparison is crazy, for one most of them are done by content creators, so if the iPhone supposedly does better video recording, which phone do you think a content creator is going to choose, and also if a person is not able to use a smartphone obviously you do not need a smartphone, people get everything else in life and want to make it their only, a car, a house, but they will get a phone and want it plain and simple like the iPhone, that is crazy to me smh	pos_pos
13	I would have iPhone	pos_pos
14	I've always told people that apple vs android is always subjective like I could write an article on this as a well rounded person that knows the differences and experiences with both phones.	neut_neut
15	Iâ€™ve been one of those who keeps switching between Samsung and IPhone. Samsung is a better phone for me so much customisation whereas IPhone has better icons, ringtones and IMessage which is great.	pos_pos
	I just went to the galaxy s23 about a month ago and I kind of regret leaving iphone. It just felt more user friendly where as this s23 feels more finicky.. I like both but the iphone just had	

Raila Odinga Vs. William Ruto							
		OPINION LABELS					
ID	COMPARATIVE OPINION TEXT	RAO POS	RAO NEG	RAO NEU	RUTO POS	RUTO NEG	RUTO NEU
1	Raila has been holding public barazas with an aim of delegitimizing the presidency of William Ruto. https://t.co/MM9e90gkEx	0	1	0	1	0	1
2	@OmwambuKE @Wakabando You just said Raila is hated by more than half of the country? Raila beat Ruto in all regions except Mt Kenya and RV... Many of Raila's supporters did not turn up to vote but they're with him.	1	0	1	0	1	0
3	180a. Smears, Torture, Deathreats: Prayers for Humanitarian Visa, Asylum Protection, Self-Determination and Rule-of-Law from Ruto, Uhuru, Raila, co-conspirators and their Jinn-Theocracy GoK Sanctioned political-legal abuse, delayed-denied justice, outla... https://t.co/JChFZeMzY4 https://t.co/xQnpGohw1I	1	0	1	1	0	0
4	@isaac_otwoma @RailaOdinga Ruto William very early in the morning..I can never vote for anyone who has been associated with @RailaOdinga .. even if pastor ng'ang'a runs against Raila I will vote in ng'ang'a...	0	1	0	0	0	1
5	2018-22,Uhuru to Ruto; lets stop politics and work first. Ruto; Hatupangwingwi!!! 2023, Ruto to Raila; stop rallies and let me work. Raila; Ruto must go!!! What goes around comes around. Do to others what you would like others do to you.	1	0	1	0	1	0
6	@RailaOdinga Raila you are old, you are tired, just go home retire. Leave kalonzo and Kalua to fight Ruto. Stop being delusional	0	1	0	1	0	1
7	Peter Salasya you think we don't know you? Umeanza kuvaa nguo juzi Sasa unajiskia umefika. Salasya abandoned his family completely and he's now busy abusing women. ðŸ¤"ðŸ¤" President Ruto, Sonko and Toto must be respected by Raila's cows. https://t.co/9tYDy5P5Bs	0	1	0	1	0	0
8	President Ruto, please roll up your sleeves and work. Just ignore Raila unless your fear him so much.	0	1	0	1	0	0
9	Sasa Raila, how does it help the country you telling us your votes were stolen. As @ReubenKigame has said, if your mission on those rallies isn't the plight of Kenyans, go home. Take on Ruto to lower standards of living siyo siasa ya Kila siku Aisha Jumwa Danston Omari Pkosing	0	1	0	1	0	0
10	@karoba_john Raila won. It's so clear. Everyone in the world was bought by Ruto to sanitize this. Even Uhuru was bribed to step down otherwise he would have given the sword to Baba as agreed. He won 80% in Mt Kenya and 60% in Rift Valley. Baba is loved. E	1	0	1	0	1	0
11	@Wakabando The only problem with u is hypocrisy did all this you have stated here start with Ruto ? Of course not ! Give Ruto a chance of 1 year because you kept quiet when uhuru n raila were rapping this country financially.	1	0	1	1	0	1
12	Baba okays Azimio governors to meet with President Ruto. https://t.co/jtLiS44NnF	1	0	1	1	0	1
13	Begging for mercy, Kioni's 59% votes for Raila presidency is the punisher in his soul, the result boil down to vote of no confidence by 30 Jubilee MPs who suspended him and his brother Murathe and voe to join Ruto UDA in mass, Kanini is acting as Jubilee SG. https://t.co/bEsAH374FD	1	0	1	1	0	1
14	@Kenyans So governors should meet president Ruto but not MPs ?? I think I know why!! Itâ€™s because governors control big budget and Raila like Judas lives on syphoning and drawing by large and big from that kitty which unfortunately MPs donâ€™t have!!	0	1	0	1	0	1
15	Joseph Kinyua is back. William Ruto will soon realize that favoritism and political appointments are a drawback to his administration. And before the end of next year, many Uhuru's men will be back. Raila Odinga will also get the handshake he's been longing for. https://t.co/aTWgrjLfgI	1	0	1	0	1	0

Appendix IV – Data for Hypothesis Testing

Data for Experiment H1 – Single ML Models

Hybrid ML Model	Dataset	Accuracy
LR	D1	86
SGD	D1	86.6
MNB	D1	80.4
KNN	D1	47.1
SVM	D1	76.6
DT	D1	83.9
RF	D1	85.3
MLP	D1	86.2
LR	D2	92
SGD	D2	91.8
MNB	D2	88.8
KNN	D2	53.7
SVM	D2	86.3
DT	D2	91.8
RF	D2	92.3
MLP	D2	92.6
LR	D3	78.2
SGD	D3	78.6
MNB	D3	73.5
KNN	D3	55.9
SVM	D3	74.1
DT	D3	76
RF	D3	76.7
MLP	D3	77.9

Data for Hypothesis 2 – Hybrid ML Models

Hybrid ML Model	Dataset	Accuracy
MLP + DT	D1	86
MLP + DT	D2	92
MLP + DT	D3	98.4
MLP + RF	D1	86.6
MLP + RF	D2	92.6
MLP + RF	D3	99.7
MLP + SGD	D1	86.9
MLP + SGD	D2	92.3
MLP + SGD	D3	91.8
MLP + SVM	D1	85.6
MLP + SVM	D2	91.9
MLP + SVM	D3	99.7
SGD + DT	D1	63.1
SGD + DT	D2	92
SGD + DT	D3	99.7
SGD + MLP	D1	84.6
SGD + MLP	D2	91
SGD + MLP	D3	95.7
SGD + RF	D1	85.6
SGD + RF	D2	91.3
SGD + RF	D3	100
SGD + SVM	D1	85.7
SGD + SVM	D2	92.1
SGD + SVM	D3	100

Data for Hypothesis 5 – Feature Extraction Techniques vs Accuracy

Hybrid ML Model	Feature Extraction Technique	Accuracy
MLP + RF	CV1	92.0
MLP + RF	CV2	92.8
MLP + RF	CV3	92.9
SGD + RF	CV1	91.3
SGD + RF	CV2	90.8
SGD + RF	CV3	92.3
MLP + RF	TF1	91.1
MLP + RF	TF2	92.6
MLP + RF	TF3	92.7
SGD + RF	TF1	91.3
SGD + RF	TF2	90.8
SGD + RF	TF3	92.3
MLP + RF	CB1	43.4
MLP + RF	CB5	43.0
SGD + RF	CB1	42.3
SGD + RF	CB5	42.0

Annotations:

CV - Count Vectorizer; TF - TFIDF; CB = CBOW

The No. after the annotation is the n-gram range or window-size

All the Data for Hypothesis Testing

ML Model	Dataset	Feature Extraction Technique (FET)	Accuracy (%)
LR	D1	CV3	86
SGD	D1	CV3	85
DT	D1	CV3	84
RF	D1	CV3	85
SVM	D1	CV3	77
KNN	D1	CV3	47
GNB	D1	CV3	80
MLP	D1	CV3	86
LR	D1	TFIDF3	66
SGD	D1	TFIDF3	86
DT	D1	TFIDF3	85
RF	D1	TFIDF3	86
SVM	D1	TFIDF3	81
KNN	D1	TFIDF3	54
GNB	D1	TFIDF3	80
MLP	D1	TFIDF3	86
LR	D1	CBOW5	42
SGD	D1	CBOW5	14
DT	D1	CBOW5	42
RF	D1	CBOW5	42
SVM	D1	CBOW5	39
KNN	D1	CBOW5	42
GNB	D1	CBOW5	4.1
MLP	D1	CBOW5	42
LR	D1	SkipGram	42
SGD	D1	SkipGram	26
DT	D1	SkipGram	42
RF	D1	SkipGram	42

SVM	D1	SkipGram	39
KNN	D1	SkipGram	42
GNB	D1	SkipGram	5
MLP	D1	SkipGram	42
MLP + DT	D1	CV3	86
MLP + RF	D1	CV3	87
MLP + SGD	D1	CV3	87
MLP + SVM	D1	CV3	86
SGD + DT	D1	CV3	63
SGD + MLP	D1	CV3	85
SGD + RF	D1	CV3	86
SGD + SVM	D1	CV3	86
MLP + DT	D1	TFIDF3	87
MLP + RF	D1	TFIDF3	86
MLP + SGD	D1	TFIDF3	86
MLP + SVM	D1	TFIDF3	85
SGD + DT	D1	TFIDF3	85
SGD + MLP	D1	TFIDF3	87
SGD + RF	D1	TFIDF3	86
SGD + SVM	D1	TFIDF3	65
LR	D2	CV3	92
SGD	D2	CV3	92
DT	D2	CV3	92
RF	D2	CV3	92
SVM	D2	CV3	86
KNN	D2	CV3	53
GNB	D2	CV3	90
MLP	D2	CV3	92
LR	D2	TFIDF3	75
SGD	D2	TFIDF3	93

DT	D2	TFIDF3	92
RF	D2	TFIDF3	93
SVM	D2	TFIDF3	91
KNN	D2	TFIDF3	53
GNB	D2	TFIDF3	89
MLP	D2	TFIDF3	92
LR	D2	CBOW5	39
SGD	D2	CBOW5	40
DT	D2	CBOW5	40
RF	D2	CBOW5	40
SVM	D2	CBOW5	30
KNN	D2	CBOW5	40
GNB	D2	CBOW5	24
MLP	D2	CBOW5	40
LR	D2	SkipGram	42
SGD	D2	SkipGram	26
DT	D2	SkipGram	42
RF	D2	SkipGram	42
SVM	D2	SkipGram	39
KNN	D2	SkipGram	42
GNB	D2	SkipGram	5
MLP	D2	SkipGram	42
MLP + DT	D2	CV3	92
MLP + RF	D2	CV3	93
MLP + SGD	D2	CV3	92
MLP + SVM	D2	CV3	92
SGD + DT	D2	CV3	92
SGD + MLP	D2	CV3	91
SGD + RF	D2	CV3	91
SGD + SVM	D2	CV3	92

MLP + DT	D2	TFIDF3	92
MLP + RF	D2	TFIDF3	92
MLP + SGD	D2	TFIDF3	92
MLP + SVM	D2	TFIDF3	92
SGD + DT	D2	TFIDF3	92
SGD + MLP	D2	TFIDF3	92
SGD + RF	D2	TFIDF3	94
SGD + SVM	D2	TFIDF3	93
LR	D3	CV3	100
SGD	D3	CV3	100
DT	D3	CV3	99
RF	D3	CV3	100
SVM	D3	CV3	100
KNN	D3	CV3	100
GNB	D3	CV3	100
MLP	D3	CV3	100
LR	D3	TFIDF3	100
SGD	D3	TFIDF3	100
DT	D3	TFIDF3	99
RF	D3	TFIDF3	100
SVM	D3	TFIDF3	100
KNN	D3	TFIDF3	100
GNB	D3	TFIDF3	100
MLP	D3	TFIDF3	100
LR	D3	CBOW5	46
SGD	D3	CBOW5	37
DT	D3	CBOW5	46
RF	D3	CBOW5	46
SVM	D3	CBOW5	33
KNN	D3	CBOW5	46

GNB	D3	CBOW5	2.3
MLP	D3	CBOW5	46
LR	D3	SkipGram	49
SGD	D3	SkipGram	44
DT	D3	SkipGram	49
RF	D3	SkipGram	49
SVM	D3	SkipGram	35
KNN	D3	SkipGram	49
GNB	D3	SkipGram	39
MLP	D3	SkipGram	49
MLP + DT	D3	CV3	98
MLP + RF	D3	CV3	100
MLP + SGD	D3	CV3	92
MLP + SVM	D3	CV3	100
SGD + DT	D3	CV3	100
SGD + MLP	D3	CV3	96
SGD + RF	D3	CV3	100
SGD + SVM	D3	CV3	100
MLP + DT	D3	TFIDF3	99
MLP + RF	D3	TFIDF3	100
MLP + SGD	D3	TFIDF3	100
MLP + SVM	D3	TFIDF3	100
SGD + DT	D3	TFIDF3	100
SGD + MLP	D3	TFIDF3	100
SGD + RF	D3	TFIDF3	100
SGD + SVM	D3	TFIDF3	100

Appendix V – Posthoc Analysis Results

Posthoc Analysis for Hypothesis: One-Way ANOVA for ML Models vs Accuracy

Post Hoc Comparisons - ML Model

Comparison							
ML Model		**ML Model**	**Mean Difference**	**SE**	**df**	**t**	**p_{tukey}**
LR	-	SGD	-6.167	5.99	80.0	-1.0302	1.000
	-	DT	-5.333	5.99	80.0	-0.8910	1.000
	-	RF	-6.167	5.99	80.0	-1.0302	1.000
	-	SVM	-2.667	5.99	80.0	-0.4455	1.000
	-	KNN	18.667	5.99	80.0	3.1184	0.147
	-	GNB	-3.333	5.99	80.0	-0.5569	1.000
	-	MLP	-6.167	5.99	80.0	-1.0302	1.000
	-	MLP + DT	-5.833	5.99	80.0	-0.9745	1.000
	-	MLP + RF	-6.500	5.99	80.0	-1.0859	0.999
	-	MLP + SGD	-5.000	5.99	80.0	-0.8353	1.000
	-	MLP + SVM	-6.000	5.99	80.0	-1.0024	1.000
	-	SGD + DT	-2.167	5.99	80.0	-0.3620	1.000
	-	SGD + MLP	-5.333	5.99	80.0	-0.8910	1.000
	-	SGD + RF	-6.333	5.99	80.0	-1.0580	0.999
	-	SGD + SVM	-2.833	5.99	80.0	-0.4733	1.000
SGD	-	DT	0.833	5.99	80.0	0.1392	1.000
	-	RF	4.44e-16	5.99	80.0	7.42e-17	1.000
	-	SVM	3.500	5.99	80.0	0.5847	1.000
	-	*KNN*	*24.833*	*5.99*	*80.0*	*4.1486*	*0.008*
	-	GNB	2.833	5.99	80.0	0.4733	1.000

Post Hoc Comparisons - ML Model

Comparison							
ML Model		**ML Model**	**Mean Difference**	**SE**	**df**	**t**	**p_{tukey}**
	-	MLP	5.33e-15	5.99	80.0	8.90e-16	1.000
	-	MLP + DT	0.333	5.99	80.0	0.0557	1.000
	-	MLP + RF	-0.333	5.99	80.0	-0.0557	1.000
	-	MLP + SGD	1.167	5.99	80.0	0.1949	1.000
	-	MLP + SVM	0.167	5.99	80.0	0.0278	1.000
	-	SGD + DT	4.000	5.99	80.0	0.6682	1.000
	-	SGD + MLP	0.833	5.99	80.0	0.1392	1.000
	-	SGD + RF	-0.167	5.99	80.0	-0.0278	1.000
	-	SGD + SVM	3.333	5.99	80.0	0.5569	1.000
DT	-	RF	-0.833	5.99	80.0	-0.1392	1.000
	-	SVM	2.667	5.99	80.0	0.4455	1.000
	-	*KNN*	*24.000*	*5.99*	*80.0*	*4.0094*	*0.012*
	-	GNB	2.000	5.99	80.0	0.3341	1.000
	-	MLP	-0.833	5.99	80.0	-0.1392	1.000
	-	MLP + DT	-0.500	5.99	80.0	-0.0835	1.000
	-	MLP + RF	-1.167	5.99	80.0	-0.1949	1.000
	-	MLP + SGD	0.333	5.99	80.0	0.0557	1.000
	-	MLP + SVM	-0.667	5.99	80.0	-0.1114	1.000
	-	SGD + DT	3.167	5.99	80.0	0.5290	1.000
	-	SGD + MLP	-4.88e−15	5.99	80.0	-8.16e−16	1.000
	-	SGD + RF	-1.000	5.99	80.0	-0.1671	1.000
	-	SGD + SVM	2.500	5.99	80.0	0.4176	1.000
RF	-	SVM	3.500	5.99	80.0	0.5847	1.000

Post Hoc Comparisons - ML Model

Comparison							
ML Model		**ML Model**	**Mean Difference**	**SE**	**df**	**t**	**p_{tukey}**
	-	*KNN*	*24.833*	*5.99*	*80.0*	*4.1486*	*0.008*
	-	GNB	2.833	5.99	80.0	0.4733	1.000
	-	MLP	4.88e-15	5.99	80.0	8.16e-16	1.000
	-	MLP + DT	0.333	5.99	80.0	0.0557	1.000
	-	MLP + RF	-0.333	5.99	80.0	-0.0557	1.000
	-	MLP + SGD	1.167	5.99	80.0	0.1949	1.000
	-	MLP + SVM	0.167	5.99	80.0	0.0278	1.000
	-	SGD + DT	4.000	5.99	80.0	0.6682	1.000
	-	SGD + MLP	0.833	5.99	80.0	0.1392	1.000
	-	SGD + RF	-0.167	5.99	80.0	-0.0278	1.000
	-	SGD + SVM	3.333	5.99	80.0	0.5569	1.000
SVM	-	*KNN*	*21.333*	*5.99*	*80.0*	*3.5639*	*0.046*
	-	GNB	-0.667	5.99	80.0	-0.1114	1.000
	-	MLP	-3.500	5.99	80.0	-0.5847	1.000
	-	MLP + DT	-3.167	5.99	80.0	-0.5290	1.000
	-	MLP + RF	-3.833	5.99	80.0	-0.6404	1.000
	-	MLP + SGD	-2.333	5.99	80.0	-0.3898	1.000
	-	MLP + SVM	-3.333	5.99	80.0	-0.5569	1.000
	-	SGD + DT	0.500	5.99	80.0	0.0835	1.000
	-	SGD + MLP	-2.667	5.99	80.0	-0.4455	1.000
	-	SGD + RF	-3.667	5.99	80.0	-0.6125	1.000
	-	SGD + SVM	-0.167	5.99	80.0	-0.0278	1.000
KNN	-	*GNB*	*-22.000*	*5.99*	*80.0*	*-3.6753*	*0.034*

Post Hoc Comparisons - ML Model

Comparison							
ML Model		**ML Model**	**Mean Difference**	**SE**	**df**	**t**	**p_{tukey}**
	-	*MLP*	*-24.833*	*5.99*	*80.0*	*-4.1486*	*0.008*
	-	*MLP + DT*	*-24.500*	*5.99*	*80.0*	*-4.0929*	*0.009*
	-	*MLP + RF*	*-25.167*	*5.99*	*80.0*	*-4.2043*	*0.006*
	-	*MLP + SGD*	*-23.667*	*5.99*	*80.0*	*-3.9537*	*0.014*
	-	*MLP + SVM*	*-24.667*	*5.99*	*80.0*	*-4.1208*	*0.008*
	-	SGD + DT	-20.833	5.99	80.0	-3.4804	0.059
	-	*SGD + MLP*	*-24.000*	*5.99*	*80.0*	*-4.0094*	*0.012*
	-	*SGD + RF*	*-25.000*	*5.99*	*80.0*	*-4.1765*	*0.007*
	-	*SGD + SVM*	*-21.500*	*5.99*	*80.0*	*-3.5918*	*0.043*
GNB	-	MLP	-2.833	5.99	80.0	-0.4733	1.000
	-	MLP + DT	-2.500	5.99	80.0	-0.4176	1.000
	-	MLP + RF	-3.167	5.99	80.0	-0.5290	1.000
	-	MLP + SGD	-1.667	5.99	80.0	-0.2784	1.000
	-	MLP + SVM	-2.667	5.99	80.0	-0.4455	1.000
	-	SGD + DT	1.167	5.99	80.0	0.1949	1.000
	-	SGD + MLP	-2.000	5.99	80.0	-0.3341	1.000
	-	SGD + RF	-3.000	5.99	80.0	-0.5012	1.000
	-	SGD + SVM	0.500	5.99	80.0	0.0835	1.000
MLP	-	MLP + DT	0.333	5.99	80.0	0.0557	1.000
	-	MLP + RF	-0.333	5.99	80.0	-0.0557	1.000
	-	MLP + SGD	1.167	5.99	80.0	0.1949	1.000
	-	MLP + SVM	0.167	5.99	80.0	0.0278	1.000
	-	SGD + DT	4.000	5.99	80.0	0.6682	1.000

Post Hoc Comparisons - ML Model

Comparison							
ML Model		**ML Model**	**Mean Difference**	**SE**	**df**	**t**	**p_{tukey}**
	-	SGD + MLP	0.833	5.99	80.0	0.1392	1.000
	-	SGD + RF	-0.167	5.99	80.0	-0.0278	1.000
	-	SGD + SVM	3.333	5.99	80.0	0.5569	1.000
MLP + DT	-	MLP + RF	-0.667	5.99	80.0	-0.1114	1.000
	-	MLP + SGD	0.833	5.99	80.0	0.1392	1.000
	-	MLP + SVM	-0.167	5.99	80.0	-0.0278	1.000
	-	SGD + DT	3.667	5.99	80.0	0.6125	1.000
	-	SGD + MLP	0.500	5.99	80.0	0.0835	1.000
	-	SGD + RF	-0.500	5.99	80.0	-0.0835	1.000
	-	SGD + SVM	3.000	5.99	80.0	0.5012	1.000
MLP + RF	-	MLP + SGD	1.500	5.99	80.0	0.2506	1.000
	-	MLP + SVM	0.500	5.99	80.0	0.0835	1.000
	-	SGD + DT	4.333	5.99	80.0	0.7239	1.000
	-	SGD + MLP	1.167	5.99	80.0	0.1949	1.000
	-	SGD + RF	0.167	5.99	80.0	0.0278	1.000
	-	SGD + SVM	3.667	5.99	80.0	0.6125	1.000
MLP + SGD	-	MLP + SVM	-1.000	5.99	80.0	-0.1671	1.000
	-	SGD + DT	2.833	5.99	80.0	0.4733	1.000
	-	SGD + MLP	-0.333	5.99	80.0	-0.0557	1.000
	-	SGD + RF	-1.333	5.99	80.0	-0.2227	1.000
	-	SGD + SVM	2.167	5.99	80.0	0.3620	1.000
MLP + SVM	-	SGD + DT	3.833	5.99	80.0	0.6404	1.000
	-	SGD + MLP	0.667	5.99	80.0	0.1114	1.000

Post Hoc Comparisons - ML Model

Comparison							
ML Model		**ML Model**	**Mean Difference**	**SE**	**df**	**t**	**p_{tukey}**
	-	SGD + RF	-0.333	5.99	80.0	-0.0557	1.000
	-	SGD + SVM	3.167	5.99	80.0	0.5290	1.000
SGD + DT	-	SGD + MLP	-3.167	5.99	80.0	-0.5290	1.000
	-	SGD + RF	-4.167	5.99	80.0	-0.6961	1.000
	-	SGD + SVM	-0.667	5.99	80.0	-0.1114	1.000
SGD + MLP	-	SGD + RF	-1.000	5.99	80.0	-0.1671	1.000
	-	SGD + SVM	2.500	5.99	80.0	0.4176	1.000
SGD + RF	-	SGD + SVM	3.500	5.99	80.0	0.5847	1.000

Note. Comparisons are based on estimated marginal means

Posthoc Analysis for Hypothesis: One-Way ANOVA for Dataset vs Accuracy

Post Hoc Comparisons - Dataset

Comparison							
Dataset		**Dataset**	**Mean Difference**	**SE**	**df**	**t**	**p_{tukey}**
D1	-	D2	-8.99	1.65	44.0	-5.43	< .001
	-	D3	-15.93	1.65	44.0	-9.63	< .001
D2	-	D3	-6.94	1.63	44.0	-4.26	< .001

Note. Comparisons are based on estimated marginal means

Posthoc Analysis for Hypothesis: One-Way ANOVA for Feature Extraction Techniques (FET) vs Accuracy

Post Hoc Comparisons - FET

Comparison							
FET		**FET**	**Mean Difference**	**SE**	**df**	**t**	**p_{tukey}**
CV3	-	TFIDF3	-0.0833	2.36	140	-0.0353	1.000
	-	CBOW5	53.6729	2.89	140	18.5873	< .001
	-	SkipGram	51.3896	2.89	140	17.7966	< .001
TFIDF3	-	CBOW5	53.7563	2.89	140	18.6162	< .001
	-	SkipGram	51.4729	2.89	140	17.8254	< .001
CBOW5	-	SkipGram	-2.2833	3.33	140	-0.6848	0.903

Note. Comparisons are based on estimated marginal means

Appendix VI – Sample Codes and Outputs

```
#Cleaning reviews for preprocessing
#General cleaning of the reviews
stop = stopwords.words("english")
def clean_data(text):
    text = re.sub("@[A-Za-z0-9_]+","",text) #Remove @ sign
    text = re.sub(r"(?:\@|http?\://|https?\://|www)\S+", "", text) #Remove http links
    #text = re.sub('\W',  " ", text)
    text = " ".join(text.split())
    #comment = ''.join(c for c in comment if c not in emoji.UNICODE_EMOJI) #Remove Emojis
    text = text.replace("#", "").replace("_", " ") #Remove hashtag sign but keep the text
    text = text.lower()
    return text
df['CleanReviews'] = df['Reviews'].apply(lambda x: clean_data(str(x)))
df['CleanReviews'] = df['CleanReviews'].apply(lambda x: " ".join([word for word in x.split() if word not in stop]))
df.head(2)
```

```
#Importing the required libraries
import numpy as np
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.metrics import f1_score, accuracy_score, precision_score, recall_score, classification_report
from sklearn.preprocessing import LabelEncoder
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
from sklearn.linear_model import LogisticRegression, SGDClassifier
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, StackingClassifier
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
from sklearn.naive_bayes import GaussianNB, MultinomialNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
import joblib
import glob
import time
import os
import re
import matplotlib.pyplot as plt
from nltk.corpus import stopwords
import warnings
warnings.filterwarnings("ignore")
```

```
df = pd.read_csv("../Facebook vs Twitter.csv")
df.columns = ['Reviews', "Polarity"]
df['BrandA_Polarity'] = df['Polarity'].apply(lambda x: str(x).split("_")[0])
df['BrandB_Polarity'] = df['Polarity'].apply(lambda x: str(x).split("_")[1])
df.head(2)
```

	Reviews	Polarity	BrandA_Polarity	BrandB_Polarity
0	Facebook has more users then Twitter	pos_neg	pos	neg
1	Twitter and Facebook are adaptable innovative ...	pos_pos	pos	pos

Appendix VII – Publications

Ondara, B., Waithaka, S., Kandiri, J., & Muchemi, L. (2022). Machine Learning Techniques, Features, Datasets, and Algorithm Performance Parameters for Sentiment Analysis: A Systematic Review. *Open Journal for Information Technology*, *5*(1), 1–16. https://doi.org/10.32591/coas.ojit.0501.01001o

Ondara, B., Waithaka, S., Kandiri, J., & Muchemi, L. (2023). Hybrid Machine Learning Techniques for Comparative Opinion Mining. Indonesian Journal of Artificial Intelligence and Data Mining (*IJAIDM)*, *6*(2).

Appendix VIII –Research Authorization

KENYATTA UNIVERSITY
GRADUATE SCHOOL

E-mail: kubps@yahoo.com
dean-graduate@ku.ac.ke
Website: www.ku.ac.ke

P.O. Box 43844, 00100
NAIROBI, KENYA
Tel. 810901 Ext. 57530

Internal Memo

FROM: Dean, Graduate School

DATE: 4th May, 2021

TO: Mr. Bernard O. Ondara
C/o Department of Computing & Information Tech.
KENYATTA UNIVERSITY

REF: J98/25718/18

SUBJECT: APPROVAL OF RESEARCH PROPOSAL

This is to inform you that the Graduate School Board at its meeting 28th April, 2021 approved your Ph.D. Research Proposal entitled "Hybrid Machine Learning Model for Comparative Opinion Mining in Brand Reputation Monitoring".

You may now proceed with your Data collection, subject to clearance with the Director General, National Commission for Science, Technology & Innovation.

As you embark on your data collection, please note that you will be required to submit to Graduate School completed supervision Tracking and Progress Report Forms. The Forms are available at the University's Website under Graduate School webpage downloads.

By copy of this letter, the Registrar (Academic) is hereby requested to grant you substantive registration for your Ph.D. studies.

Thank you.

REUBEN MURIUKI
FOR: DEAN, GRADUATE SCHOOL

c.c. Chairman, Department of Computing & Information Technology
Registrar (Academic) Att; Mr. Richard Chweya

Supervisors:

1. Dr. Stephen Waithaka
C/o Department of Computing & Information Tech.
KENYATTA UNIVERSITY

2. Dr. John Kandiri
C/o Department of Computing & Information Tech.
KENYATTA UNIVERSITY

3. Dr. Lawrence Muchemi
School of Computing & Informatics
University of Nairobi
C/o Department of Computing & Information Tech.
KENYATTA UNIVERSITY

KENYATTA UNIVERSITY
GRADUATE SCHOOL

E-mail: kubps@yahoo.com
dean-graduate@ku.ac.ke
Website: www.ku.ac.ke

P.O. Box 43844, 00100
NAIROBI, KENYA
Tel. 8710901 Ext. 57530

Our Ref: J98/25718/18

Date: 4th May, 2021

The Director General,
National Commission for Science, Technology & Innovation,
P.O. Box 30623-00100,
NAIROBI

Dear Sir/Madam,

RE: RESEARCH AUTHORIZATION FOR BERNARD O. ONDARA - REG. NO. J98/25718/18

I write to introduce Ondara who is a Postgraduate Student of this University. The Student is registered for a Ph.D. degree programme in the **Department of Computing & Information Technology in the School of Engineering.**

Ondara intends to conduct research for Ph.D. thesis entitled, **"Hybrid Machine Learning Model for Comparative Opinion Mining in Brand Reputation Monitoring".**

Any assistance given will be highly appreciated.

Yours faithfully,

PROF. ELISHIBA KIMANI
DEAN, GRADUATE SCHOOL

RM/cao

Appendix IX – Research License

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: 268482

Date of Issue: 17/June/2021

RESEARCH LICENSE

This is to Certify that Mr.. Bernard Ondara of Kenyatta University, has been licensed to conduct research in Kajiado, Kiambu, Machakos, Nairobi, Nakuru on the topic: Hybrid Machine Learning Model for Comparative Opinion Mining in Brand Reputation Monitoring for the period ending : 17/June/2022.

License No: NACOSTI/P/21/11203

268482

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION