International Journal of Computing Sciences Research (ISSN print: 2546-0552; ISSN online: 2546-115X)
Vol. 9, pp. 3392- 3417
doi: 10.25147/ijcsr.2017.001.1.224
https://stepacademic.net

Short Paper

Empirical Analysis of the State-of-the-Art Models for Handling Polarity Shifts Due to Implicit Negation in Mobile Phone Reviews

Millicent K. Murithi
Department of Computer Science, Murang'a University of Technology, Kenya
mkathambi@mut.ac.ke
(corresponding author)

Aaron M. Oirere
Department of Computer Science, Murang'a University of Technology, Kenya

Rachael N. Ndung'u
Department of Information Technology, Murang'a University of Technology, Kenya

Date received: April 23, 2024
Date received in revised form: September 17, 2024
Date accepted: October 15, 2024

Recommended citation:

> Murithi, M. K., Oirere, A. M., & Ndung'u, R. N. (2025). Empirical analysis of the state-of-the-art models for handling polarity shifts due to implicit negation in mobile phone reviews. *International Journal of Computing Sciences Research*, 9, 3392-3417. https://doi.org/10.25147/ijcsr.2017.001.1.224

Abstract

Purpose – This paper presents a comprehensive empirical analysis focusing on sentiment flux within state-of-the-art models designed for handling polarity shifts due to implicit negation in Amazon mobile phones' reviews.

Method – The research evaluates diverse models across five categories: traditional machine learning (ML), deep learning (DL), and hybrid models combining both approaches. Various feature extraction, feature selection, and data augmentation techniques are tested on Amazon mobile phone reviews dataset. BERT and LSTM are used for deep learning while SVM and Naive Bayes are used for traditional ML. ANOVA is used to identify the presence or absence of significant differences and interactions among these entities.

Results – DL shows superior performance compared to traditional ML models. ANOVA analysis shows significant performance differences between conventional ML and DL

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

models. Traditional ML models interact significantly with feature extraction and selection techniques while DL models do not. Traditional ML models do not interact significantly with data augmentation methods while DL models do. FastText extraction outperforms word2vec; Back translation outperforms synonym replacement while recursive feature selection (RFE) surpasses TF-IDF (Term Frequency-Inverse Document Frequency). The BERT and LSTM exhibit one of the strongest performances.

Conclusion – The study concludes that DL models are more effective. Data augmentation techniques significantly impact the performance of DL models, with back translation showing superior performance over synonym replacement. This provides a leverage point in developing an improved model in the future.

Recommendations – Future research should focus on developing a hybrid model for Enhanced Polarity Shift Management of Mobile Phone Reviews using Contextual Back Translation Augmented by Seq2seq Perturbations. This aims at leveraging contextual back translation and Seq2seq perturbations to generate a diverse interpretation that consequently improves the model's ability to handle nuanced expressions of sentiments due to implicit negation with enhanced accuracy, generalizability, robustness to polarity shifts, and contextual understanding.

Research Implications – The findings provide valuable insights into the development of state-of-the-art models, offering a promising direction for further research in sentiment analysis.

Keywords – empirical analysis, hybrid, perturbations, implicit negation, sentiment flux

INTRODUCTION

Sentiment analysis, a pivotal aspect of Natural Language Processing (NLP), plays a crucial role in deciphering the attitude expressed in textual content (Khan et al., 2016). Understanding and accurately interpreting sentiment in textual data is challenging yet crucial (Wankhade et al., 2022). The subtleties and dynamism embedded in human language often led to complexities, especially when it comes to handling implicit negation and polarity shifts (Israel,2011). Implicit negation, wherein the sentiment conveyed is contrary to the literal meaning of the words used, adds an extra layer of intricacy to sentiment analysis (Kumar & Garg, 2020). Recent advancements in the field of NLP have witnessed the emergence of sophisticated models that leverage deep learning techniques, pre-trained embeddings and attention mechanisms (Torfi et al., 2020). These models, often touted as the pinnacle of sentiment analysis, claim to possess an inherent understanding of context and semantic complexities (Xiang et al., 2015). However, the effectiveness of these models in handling sentiment flux induced by implicit negation remains an area that demands rigorous investigation (Van de Kauter et al., 2015).

Considering this, the primary objective of this empirical analysis is two-fold. Firstly, we aim to scrutinize the efficacy of a variety of state-of-the-art sentiment analysis models in accurately capturing sentiment shifts caused by implicit negation. By subjecting these models to a benchmark dataset containing implicit negations, varied feature extraction, feature selection and data augmentation techniques, we seek to uncover the differences in their performance in adapting to the dynamic nature of sentiment expression and polarity shift. The exploration of sentiment flux and the dynamic changes in sentiment polarity is essential for enhancing the robustness and accuracy of sentiment analysis models (Marrapu et al., 2024). As language evolves and adapts to new contexts, the ability of models to discern sentiment flux becomes paramount for applications ranging from customer feedback analysis to social media monitoring (Moon et al., 2021). Secondly, we endeavour to contribute insights, based on the statistical analysis of the experimental data, that shed more light and provide a deeper understanding of the underlying mechanisms influencing sentiment prediction. Through this exploration, we aim to pave the way for advancements in sentiment analysis methodologies, fostering the development of improved models that can adeptly navigate the intricate landscape of sentiment flux induced by implicit negation.

LITERATURE REVIEW

This section focuses on the existing sentiment analysis models that are used to handle polarity shift as well as the various feature extraction, feature selection and data augmentation techniques used with these models. Using numerous technologies, the internet has become more accessible in the modern era, making it possible for people to read product and service reviews and share ideas (Astya, 2017). It is noted that polarity shift is one of the problems with sentiment analysis (Xia et al., 2016). Polarity shifters are the factors that can change a word's prior polarity in one of three ways i.e. rise, decrease, or neutral (Xia et al., 2016). Madhuli and Rahuli (2020) note that there is a need to consider all the factors that are responsible for polarity shifts. These polarity shifters can have a detrimental effect on the classification performance of a sentiment analysis model used for handling polarity shifts if not detected and handled effectively (Xia et al., 2016). Some of the polarity shifters include negation, contrast, context, ambiguity, evolving language, subjectivity, cultural and regional differences, news and events, authors tones, review fraud, pragmatics and domain-specific knowledge (Abdi et al., 2019; Eke et al., 2021; Japhne & Murugeswari, 2020). Sentiment analysis methods are categorized as Lexicon, machine learning and hybrid-based. Lexicon methods generate a list of negative and positive terms to deduce any polarity in a message automatically or manually. Machine learning such as Support Vector Machine, Naïve Bayes, Random Forest and Maximum Entropy have been mainly used in sentiment analysis (Abirami & Gayathri, 2015). Various researchers have used machine learning methods due to the availability of labelled data. D'Souza and Sonawane (2019) developed a Dual Sentiment Analysis (DSA) method to handle both actual reviews and reversed reviews. Liu (2020) in their study combined CNN-LSTM to achieve better classification. CNN allowed for sequential features to be fed into LSTM which handled the long-dependence of the sentiments.

Feature selection techniques have been used in sentiment analysis models to minimize the dimensionality of the feature space and identify the most important features (Mhatre et al., 2017; Pongthanoo & Songpan, 2020; Prastyo et al., 2020). The feature selection phase in sentiment analysis approaches varies greatly. Some use metaheuristic techniques while others employ statistical and filter techniques (Miao & Niu, 2016). There are three types of feature selection techniques in sentiment analysis. These include filter, wrapper, and hybrid approaches. Mohd Nafis and Awang (2021) note that the TF-IDF (Term Frequency and Inverse Document Frequency) approach has been used as a feature selection to reduce the influence of irrelevant words in a document. Other researchers like Japhne and Murugeswari (2020), Ahmad and Aftab (2017), and Kumar and Garg (2023) have also used the TF IDF technique in their work.

Hegde and Seema (2017) propose the use of the bag-of-words (BOW) method in NLP which expresses a text document as an array of fixed-length vectors. Kumar and Rajini (2019) propose the use of word embedding that includes Word2Vec, GloVe, and FastText as feature extraction techniques. The "word embeddings" feature extraction methods are applied to several sentiment analysis models used in handling polarity shift problems. These include studies by Avinash and Sivasankar (2019), Singh and Paul (2021), Deho et al. (2018), and Rajabi et al. (2020).

Data augmentation is a technique that has been used in recent years to automatically generate more training data (Sennrich et al., 2016). Sugiyama and Yoshinaga (2019) used the back translation technique to improve the performance of the translation model by generating more training data. Fadaee et al. (2016) proposed a novel approach that augments the training by generating new sentence pairs containing rare words. Kobayashi (2018) proposed contextual augmentation to stochastically replace words with other words. Hou et al. (2018) used a sequence-to-sequence model to augment the training data by generation of lexical and syntactic alternatives. Duong and Nguyen-Thi (2021) note that back translation and Syntax-Tree transformation are the data augmentation techniques that have the potential to improve sentiment polarity classification.

Within the scope of these studies, it is noted that improved sentiment classification accuracy is achieved with hybrid algorithms compared to conventional machine learning and simple deep learning models. Additionally, it is notable that Word2Vec, Glove, FastText and BoW are the most prevalent feature extraction methods (Salur & Aydin, 2020). It is noted that the main disadvantage of Word2Vec and Glove is their ability to generate a random vector in a word, not in the dataset. FastText on the other hand is a continuation of Word2Vec that can overcome this disadvantage. Salur and Aydin (2020), Long et al. (2019), Tang et al. (2015), and Yang et al. (2016) note that deep learning is the most prevalent sentiment with LSTM, GRU, BiLSTM and CNN being the most prevalent analysis models.

METHODOLOGY

This section discusses the steps that were followed when carrying out the empirical analysis.

Research Design

The research study is based on the positivist research philosophy. Positivism research philosophy is a school of thought that confines itself to the fact that the knowledge of a specific phenomenon is based on what can be observed, measured and recorded, in the same way as in natural science (Howell, 2012). On the other hand, the empirical research design is used to conduct and analyze the experiments quantitatively. The research objectives, as well as the hypothesis, guide the empirical research design process and have been formulated as follows:

- **RO1:** How does the performance of the different categories of sentiment analysis models compare in a variety of feature extraction, feature selection and data augmentation techniques?
- **RO2:** What practical implications arise from the analysis of the results of the experimental data on the performance of different categories of sentiment analysis models?
- **RO3:** How can the valuable insights collected from the analysis of the experimental data be leveraged in the development of a novel sentiment analysis model for polarity shift management due to implicit negation in mobile phone reviews?

Consequently, the research hypotheses were formulated as follows:

- **Ha0:** There are significant differences among the different categories of the sentiment analysis models used for handling polarity shift in a variety of feature extraction, feature selection and data augmentation techniques.
- **Ha1:** There are no significant differences among the different categories of the sentiment analysis models used for handling polarity shift in a variety of feature extraction, feature selection and data augmentation techniques.

Experimental Variables

The experimental algorithms, feature extraction methods, feature selection methods and data augmentation techniques consist of the state of the art used in handling implicit negations. The research adopts a comprehensive approach by evaluating diverse sentiment analysis models across five categories: Traditional Machine Learning (ML) Models, Deep Learning Models, Hybrid Models combining traditional ML and deep learning, Hybrid Models exclusively leveraging deep learning methodologies and lastly, the Hybrid Models exclusively utilizing traditional ML approaches The specific Traditional Machine Learning and Deep Learning Models include: BERT (Bidirectional Encoder Representations from Transformers), Long Short Term Memory (LSTM), Naïve Bayes and SVM (Support Vector Machines). The feature extraction methods include Fast Text and Word2Vec while the feature selection methods include the wrapper-based recursive feature selection (RFE) performs better than TF-IDF. Lastly, Back translation and synonym replacement are used as the data augmentation methods. The choice of these experimental variables is in line with the observations made in the systematic review paper by Murithi et al. (2024).

Experimental Dataset

The models are vigorously tested using the Amazon mobile phone reviews dataset. This is a large-scale collection of user-generated product reviews from the Amazon e-commerce platform. It contains textual reviews along with additional metadata like products, reviews and their products. Key components of this dataset include text reviews, ratings, product metadata, reviewer metadata, votes, etc. (AlQahtani, 2021). Of the most importance is the fact that it contains instances where sentiments may shift due to negation or implicit contextual cues. The choice of the Amazon reviews dataset in the empirical analysis process is because it exhibits several aspects like diversity in language and style, difference in user demographics, varying lengths of text reviews, presence of polarity shift patterns as well as bias and noise differences. Besides these, its large size offers rich data for training a machine learning sentiment analysis model, focusing on implicit negations. This dataset is accessible from the Kaggle repository. Both the feature extraction techniques, data augmentation techniques and feature selection techniques are varied during the preprocessing of this dataset, in each experiment.

Development tools

Google Colab is used as Python's Integrated Development Environment (IDE) because of the flexibility that it offers as well as the computational resources. Several Python libraries are used to run the experiments. TensorFlow, Keras and PyTorch are the deep learning libraries used to implement and train the LSTM and the transformer-based BERT models. They provide efficient tools for building neural network architectures, handling large datasets, and training complex models (Erickson et al., 2017). Scikit-learn is used with the traditional machine learning tasks i.e. Support Vector Machine (SVM) and Naive Bayes. It provides implementations of various machine learning algorithms, as well as tools for pre-processing data, feature selection, and model evaluation (Kramer & Kramer, 2016). FastText library, developed by Facebook Research, provides efficient tools for text representation and classification.

NLTK (Natural Language Toolkit) is used across all the experiments to pre-process text data before training the models. NLTK is a popular library for natural language processing tasks in Python (Hardeniya et al., 2016). It provides tools for tokenization, stemming, part-of-speech tagging, and other text-processing tasks (Wang & Hu, 2021). Pandas and NumPy libraries are used for data manipulation and numerical computations.

Panda's libraries are particularly useful for handling structured data and data frames, while NumPy provides support for array operations and mathematical computations (Lemenkova, 2019). SciPy library is used for model evaluation. Matplotlib and Seaborn are used for visualization. Sea-born libraries are useful for presenting performance metrics, data distributions, and comparison visuals (Waskom, 2021).

Performance Metrics

To evaluate the performance of these models, three performance metrics are used, i.e. accuracy, Cohen's kappa and Mathew's correlation coefficient. The choice of these metrics in the evaluation process is in line with the systematic review works by Murithi et al. (2024). These scores are first normalized before computing their average score. Equation 1 and 2 below represents the normalization formula and the average performance score formula, respectively.

$$X_{normalized} = \frac{(X - X_{min})}{(X_{max} - X_{min})}$$

Equation 1

Performance = (AccuracyNormalized + Cohen's kappaNormalized +Mathew's CorrelationNormalized) /3 *Equation* 2

Data Analysis

The spreadsheet software is used to compute the normalization as well as the average performance scores while the R software is used to compute the ANOVA (Analysis of variance) analysis. ANOVA is a collection of statistical models and their associated estimation procedures used to analyze the differences among means. Three-way ANOVA analysis is conducted to analyze the experimental data and identify significant performance differences and interactions between model categories, feature extraction/selection techniques, and data augmentation methods. The results are interpreted to conclude the performance hierarchy of different model categories, the impact of feature extraction/selection techniques, and the effectiveness of data augmentation methods. Varying data augmentation techniques, feature extraction and selection techniques are used to identify differences in the performance of the different sentiment analysis models. The observation checklists as well as the experimental checklists are used as the data collection instruments in this research.

The observation checklists provide a mechanism through which the data from the various experiments are systematically recorded and documented to maintain consistency, objectivity, and thoroughness in the observations. This enhances the rigour, validity and reliability of the research study (Kapoor et al., 2023). On the other hand, the experimental checklists ensure that the experiments on the performance of the different categories of sentiment analysis models are conducted systematically, consistently, and with attention to important details (Freeman, 1999). They also help in effectively planning, executing, and documenting the results from the various experiments on the sentiment analysis models, in a variety of feature extraction techniques, feature selection techniques as well and data augmentation techniques. Key findings and insights are highlighted based on empirical observations and statistical analysis. Based on the observations and insights gained from the empirical analysis, recommendations are made for future research. This includes the proposal for developing a Hybrid model for Enhanced Polarity Shift Management of Mobile phone Reviews using Contextual Back Translation Augmented by Seq2seq Perturbations, aimed at further improving sentiment analysis accuracy and robustness.

RESULTS

The results of the data analysis from both the spreadsheets and the R software are presented in the subsequent sections.

Table 1. Three-way ANOVA of the sentiment analysis models, feature extraction techniques and feature selection techniques

ANOVA - Average Normalized Performance

	Sum of Squares	**df**	**Mean Square**	**F**	**p**
Model	4.03673	9	0.44853	43.743	< .001
Feature Extraction	0.10878	1	0.10878	10.609	0.002
Feature Selection	0.10878	1	0.10878	10.609	0.002
Model ✻ Feature Extraction	0.21528	9	0.02392	2.333	0.032
Model ✻ Feature Selection	0.21528	9	0.02392	2.333	0.032
Feature Extraction ✻ Feature Selection	0.00465	1	0.00465	0.454	0.504
Model ✻ Feature Extraction ✻ Feature Selection	0.02301	9	0.00256	0.249	0.984
Residuals	0.41015	40	0.01025		

Table 1 exhibits significant differences in the sentiment analysis models, feature extraction techniques as well as feature selection techniques, all at their levels. Moreover, there are significant interactions between model and feature extraction techniques as well as between models and feature selection techniques. Further post-hoc analysis demonstrates that the significant interactions between the models and the feature extraction/selection techniques only apply to the traditional machine learning models. These interactions do not apply to the deep learning models.

Table 2. Three-way ANOVA of the sentiment analysis models, feature Extraction techniques and data augmentation techniques

ANOVA - Average Normalized Performance

	Sum of Squares	df	Mean Square	F	P
Model	4.03673	9	0.44853	48.83906	< .001
Feature Extraction	0.10878	1	0.10878	11.84497	0.001
Data Augmentation	0.20503	1	0.20503	22.32544	< .001
Model ✻ Feature Extraction	0.21528	9	0.02392	2.60462	0.018

ANOVA - Average Normalized Performance

	Sum of Squares	df	Mean Square	F	P
Model ✻ Data Augmentation	0.18168	9	0.02019	2.19810	0.043
Feature Extraction ✻Data Augmentation	3.12e-5	1	3.12e-5	0.00340	0.954
Model ✻ Feature Extraction ✻ Data Augmentation	0.00778	9	8.65e-4	0.09414	1.000
Residuals	0.36735	40	0.00918		

Table 2 exhibits significant differences in the sentiment analysis models, feature extraction techniques as well as data augmentation techniques, all at their levels. Moreover, there are significant interactions between model and data augmentation techniques as well as between models and feature extraction techniques. Further post-hoc analyses demonstrate that the significant interactions between the models and the feature extraction techniques only apply to the traditional machine learning models. These interactions do not apply to the deep learning models. On the other hand, further post-hoc analysis demonstrates that the significant interactions between the sentiment analysis models and the data augmentation techniques only apply to the deep learning models. These interactions do not apply to the traditional machine learning models.

Table 3. Three-way ANOVA of the sentiment analysis models, feature selection techniques and data augmentation techniques

ANOVA - Performance Score

	Sum Squares	ofdf	Mean Square	F	p
Model	4.03673	9	0.44853	48.83906	< .001
Feature Selection	0.10878	1	0.10878	11.84497	0.001
Data Augmentation	0.20503	1	0.20503	22.32544	< .001
Model ✻ Feature Selection	0.21528	9	0.02392	2.60462	0.018
Model ✻ Data Augmentation	0.18168	9	0.02019	2.19810	0.043
Feature Selection ✻ Data Augmentation	3.12e-5	1	3.12e-5	0.00340	0.954
Model ✻ Feature Selection ✻ Data Augmentation	0.00778	9	8.65e-4	0.09414	1.000

Residuals	0.36735	40	0.00918

Table 3 exhibits significant differences in the sentiment analysis models, feature selection techniques as well as data augmentation techniques, all at their levels. Moreover, there are significant interactions between model and data augmentation techniques as well as between models and feature selection techniques. Further post-hoc analysis demonstrates that the significant interactions between the models and the feature selection techniques only apply to the traditional machine learning models. These interactions do not apply to the deep learning models. On the other hand, further post-hoc analysis demonstrates that the significant interactions between the sentiment analysis models and data augmentation techniques only apply to the deep learning models. These interactions do not apply to the traditional machine learning models.

Table 4. Three-way ANOVA of the feature extraction techniques, Feature Selection Techniques and the data augmentation techniques

ANOVA - Average Normalized Performance

	Sum of Squares	**df**	**Mean Square**	**F**	**P**
Feature Extraction	0.10878	1	0.10878	1.6686	0.201
Data Augmentation	0.20503	1	0.20503	3.1450	0.080
Feature Selection	0.10878	1	0.10878	1.6686	0.201
Feature Extraction ✻ Data Augmentation	3.13e-5	1	3.13e-5	4.79e-4	0.983
Feature Extraction ✻ Feature Selection	0.00465	1	0.00465	0.0713	0.790
Data Augmentation ✻ Feature Selection	3.13e-5	1	3.13e-5	4.79e-4	0.983
Feature Extraction ✻ Data Augmentation ✻ Feature Selection	0.00153	1	0.00153	0.0235	0.879
Residuals	4.69383	72	0.06519		

Table 4 exhibits a lack of significant differences in the feature extraction techniques, data augmentation techniques as well and the feature selection techniques, all at their levels. Moreover, there lack of significant interactions between feature extraction and data augmentation techniques, feature extraction and feature selection techniques as well as data augmentation and feature selection techniques. There are also no significant interactions among the three variables combined i.e. feature extraction techniques, data augmentation techniques and feature selection techniques. Based on this observation, further post-hoc analysis is not necessary.

Analysis of Different Categories of Sentiment Analysis Models and Their Average Normalized Performance Scores

In Figure 1, the acronyms for the different categories of models refer to the following: TML1 (first traditional machine learning model), TML2 (second traditional machine learning model), DL1 (first Deep learning model), DL2 (second Deep learning model), HTML1 (first hybrid made up of traditional machine learning models exclusively), HTML2 (second hybrid made up of traditional machine learning models exclusively), HDL1 (first hybrid made up of deep learning models exclusively), HDL2 (second hybrid made up of deep learning models exclusively), HTMDL1 (first hybrid made up of both the traditional and deep learning models) and lastly, HTMDL2 (second hybrid made up of both the traditional and deep learning models). This graph demonstrates that deep learning models and their respective hybrids outperform the traditional machine learning models and their hybrids.

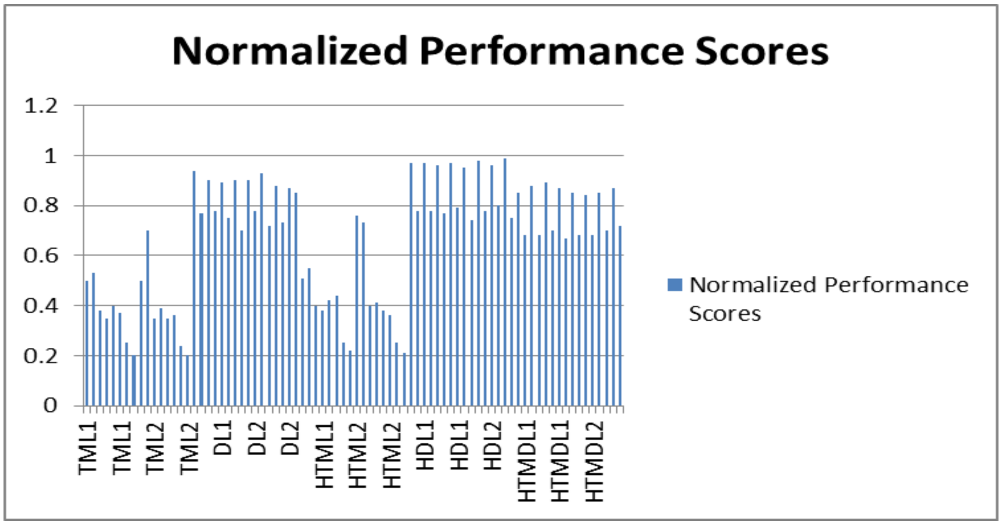

Figure 1. Bar graph of normalized performance scores of sentiment analysis models

Analysis of Different Categories of Feature Extraction Techniques and Their Average Normalized Performance Scores

In the below bar graph (Figure 2), the acronyms for the different categories of feature extraction techniques are as follows: FE1 (FastText) and FE2 (Word2Vec). This bar graph demonstrates that the FastText feature extraction technique outperforms the word2vec feature extraction technique in the performance of the traditional models.

Analysis of the Total Normalized Scores for the Different Feature Selection Techniques

In Figure 4, the acronyms for the different categories of feature selection techniques are as follows: FS1 (recursive feature selection) and FS2 (Term Frequency- Inverse Document Frequency). This bar graph demonstrates that the recursive feature selection (RFE) technique surpasses the performance of the TF-IDF (Term Frequency- Inverse Document Frequency) in the performance of the traditional machine learning-based sentiment analysis models.

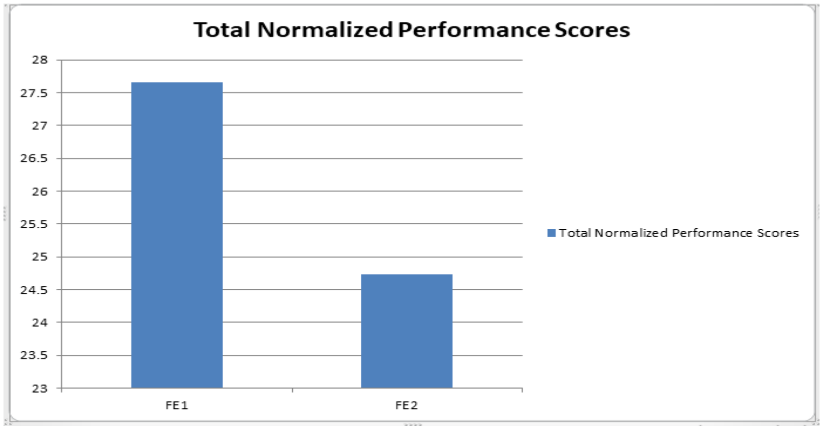

Figure 2. Bar graph of the total normalized average scores for feature extractions

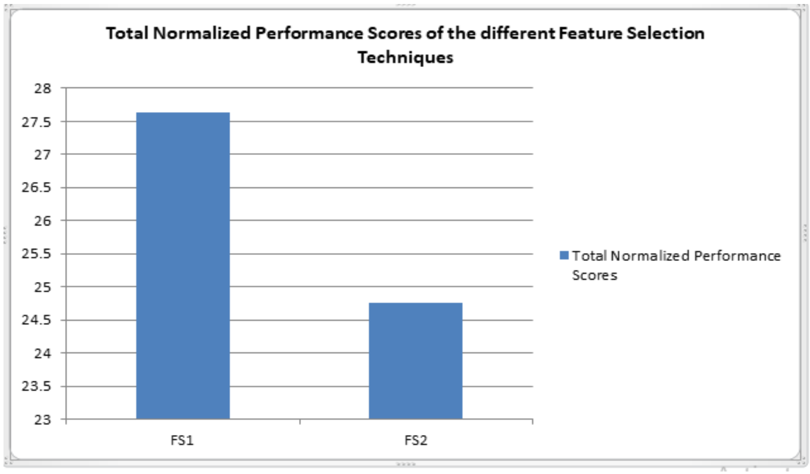

Figure 3. Bar graph of the total normalized average scores for the different feature selection techniques

Analysis of the Total Normalized Scores for the Different Data Augmentation Techniques

In the above bar graph (Figure 4), the acronyms for the different categories of data augmentation techniques are as follows: DA1 (Back translation) and DA2 (synonym replacement). This bar graph demonstrates that the Back translation data augmentation technique outperforms the synonym replacement technique in the performance of the deep learning-based sentiment analysis models.

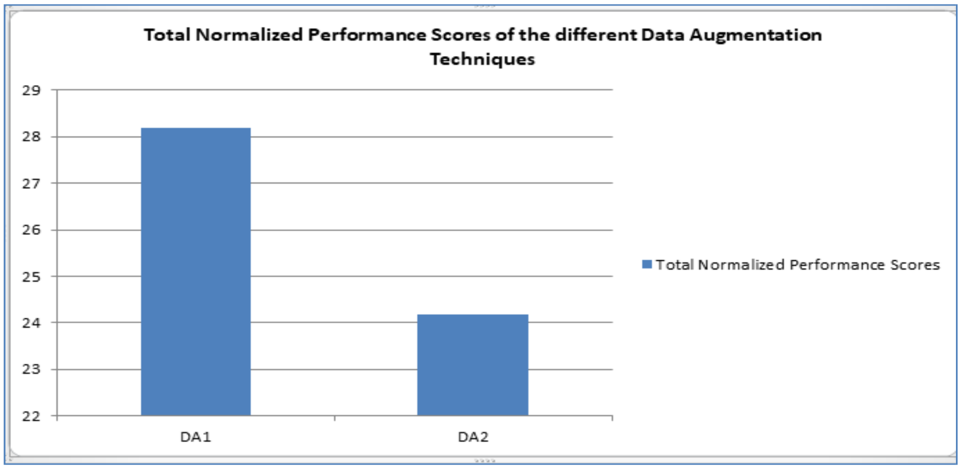

Figure 4. Bar graph of the total normalized average scores for the different feature selection

DISCUSSION

The experiments in this paper embark on a comprehensive empirical exploration of sentiment flux within the context of state-of-the-art models, specifically engineered to handle polarity shifts induced by implicit negation. Implicit negation, a linguistic phenomenon where negation is subtly conveyed without explicit negation terms, poses a unique challenge to sentiment analysis models. Understanding the complex landscape of sentiment flux under such conditions is imperative for developing robust and accurate sentiment analysis models. This empirical research delves into the experimentation and performance evaluation of diverse sentiment analysis models, broadly categorized into five groups: traditional machine learning models, deep learning models, hybrid models integrating both machine learning and deep learning, hybrid models predominantly utilizing deep learning methodologies, and hybrid models exclusively employing machine learning approaches. The empirical analysis is geared towards unravelling the impact of various feature extraction, feature selection, and data augmentation techniques on the models' efficacy in handling polarity shifts induced by implicit negation. Through rigorous experimentation, advanced ANOVA statistical analyses are employed using the R software to extract key findings and insights.

The findings reveal significant performance disparities between traditional machine learning-based models and their deep learning counterparts. Notably, interactions between traditional machine learning models and feature extraction techniques underscore the pivotal role played by these techniques in shaping model performance. In contrast, the analyses suggest that most feature extraction in deep learning models occurs within the deep network itself, minimizing the influence of external feature extraction techniques. Further exploration uncovers meaningful interactions between traditional machine learning models and feature selection techniques, while no such interactions are observed within the realm of deep learning models. Additionally, this investigation reveals the absence of significant interactions between traditional machine learning models and data augmentation techniques, signalling potential limitations in leveraging certain augmentation strategies for this category. Conversely, deep learning models exhibit substantial interactions with data augmentation techniques, with the back translation method outperforming synonym replacement in augmenting data and enhancing model performance. These empirical observations provide a foundational understanding and insights that pave the way for the future development of an advanced sentiment analysis model. We envision a transformer-based BERT-based deep learning model that strategically configures hyperparameters for effective feature extraction within the deep network.

Coupled with a back translation data augmentation hybrid, incorporating perturbations tailored for implicit negation, such a model holds promise for improved accuracy and adaptability in handling nuanced shifts in sentiment induced by implicit negation. The following provides detailed practical implications for the observations made from the experimental data analysis.

Significant differences between traditional ML and deep learning models

The ANOVA analysis reveals notable performance distinctions between traditional machine learning (ML) models and deep learning (DL) models. This suggests that the architectures and methodologies employed in deep learning models are particularly effective in handling polarity shifts due to implicit negation compared to traditional machine learning approaches. Deep learning models inherently capture more complex patterns in sentiment analysis tasks, which are crucial for handling nuanced sentiment flux. These observations are in line with the ones made by Pathak et al. (2020).

Interactions between traditional ML models and feature extraction techniques

Significant interactions between traditional ML models and feature extraction techniques indicate that the choice of feature extraction methods significantly impacts the performance of these models. This observation suggests that traditional ML models heavily rely on feature engineering to extract relevant information from textual data for sentiment analysis, highlighting the importance of tailored feature extraction techniques in traditional ML approaches. This observations in line with the one made by Symeonidis et al. (2018)

Lack of interactions between DL models and feature extraction techniques

Conversely, the absence of significant interactions between deep learning models and feature extraction techniques implies that most of the feature extraction process occurs within the deep network architecture itself. Deep learning models are capable of automatically learning hierarchical representations of text, eliminating the need for explicit feature engineering. This finding underscores the self-sufficiency of deep learning models in extracting meaningful features from raw data. These observations are in line with the ones made by Leng and Jiang (2016).

Interactions between traditional ML models and feature selection techniques

Meaningful interactions between traditional ML models and feature selection techniques suggest that feature selection plays a crucial role in enhancing the performance of these models. Traditional ML models benefit from selecting relevant features to improve their discriminative power in sentiment analysis tasks, indicating the importance of feature selection strategies in optimizing model performance. These observations are in line with the ones made by Cai et al. (2018).

Lack of interactions between DL models and feature selection techniques

Conversely, the absence of significant interactions between deep learning models and feature selection techniques implies that deep learning models are less sensitive to feature selection processes. This finding suggests that deep learning models inherently learn to focus on informative features during training, reducing the need for explicit feature selection techniques. These observations are in line with the ones made by Zou et al. (2015).

Lack of interactions between traditional ML models and data augmentation techniques

The absence of significant interactions between traditional ML models and data augmentationtechniques indicates that data augmentationmethods may not significantly impact the performance of traditional ML models in handling polarity shifts due to implicit negation. This suggests that traditional ML models may not benefit as much from data augmentation techniques compared to deep learning models. These observations are in line with the ones made by Sakai et al. (2017).

Interactions between DL models and data augmentation techniques

Significant interactions between deep learning models and data augmentation techniques highlight the importance of data augmentation in enhancing the performance of deep learning models for sentiment analysis. Specifically, the superior performance of the back translation method compared to synonym replacement suggests that generating diverse and semantically meaningful augmented data through back translation contributes significantly to improving the robustness of deep learning models in capturing polarity shifts. These observations are in line with the ones made by Lee (2021).

The superior performance of Fast Text Feature Extraction Technique as opposed to Word2Vec

The observation of superior performance of the FastText as opposed to the Word2vec is in line with the one made by Hasan et al. (2020). This is attributed to the fact that FastText considers subword information (character n-grams) when generating word embeddings. This means that even for words that are not present in the training corpus or are out-of-vocabulary (OOV), FastText can still create meaningful embeddings based on their constituent character sequences. In sentiment analysis, words affected by implicit negation usually have variations due to prefixes, suffixes, or internal character changes. FastText's ability to capture subword information helps in understanding such morphological variations, leading to more robust representations of words affected by implicit negation cues. The inclusion of subword information enhances the model's generalization capabilities. FastText embeddings can generalize better to unseen words or variations of words not encountered during training. Implicit negation introduces new linguistic patterns and variations that were not explicitly seen in the training data.

FastText's ability to generalize across subword representations helps in adapting to these variations and maintaining robustness in sentiment analysis predictions. FastText embeddings provide flexibility in capturing contextually similar words based on their subword similarities. This is useful in capturing sentiment-related prefixes, suffixes, or morphological variations affected by implicit negation cues. Implicit negation often involves contextually shifting sentiment orientations within sentences. FastText's contextual flexibility at the subword level can aid in capturing these subtle changes and improving the model's sensitivity to implicit negation cues. Overall, FastText's strength lies in its ability to capture morphological variations, handle rare or domain-specific terms, and provide contextual flexibility through subword embeddings. These qualities contribute to the generation of more robust features for sentiment analysis models dealing with polarity shift and implicit negation in datasets like Amazon mobile reviews.

The superior performance of the wrapper-based recursive feature selection Technique as opposed to TFIDF

The observation of superior performance of the wrapper-based recursive feature selection (RFE) technique as opposed to the TF-IDF is in line with the observation made by Onan (2016). This is attributed to the RFE's customization, model performance optimization, interpretability, and alignment with domain-specific nuances, making it well-suited for sentiment analysis tasks focused on handling polarity shifts in such datasets. Amazon reviews dataset contains nuanced language, domain-specific terms, and implicit sentiment expressions influenced by negation. These complexities require a feature selection method that can capture subtle linguistic nuances and optimize model performance accordingly. RFE allows for customized feature subset selection based on the iterative evaluation of feature importance. This customization is beneficial for capturing features related to polarity shifts, implicit negation cues, and context-dependent sentiment expressions present in Amazon reviews. RFE is designed to optimize model performance by selecting feature subsets that contribute most effectively to the predictive power of the model. In sentiment analysis tasks where polarity shift detection is crucial, optimizing model performance can lead to more accurate sentiment predictions. RFE provides insights into which features are most relevant for the sentiment analysis model. This enhances interpretability and understanding of the linguistic cues and patterns influencing polarity shifts and implicit negation in Amazon reviews. While RFE may require some domain-specific knowledge for fine-tuning, it leverages this knowledge to select features that align with the specific nuances of sentiment analysis in the Amazon reviews domain. This targeted approach leads to better feature selection outcomes.

The superior performance of the Back Translation data Augmentation Technique as opposed to the Synonym replacement

The observation of the higher performance of the Back translation as compared to the Synonym replacement is in line with the one made by Beddiar et al. (2021). This is attributed to the fact that back translation introduces significant semantic variations by translating sentences into another language and then back to the original. This variation helps deep learning models learn diverse linguistic patterns and capture implicit nuances in sentiment. Deep learning models, particularly those based on transformers like BERT, rely heavily on contextual information. Back translation augments the data with diverse contexts, which can enhance the model's ability to understand and generalize well across different sentence structures and sentiment expressions. Amazon's mobile reviews dataset often contains a wide range of language styles, sentiments, and implicit expressions. Back translation simulates different ways that the customers express their opinions, helping the model learn to handle polarity shifts due to implicit negations. While synonym replacement is beneficial, especially in scenarios where specific word choices impact sentiment, it does not capture the same level of semantic and contextual variations as back translation. Amazon reviews often contain diverse language patterns beyond simple word replacements. Considering the complexity and diversity of language in Amazon reviews, back translation tends to be a better choice for deep learning models. It offers richer semantic variations and contextual diversity, aligning well with the challenges posed by implicit negations and nuanced sentiment expressions commonly found in such datasets. By leveraging back translation, the sentiment analysis model can learn from a more diverse set of examples, improving its ability to handle polarity shifts effectively and generalize well to unseen data scenarios within the Amazon mobile reviews domain.

While back translation is a valuable data augmentation technique for handling polarity shifts in sentiment analysis, especially in datasets like Amazon mobile reviews, it does come with its own set of challenges. For example, Back translation may not always preserve the specific context, or idiomatic expressions present in the original language. This can lead to the generation of sentences that are grammatically correct but lose some nuanced meanings or cultural nuances. Developing context-aware back translation approaches that consider the context of the original text and try to preserve idiomatic expressions or domain-specific language can help mitigate this challenge. The quality of translations can vary depending on the translation model and language pair used. Poor translations can introduce noise or incorrect sentiment cues into the augmented data, affecting model performance. Using high-quality translation models or fine-tuning translation models specifically for sentiment-related tasks can improve the quality of back-translated data. Leveraging multiple translation models and incorporating quality checks can also enhance the reliability of the augmentation process. Back translation relies on diverse translations to introduce variability in the data. However, if the translated data is not diverse enough or if certain linguistic patterns are not adequately represented, it can lead to biased or skewed augmentation. Incorporating multiple translation models or language resources can improve data diversity. Additionally, post-augmentation data analysis and bias detection techniques can help identify and mitigate biases introduced through back translation. Back translation may not always capture domain-specific language or sentiments accurately, especially in specialized domains like product reviews. Adapting translations to domain-specific vocabulary and sentiment expressions is crucial. Fine-tuning translation models on domain-specific data or incorporating domain-specific dictionaries and lexicons during back translation can improve relevance and accuracy.

Overall, while back translation is a powerful augmentation technique, addressing these challenges through advancements in translation models, context-aware approaches, quality checks, and domain adaptation strategies can further enhance its effectiveness in handling polarity shifts and improving sentiment analysis model performance on diverse datasets like Amazon mobile reviews. Among the challenges mentioned, one major challenge in handling datasets used by the sentiment analysis models used to handle polarity shift due to implicit negation through back translation is preserving specific context and idiomatic expressions.

For example, the Amazon reviews dataset often contains colloquial language, domain-specific terms, and nuanced sentiments that are crucial for accurate sentiment analysis and polarity shift management. However, back translation may not always capture these aspects effectively, leading to a loss of context and potentially affecting model performance. Improvement through perturbations can help address this challenge by introducing controlled variations in the back-translated data while preserving important contextual and idiomatic elements. Perturbations involve making small modifications or additions to the data to create diverse yet contextually relevant examples. Improvement through perturbations can be beneficial in several ways: Perturbations can be designed to preserve specific context and idiomatic expressions during back translation, for example, idiom retention, domain-specific terms and cultural sensitivity. Identifying common idiomatic expressions in the original language and ensuring that they are retained or translated appropriately during back translation is critical. Incorporating domain-specific dictionaries or lexicons to guide back translation and ensure that important terms related to mobile devices or product features are accurately translated is also critical. Lastly, considering cultural nuances and sensitivities to avoid mistranslations or misinterpretations that may arise due to cultural differences is also important. Perturbations can introduce semantic variability in the back-translated data without compromising on context. This helps in creating a diverse training set for deep learning models, enhancing their ability to handle polarity shifts and nuanced sentiment analysis. Implementing quality control measures within the perturbation process to validate the effectiveness of back translation and ensure that perturbed examples maintain their relevance and authenticity is important. Combining perturbations with the back translation data augmentation strategy to further enrich the training data with diverse linguistic patterns and sentiment expressions could be also beneficial.

By incorporating perturbations into the back translation process, researchers and practitioners can tailor the augmentation process to the specific challenges of Amazon reviews or similar datasets, thereby improving the quality and relevance of augmented data for sentiment analysis and polarity shift management tasks. Enhancing Sentiment Analysis through Contextual Back Translation Augmented by Seq2seq Perturbations could be worth investigating during future research. Combining Seq2seq models with perturbation techniques during back translation helps in creating more diverse and contextually relevant augmented datasets for training NLP models, enhancing their performance in tasks like sentiment analysis on datasets such as Amazon reviews.
These improvements can lead to more robust and accurate sentiment analysis models for polarity shift management due to implicit negation, especially in domains with complex language structures and nuanced sentiments. Several practical implications arise: The significant differences in performance between traditional machine learning (ML) models and deep learning (DL) models suggest that DL models are particularly effective in handling polarity shifts due to implicit negation.

This implies that the inherent capacity of DL models to capture complex patterns and hierarchical representations in textual data makes them well-suited for sentiment analysis tasks that involve nuanced sentiment flux. The significant interactions observed between traditional ML models and feature extraction techniques highlight the reliance of traditional ML models on feature engineering. This underscores the importance of tailored feature extraction techniques in traditional ML approaches to extract relevant information from text data effectively. The lack of significant interactions between DL models and feature extraction techniques suggests that most of the feature extraction process occurs within the deep network architecture itself. Deep learning models can automatically learn hierarchical representations of text, reducing the need for explicit feature engineering and indicating the self-sufficiency of DL models in extracting meaningful features from raw data. Meaningful interactions between traditional ML models and feature selection techniques emphasize the importance of feature selection in enhancing the performance of these models. Feature selection strategies can improve the discriminative power of traditional ML models in sentiment analysis tasks, highlighting the significance of feature selection techniques in optimizing model performance. The absence of significant interactions between DL models and feature selection techniques suggests that DL models are less sensitive to feature selection processes. This implies that DL models inherently learn to focus on informative features during training, reducing the need for explicit feature selection techniques compared to traditional ML models. The significant interactions between DL models and data augmentation techniques underscore the importance of data augmentation in enhancing the performance of DL models for sentiment analysis. The superior performance of the back translation method compared to synonym replacement suggests that generating diverse and semantically meaningful augmented data contributes significantly to improving the robustness of DL models in capturing polarity shifts. Overall, these practical implications provide valuable insights, guiding the future development of improved sentiment analysis models for handling polarity shifts due to implicit negation.

CONCLUSIONS AND RECOMMENDATIONS

In conclusion, our comprehensive empirical analysis has provided valuable insights into the complex nuanced landscape of sentiment flux within the realm of state-of-the-art models designed for handling polarity shifts due to implicit negation, within the context of mobile phones reviews. The key findings stemming from the ANOVA analysis highlight significant differences in performance between traditional machine learning-based models and their deep learning counterparts, underlining the importance of model selection in addressing the intricacies of sentiment analysis.

Specifically, the findings exhibit the superiority of deep learning models, particularly hybrids, in handling polarity shifts and implicit negation compared to traditional machine learning (ML) models. This suggests that incorporating advanced neural network architectures can significantly enhance sentiment analysis accuracy and robustness. The observed interactions between traditional machine learning models and feature extraction techniques emphasize the need for tailored approaches in enhancing feature representation. Conversely, the lack of significant interactions between deep learning models and feature extraction techniques suggests that feature extraction is predominantly occurring within the deep network itself, urging researchers to delve deeper into the inner workings of these models. Notably, our identification of meaningful interactions between traditional machine learning models and feature selection techniques provides avenues for refining model interpretability and generalization. The absence of significant interactions in the deep learning models, on the other hand, prompts further exploration into optimizing feature extraction and selection strategies specific to these models. The highlighted superiority of the back translation method over synonym replacement in data augmentation for deep learning models signals a promising avenue for improving model robustness. These observations lay the groundwork for the future development of an enhanced sentiment analysis model tailored to handle polarity shifts due to implicit negation.

Based on these findings, it is recommended to focus on developing and deploying hybrid models that leverage deep learning methodologies, such as BERT and LSTM, combined with effective data augmentation techniques like back translation. Such an approach aligns with the identified strengths of deep learning models and their interactions with data augmentation methods, which have been shown to improve sentiment analysis performance, and would offer a promising avenue to further enhance model performance and generalization. Furthermore, besides the use of accuracy, Cohen's kappa and Mathew's correlation in the evaluation of models in this empirical paper, future evaluation of the proposed novel algorithm should also be based on Kruskal Wallis H statistic besides visualizations like confusion matrix in the form of a table ROC curve and AUC, precision, Recall curve, word clouds, and attention maps. This will provide a comprehensive evaluation of its novelty. Lastly, the validation process of the novel model should incorporate additional datasets, besides Amazon Reviews, to test its generalizability.

IMPLICATIONS

The implications of this research are substantial for the field of sentiment analysis and natural language processing (NLP) in the domain of mobile phone reviews. The observed performance differences between traditional ML and deep learning models underscore the need for tailored approaches and the adoption of state-of-the-art architectures for effective sentiment analysis tasks.

Additionally, the significant interactions identified between traditional ML models and feature extraction/selection techniques highlight the importance of these pre-processing steps in capturing nuanced sentiment nuances present in mobile phone reviews. This implies that careful consideration and optimization of feature extraction and selection methods can lead to improved sentiment analysis outcomes. Moreover, the findings regarding data augmentation techniques, particularly the superiority of back translation over synonym replacement, emphasize the critical role of contextually rich data augmentation methods in enhancing sentiment analysis accuracy and model performance. Further improvement on the performance of the back translation offers a good leverage point in the development of more effective models in the future.

Overall, these implications provide valuable guidance for researchers and practitioners aiming to develop improved sentiment analysis models capable of handling polarity shifts, implicit negation, and nuanced sentiment expressions effectively in diverse review datasets like those from Amazon mobile phone reviews.

The theoretical implications of our findings extend beyond model development, offering a deeper understanding of the interplay between different techniques in the context of sentiment analysis in the face of implicit negation. Practical applications of this research include the refinement of sentiment analysis systems across diverse domains, where implicit negation plays a pivotal role in shaping the sentiment landscape. Future research endeavours should continue to investigate the evolving challenges and opportunities in sentiment analysis, exploring possibilities for the development of novel methodologies and techniques to advance the field and address the complexities introduced by implicit negation.

One such possibility is the recommendation of a hybrid model leveraging Contextual Back Translation Augmented by Seq2seq Perturbations. This suggests a promising avenue for future research. This approach aims to address the identified challenges by integrating advanced techniques to improve sentiment analysis accuracy, robustness, and contextual understanding in mobile phone reviews. Overall, these implications provide valuable insights into developing more effective sentiment analysis models capable of handling nuanced sentiment expressions, polarity shifts, and implicit negation in diverse review datasets such as those from Amazon mobile phone reviews.

ACKNOWLEDGEMENT

The researcher is indebted to her Ph.D. supervisors, the entire faculty members and the entire family for their financial and moral support.

FUNDING

The study did not receive funding from any institution.

DECLARATIONS

Conflict of Interest

The author declared that there is no conflict of interest.

Informed Consent

Due to the nature of this study, which involves the analysis of existing public data from Amazon phone reviews, individual informed consent from participants is not required. The data used in this study are publicly available and do not contain personally identifiable information. Therefore, there are no direct interactions with human subjects, and informed consent is not applicable in this context.

Ethics Approval

This is not applicable because ethical considerations related to human subjects, confidentiality, and privacy do not apply in this study. This study utilizes publicly available data and does not involve human subjects or interventions. For this reason, it does not require ethics approval from a research ethics committee. The research focuses on analysing data and evaluating machine learning models based on established methodologies and practices in the field of natural language processing (NLP) and sentiment analysis.

REFERENCES

Abdi, A., Shamsuddin, S. M., Hasan, S., & Piran, J. (2019). Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion. *Information Processing & Management*, 56(4), 1245–1259. https://doi.org/10.1016/j.ipm.2019.02.018

Abirami, A. M., & Gayathri, V. (2017, January). A survey on sentiment analysis methods and approach. In *2016 Eighth International Conference on Advanced Computing (ICoAC)* (pp. 72-76). IEEE.

Ahmad, M., & Aftab, S. (2017). Analyzing the Performance of SVM for Polarity Detection with Different Datasets. *International Journal of Modern Education and Computer Science*, 9(10), 29–36. https://doi.org/10.5815/ijmecs.2017.10.04

AlQahtani, A. S. (2021). Product sentiment analysis for Amazon reviews. *International Journal of Computer Science & Information Technology (IJCSIT)* Vol, 13.

Astya, P. (2017, May). Sentiment analysis: approaches and open issues. In *2017 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 154-158). IEEE.

Avinash, M., & Sivasankar, E. (2019). A Study of Feature Extraction Techniques for Sentiment Analysis. In A. Abraham, P. Dutta, J. K. Mandal, A. Bhattacharya, & S. Dutta (Eds.), *Emerging Technologies in Data Mining and Information Security* (Vol. 814, pp. 475–486). Springer Singapore.

Beddiar, D. R., Jahan, M. S., & Oussalah, M. (2021). Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24, 100153.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300, 70-79.

Deho, B. O., Agangiba, A. W., Aryeh, L. F., & Ansah, A. J. (2018, August). Sentiment analysis with word embedding. In *2018 IEEE 7th International Conference on Adaptive Science & Technology (ICAST)* (pp. 1-4). IEEE.

Duong, H. T., & Nguyen-Thi, T. A. (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, 8(1), 1.

Erickson, B. J., Korfiatis, P., Akkus, Z., Kline, T., & Philbrick, K. (2017). Toolkits and libraries for deep learning. *Journal of digital imaging*, 30, 400-405.

D’souza, S. R., & Sonawane, K. (2019, March). Sentiment analysis based on multiple reviews by using machine learning approaches. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 188-193). IEEE.

Eke, C. I., Norman, A. A., & Shuib, L. (2021). Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and BERT model. *IEEE Access*, 9, 48501-48518.

Freeman, J. (1999). Teaching gifted pupils. *Journal of Biological Education*, 33(4), 185-190.

Hardeniya, N., Perkins, J., Chopra, D., Joshi, N., & Mathur, I. (2016). Natural language processing: python and NLTK. *Packt Publishing Ltd.*

Hasan, H., Qorina, E. S., Hulliyah, K., Zamhari, A., & Saepudin, D. (2020,October). Comparative Analysis of the Performance of the Fasttext and Word2vec Methods on the Semantic Similarity Query of Sirah Nabawiyah Information Retrieval System: A systematic literature review. The *8th International Conference on Cyber and IT Service Management (CITSM 2020) On Virtual*, 23-24.

Hou, Y., Liu, Y., Che, W., & Liu, T. (2018). Sequence-to-sequence data augmentation for dialogue language understanding. *arXiv preprint*. arXiv:1807.01554.

Howell, K. E. (2012). An introduction to the philosophy of methodology. *An Introduction to the Philosophy of Methodology*, 1-248.

Israel, M. (2011). The grammar of polarity: Pragmatics, sensitivity, and the logic of scales (Vol. 127). *Cambridge University Press*.

Japhne, A., & Murugeswari, R. (2020, February). Opinion mining-based complex polarity shift pattern handling for improved sentiment classification. In *2020 International Conference on Inventive Computation Technologies (ICICT) (pp. 323-329)*. IEEE.

Kapoor, S., Cantrell, E., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., & Narayanan, A. (2023). Reforms: Reporting standards for machine learning-based science. *arXiv preprint*. arXiv:2308.07832.

Khan, M. T., Durrani, M., Ali, A., Inayat, I., Khalid, S., & Khan, K. H. (2016). Sentiment analysis and the complex natural language. *Complex Adaptive Systems Modeling*, 4, 1-19.

Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint* arXiv:1805.06201.

Kramer, O., & Kramer, O. (2016). Scikit-learn. *Machine learning for evolution strategies*, 45-53.

Kumar, A., & Garg, G. (2023). Empirical study of shallow and deep learning models for sarcasm detection using context in benchmark datasets. *Journal of Ambient Intelligence and Humanized Computing*, 14(5), 5327–5342.

Kumar, A., & Garg, G. (2020). Systematic literature review on context-based sentiment analysis in social multimedia. *Multimedia Tools and Applications*, 79(21), 15349-15380.

Kumar, S. S., & Rajini, A. (2019, July). Extensive survey on feature extraction and feature selection techniques for sentiment classification in social media. *In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

Lee, M. A. (2021). *Examining Machine Translation Systems and Translation Quality using the Back-Translation Method* (Doctoral dissertation, University of Wisconsin--Stout).

Lemenkova, P. (2019). Processing oceanographic data by Python libraries NumPy, SciPy and Pandas. *Aquatic Research*, 2(2), 73-91.

Leng, J., & Jiang, P. (2016). A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm. *Knowledge-Based Systems*, 100, 188-199.

Liu, B. (2020). Text sentiment analysis based on CBOW model and deep learning in big data environment. *Journal of Ambient Intelligence and Humanized Computing*, 11(2), 451–458. https://doi.org/10.1007/s12652-018-1095-6

Long, F., Zhou, K., & Ou, W. (2019). Sentiment analysis of text based on bidirectional LSTM with multi-head attention. *IEEE Access*, 7, 141960-141969.

Marrapu, S., Senn, W., & Prybutok, V. (2024). Sentiment Analysis of Twitter Discourse on Omicron Vaccination in the USA Using VADER and BERT. *Journal of Data Science and Intelligent Systems.*

Mhatre, M., Phondekar, D., Kadam, P., Chawathe, A., & Ghag, K. (2017, July). Dimensionality reduction for sentiment analysis using pre-processing techniques. In *2017 International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 16-21). IEEE.

Miao, J., & Niu, L. (2016). A Survey on Feature Selection. *Procedia Computer Science*, 91, 919-926.

https://doi.org/10.1016/j.procs.2016.07.111

Mohd Nafis, N. S., & Awang, S. (2021). An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification. *IEEE Access*, 9, 52177–52192. https://doi.org/10.1109/ACCESS.2021.3069001

Moon, S., Kim, M. Y., & Iacobucci, D. (2021). Content analysis of fake consumer reviews by survey-based text categorization. *International Journal of Research in Marketing*, 38(2), 343-364.

Murithi, M. K., Oirere, A. M., & Ndung'u, R. N. (2024). A Systematic Review of the Sentiment Analysis Models Used in Handling Polarity Shift. *International Journal of Computing Sciences Research*, 8, 2635-2676.

Onan, A. (2016). Classifier and feature set ensembles for web page classification. *Journal of Information Science*, 42(2), 150-165.

Pathak, A. R., Agarwal, B., Pandey, M., & Rautaray, S. (2020). Application of deep learning approaches for sentiment analysis. *Deep learning-based approaches for sentiment analysis*, 1-31.

Pongthanoo, P., & Songpan, W. (2020, May). Feature selection and reduction based on SMOTE and information gain for sentiment mining. In *2020 5th International Conference on Computer and Communication Systems (ICCCS)* (pp. 109-114). IEEE..

Prastyo, P. H., Ardiyanto, I., & Hidayat, R. (2020, September). A Review of Feature Selection Techniques in Sentiment Analysis Using Filter, Wrapper, or Hybrid Methods. In *2020 6th International Conference on Science and Technology (ICST)* (Vol. 1, pp. 1-6). IEEE.

Rajabi, Z., Valavi, M. R., & Hourali, M. (2020). A Context-Based Disambiguation Model for Sentiment Concepts Using a Bag-of-Concepts Approach. *Cognitive Computation*, 12(6), 1299–1312. https://doi.org/10.1007/s12559-020-09729-1

Sakai, A., Minoda, Y., & Morikawa, K. (2017, August). Data augmentation methods for machine-learning-based classification of bio-signals. In *2017 10th Biomedical Engineering International Conference (BMEiCON) (pp. 1-4)*. IEEE.

Salur, M. U., & Aydin, I. (2020). A novel hybrid deep learning model for sentiment classification. *IEEE Access*, 8, 58080-58093.

Singh, P. K., & Paul, S. (2021). Deep Learning Approach for Negation Handling in Sentiment Analysis. *IEEE Access*, 9, 102579–102592. https://doi.org/10.1109/ACCESS.2021.3095412

Sugiyama, A., & Yoshinaga, N. (2019, November). Data augmentation using back-translation for context-aware neural machine translation. *In Proceedings of the fourth workshop on discourse in machine translation (DiscoMT 2019)* (pp. 35-44).

Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for Twitter sentiment analysis. *Expert Systems with Applications*, 110, 298-310.

Tang, D., Wei, F., Qin, B., Yang, N., Liu, T., & Zhou, M. (2015). Sentiment embeddings with applications to sentiment analysis. *IEEE transactions on knowledge and data Engineering*, 28(2), 496-509.

Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N., & Fox, E. A. (2020). Natural language processing advancements by deep learning: A survey. *arXiv preprint* arXiv:2003.01200.

Van de Kauter, M., Desmet, B., & Hoste, V. (2015). The good, the bad and the implicit: a comprehensive approach to annotating explicit and implicit sentiment. *Language resources and evaluation*, 49, 685-720.

Wang, M., & Hu, F. (2021). The application of nltk library for Python natural language processing in corpus research. *Theory and Practice in Language Studies*, 11(9), 1041-1049.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review,* 55(7), 5731-5780.

Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.

Xia, R., Xu, F., Yu, J., Qi, Y., & Cambria, E. (2016). Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Information Processing & Management*, 52(1), 36-45.

Xiang, Z., Schwartz, Z., Gerdes Jr, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction?. *International journal of hospitality management*, 44, 120-130.

Zou, Q., Ni, L., Zhang, T., & Wang, Q. (2015). Deep learning-based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 12(11), 2321-2325.

Authors' Biography

Millicent Kathambi Murithi is a Tutorial Fellow at the Department of Computer Science, Murang'a University of Technology, Kenya. She holds an MSc. Degree in Computer Systems from Jomo Kenyatta University of Science and Technology, Kenya. Her research interests include Machine Learning, software engineering and Natural Language Processing.

Aaron Mogeni Oirere is a Lecturer at the Department of Computer Science, Murang'a University of Technology, Kenya. He holds a Ph.D. Degree in Computer Science from Dr. Babasaheb Ambedkar Marathwada University, Maharashtra, India. His research interests include Automatic Speech Recognition, Human-computer Interaction, Information Retrieval, Database Management Systems (DBMS), Data Analytics and Hardware & Networking.

Rachael Njeri Ndung'u is a Lecturer at the Department of Information Technology, Murang'a University of Technology, Kenya. She holds a Ph.D. Degree in Information Technology. Her research interests include Artificial Intelligence, Data analytics and Blockchain.