IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 July 2024, accepted 11 September 2024, date of publication 20 September 2024, date of current version 11 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3464374

APPLIED RESEARCH

ChatGPT as a Text Annotation Tool to Evaluate Sentiment Analysis on South African Financial Institutions

MIEHLEKETO MATHEBULA, (Member, IEEE), ABIODUN MODUPE, (Member, IEEE), AND VUKOSI MARIVATE, (Senior Member, IEEE)
Department of Computer Science, University of Pretoria, Pretoria 0002, South Africa
Corresponding author: Miehleketo Mathebula (Miehleketo93@gmail.com)

ABSTRACT Social media platforms play a significant role in analyzing customer perceptions of financial products and services in today's culture. These platforms facilitate the immediate and in-depth sharing of thoughts and experiences, offering valuable insights into consumer behaviour. Any customer looking for such a service would surf the internet for reviews and ratings before making a decision, which usually influences their ultimate pick. Feedback and suggestions from friends, family, and coworkers improve customer experiences. Customer reviews play a crucial role in shaping the reputation and profitability of businesses and products offered by financial institutions, often serving as the final assessment of quality and satisfaction during decision-making. Therefore, it is paramount for decision-makers to carefully evaluate customer feedback and understand the sentiment expressed in a given piece of text, which could lead to equity trading, and credit market assessment, and offer invaluable insights that boost the financial performance of the institution. Previous research has used human-annotated text, such as lexicon-based methods, to train machine learning models for sentiment analysis, but the approach did not capture the full range of structure and semantic relationships in natural language. Therefore, our research aims to develop a more comprehensive and accurate sentiment analysis model using advanced natural language processing techniques that could answer questions on various subjects and tasks. To do this, we first crawled customer reviews on Hellopeter, a popular review site, and financial data on the top five financial institutions listed on the Johannesburg Stock Exchange (JSE) in South Africa. After that, we used OpenAI's ChatGPT as a zero-short learning model to generate human-like annotation tools for different sentiment tasks. The OpenAI ChatGPT feature vector was subsequently fed into BERT, BiLSTM, and a SoftMax function to detect and identify the sentiment of a given sentence. Lastly, we use feature vectors with oversampling methods to address the imbalanced data dilemma and visualise the contribution features of the given piece of text for the customer reviewers. The experiments demonstrated that the method performed as well as or better than the latest and most effective methods on the tested datasets, yielding comparable results. When OpenAI's ChatGPT was combined with pre-trained BERT and BiLSTM models, it did better overall, with an average score of 98.9%, an F1-measure of 97.7%, and an AUC of 91.90% when oversampling was used. The traditional lexicon-based model got an 86.68% score using SVM and logistic regression and an AUC of 91.90%. The study shows the exceptional performance of OpenAI ChatGPT in detecting the emotional tone or polarity of a given sentence in a customer review, which helps with annotation and understanding the sentiment analysis of an event and how it influences decisions and outcomes. In conclusion, these results underscore the significant advantages of incorporating customer sentiment analysis into financial analysis and decision-making processes as a valuable tool for understanding and prioritizing customer needs and preferences.

The associate editor coordinating the review of this manuscript and approving it for publication was Ikramullah Lali.

INDEX TERMS Sentiment analysis, Hellopeter, online media, SMOTE, OpenAI ChatGPT, BiLSTM, BERT, NLP.

I. INTRODUCTION

The growing usage of social media has resulted in a world saturated with devices and cellphones. With a simple tap on a screen, we can now communicate our ideas, emotions, and experiences in real time. Getting information is as simple as utilizing a smartphone with an internet connection, which is constantly within reach. Reading books, listening to music, taking pictures, watching films, playing games, generating and editing documents, and receiving medical advice are all options available on smart communication devices. Regardless of where we are, we can communicate with friends, family, coworkers, and millions of other people. When it comes to what we can accomplish with our communication gadgets, the options are limitless. For example, we use social media to remain up-to-date on current events, voice our perspectives, and exchange information with others around the world. Platforms such as Instagram, Facebook, Twitter (now incorporated as X[1]), Yelp,[2] and Fibre Tiger[3] have enabled a larger audience to participate in these by posting their experiences, making recommendations, and influencing purchasing preferences through virtual communities. As a result, social media platforms have become indispensable parts of our daily lives by transforming the manner in which we communicate and interact with others.

Customer reviews on the Internet and social media, also known as electronic word-of-mouth (eWOM), are essential to modern marketing strategies. eWOM has significantly influenced consumer purchasing decisions. This involves addressing negative feedback promptly and professionally, as well as thanking customers for positive reviews. Consumers can rate and comment on various products and services on social media platforms. This activity has significantly impacted consumer purchasing decisions and assisted businesses in upholding a positive brand value. Establishing a robust online presence and reputation is crucial for businesses in the current digital era, as it directly influences informed purchasing decisions. For example, consumers often use social media to gather opinions on relocation, online purchases, and service delivery. They consider user ratings and financial institution charges to make informed decisions, such as fees and interest rates. Information collected from social media can greatly influence consumer choices. According to a report in [1], 79% of businesses adhere to customer testimonials on social media, which is a great way to build credibility and trust followers while increasing the likelihood that prospective clients or customers will use the service of a real estate agent to find a house, open an account to purchase a particular product from the company, or invest in their services or products [2]. In South Africa, for instance, customers use Hellopeter[4] [3], a leading online review platform that facilitates consumer-business connections, allowing open communication, feedback experiences from previous users, and the discovery of exceptional enterprises, such as top financial institutions listed on the Johannesburg Stock Exchange (JSE),[5] where details of their interactions or contact with businesses are either good, bad, or average. These pieces of feedback significantly influence their financial decisions, as they democratize marketing, increase people-centric engagement, and enhance transparency with the abundance of data that can be harnessed to make informed marketing decisions [4]. Furthermore, by utilizing this data, companies can better tailor their marketing strategies to meet the needs and preferences of their target audience. Trends and consumer preferences can shift almost overnight, and social media platforms are uniquely positioned to help businesses adapt swiftly to these changes. However, because social media text lacks structure, machine learning algorithms (MLAs) face difficulty accurately interpreting context, tone, and sentiment, making it hard to analyze and derive valuable insights. Additionally, these results can be misclassified due to biased and unreliable labelling of text or customer reviews. Improving sentiment analysis (SA) and customer feedback interpretation requires developing a tool to analyze subjectivity in a snippet of online text in a natural language [5], extract the sentiment semantically, and preserve the crucial information needed for accurate analysis. This tool helps businesses gain valuable insights and make informed decisions by utilizing customer feedback.

Without indecisiveness, (SA), also known as opinion mining, is a computational method that analyzes individuals' subconscious feelings, even when fragments of text are under 280 characters [5], [6], [7]. Typically, SA approaches can be classified into two main categories: Lexicon-based and machine-learning methods. Lexicon-based methods, like SentiWordNet [8], [9], [10], calculate sentiment scores based on how often the word appears or the frequency of the lexicon's terms in the given text. They then use these scores to determine how each piece of text in a sentence can be categorized or labelled. For instance, the word ''sick'' can have a negative interpretation in a health-related text but a positive or negative meaning in a slang expression, which is normally used on an online platform to substitute an acronym in particular content [11]. Identifying the subject matter preference in each piece of text is not sufficient; it is crucial to capture both local and global context embeddings and interpret them in a way that resonates with human understanding [12]. SentiWordNet provides sentiment scores indicating positivity, negativity, or neutrality, essential for SA. Thus, every synset is associated with a *pos(s)* that indicate a positivity score, while *neg(s)* are used to indicate

[1]https://twitter.com/?lang=en
[2]https://www.yelp.com/
[3]https://www.fibretiger.co.za/
[4]https://www.hellopeter.com/
[5]https://www.jse.co.za/

a negativity score, and *obj(s)* from the given text indicate an objectivity (neutrality) score [8], [9], [10]. The scores are highly accurate, considering both the word and its context. All three scores range within the values (0,1). Again, a valence-aware dictionary and sentiment reasoner (VADER) is both a lexicon and a rule-based SA tool that is specifically attuned to sentiments expressed in social media [13]. VADER is open-source and can be directly utilized on unlabeled text data through the NLTK package. VADER is capable of detecting the polarity and intensity of emotion. AFINN is a popular wordlist-based approach consisting of 3382 words, each with a polarity score used for SA. The limitation of these approaches is their inability to model inherent subjectivity in natural language, and their inability to handle sarcasm, irony, and other forms of figurative language in social media texts or customer reviews is the problem. Still, supervised classification methods (SCM) like support vector machines (SVM), logistic regression (LR), and others have been used to train SA. These labels involve the polarity, subjectivity, and objectivity of input data, along with preprocessing tasks such as removing punctuation, HTML tags, and numbers, converting accented characters to ASCII, and converting all texts to lowercase. No matter the research efforts in SA, the existing solutions to employ SCM or MLA usually involve feature engineering, the solution design process, and the experimental evaluation process. The feature engineering process involves turning raw data into a set of features to represent each unit of textual data as a numeric vector [14], [15]. To achieve these, bag-of-words (BOW), or simple statistics of some order word combination (e.g., n-grams), is a method of extracting features from text for ML modelling. In BOW, the words in a text are extracted, and a list of all the words and their frequencies is made, meaning that a dictionary of all the words contained in the text is subsequently created. These are manual feature engineering approaches that fail to consider word order because different sentences may have the same representation [5]. At the same time, the approach is incapable of interpreting or detecting the sentiment of words and phrases for opinions expressed (positive, negative, neutral, or using assessments) by users of online platforms, and it is a time-consuming and labor-intensive task [15], [16]. Although bag-of-n-grams considers word order in a short context, it does not apply to SA tasks due to the sparse and high-dimensional data representations. In the second phase, a classification model is adopted to train the data and evaluate it to see which is most appropriate for different categories of sentiments (or use cases). Another challenge is ensuring the quality of labeled data and how accurately SCM or MLA can align with the ground truth. One popular method to get the gold standard-labeled dataset is using crowd workers to annotate each sentence's content using Amazon Mechanical Turk (AMT) [17]. Annotation with AMT is a time-consuming and costly process, which can be a hindrance for researchers. Moreover, the quality of crowd workers may decrease when dealing with large-scale text data [18], [19].

Recently, with the increasing usage of large language models (LLMs), researchers have developed a method known as zero-shot learning from a deep learning approach that can be applied to tasks like SA, text classification, and other domain-specific tasks. This approach combined reinforcement learning with human feedback (RLHF) to improve performance across a range of tasks. This approach is used in conjunction with ChatGPT, a new chatbot by OpenAI trained on GPT 3.5 that uses reinforcement learning from human feedback (RLHF) strategies to match human behavior better than crowd workers for annotating data [19], [20], [21]. Say, for example, given a review phrase from collected customer comments from Hellopeter, ''The battery life of this laptop is superb'', if we want to figure out the subjective feature ''battery life'', one needs to have a large quantity of annotated data with terms such as ''screen'', ''keyboard'' for the laptop domain or category. When it comes to SA, several public datasets have been released that are in the domain of restaurants [22] and movies [23]. However, there is a noticeable lack of datasets specifically designed for measuring customer feedback on financial products within the South African environment.To address this gap, we have created a comprehensive dataset from customer reviews on financial products, which is made publicly available at Data Science for Social Impact (DSFSI) Hugging Face group.[6] This paper aims to address the issues of annotating large-scale datasets and the semantic dependencies between data points and each sentence of customer reviews to determine the sentiment value. To do this, we proposed using ChatGPT, developed by OpenAI and trained on GPT 3.5, as a zero-shot learning (ZSL) model for labelling. Then, implement a deep learning model to learn how the human mind represents information in a piece of snippet text on social media (e.g., customer review) and add domain knowledge to visualise the contributing features to enhance better decision-making and make them easy to understand.

The unique contributions of this study to add to the theory and practice are the following:

- We present a method tailored to gather data from the Hellopeter website, organizing customer reviews for tasks like sentiment analysis. This method uses web scraping to efficiently extract the required information.
- To address the subjectivity in language and determine the sentiment of each sentence, we utilized OpenAI's ChatGPT as a tool for data labelling in various sentiment analysis tasks without prior training.
- We utilize OpenAI ChatGPT to understand the structure and context of customer reviews, combined with deep learning models, to predict sentiment for each sentence. Additionally, we apply oversampling techniques to address dataset imbalances. The outcomes demonstrate high accuracy and efficiency in performing sentiment analysis tasks.

[6]https://huggingface.co/datasets/dsfsi/hellopeter_financial_reviews

- An interactive system was created to illustrate the impact of sarcasm prediction on data collection and to showcase its performance through experimental comparison with established methods. The interactive system successfully showcases the value of sarcasm prediction in an online social media dataset and how AI technologies are used to outperform existing methods by accurately identifying sarcastic comments.

The content of this paper is succinctly organised as follows: Section I provides the introductory message. In Section II, we provide a comprehensive literature review on transfer learning, deep learning algorithms (BERT and LSTM), and traditional methods used in sentiment analysis. In Section III, we introduced the proposed OpenAI ChatGPT BRET-BiLSTM with detailed components of the system in Figure 1. In Section IV, we discuss the details of the traditional methods that include the lexical-based model (LBM), AFINN, SentiWordNet, and VADER in comparison to the proposed OpenAI ChatGPT BRET-BiLSTM model. In Section V, we examine the exploratory data analysis of the dataset, which includes crawling from the HelloPeter website and comparing it with the JSE's financial information to highlight the financial indicator. In addition, we use statistical methods to study the distribution of the dataset in order to evaluate and visually interpret it. The statistical analysis allows us to identify trends or patterns in the data. Section VI introduces our experiments and their outcomes, covering the experimental conditions, assessment criteria, evaluation procedures, including methods to address imbalances, and result interpretation. Lastly, Section VII summarises the paper and provides suggestions for future research in this area.

II. RELATED WORK

Customer review is critical for every organisation since it allows for a better understanding of the requirements and expectations of the customers. Businesses can identify specific areas for improvement and make the necessary changes to improve customer satisfaction by analyzing customer feedback via an online snippet, either on social media or by collecting data from the Internet based on customer dissatisfaction or engagement with the company. Moreover, responding to consumer feedback quickly and efficiently can help firms retain clients and keep them from looking for alternatives. Most of the time, researchers utilize SA, which integrates natural language processing (NLP) from artificial intelligence (AI), MLA-like data mining (DA), and information retrieval (IR) [24]. SA is a valuable tool for companies as it allows them to gain insights into customer opinions and emotions towards their products or services. SA can also help companies monitor their brand reputation and identify potential crises before they escalate. Customers have used online snippets that contain far more informal language than traditional financial statements or data to express their sentiments, opinions, and perspectives regarding the service satisfaction or dissatisfaction they received. In the past, simple factorisation methods or rule-based models like BOW, term-frequency-inverse-document frequency (TFIDF), pre-trained word embedding (PTWE), and VADER did not work very well. This is because words are often used for different intended purposes in different circumstances, spelling and grammar tend to not always be correct, and there needs to be a balance between grouping words by stemming, lemmatising, stopping word removal, and other things.

In aspect-level SA, the author in [25] shows that their method works better than popular ones like Naive Bayes (NB), SVM, and neural networks (NNs). Their method leverages a sentiment lexicon to create additional features for training a linear SVM classifier specifically designed for short, informal texts like Twitter posts. This innovative approach shows promising results in effectively analyzing sentiment in tweets. In addition, the study also acknowledged the limitations of relying solely on a sentiment lexicon, as it may not capture the nuanced emotions expressed in tweets. The author suggests looking into the future by incorporating contextual information or employing more advanced NLP techniques to improve the accuracy of SA in social media posts. Again, if a customer's opinion is important on the Internet because it is given freely and to better meet customer needs and remain competitive in the market, [26] proposed a BOW method that uses NLP to determine the sentiment score and magnitude of the sentence. This way, hidden information and the feelings of the user can be retrieved from the words. The results demonstrate that using Datafiniti's hotel evaluations, around 60% of the ratings can be anticipated and 40% are unpredictable. However, in this situation, BOW disregards context by ignoring word meanings and focusing on frequency of occurrence. For SA, especially when reviewing customer feedback with short text, this is a significant quandary because the order of the words in a statement might radically influence its meaning, and the model cannot account for this. The author in [27] for example, employed tagged bag-of-concepts (TBOC), which was created to address concerns with BOW and bag-of-concepts (BoC) for determining how someone truly perceives things. It looks at all the emotional and conceptual information in the text, with a focus on the short text. The TBoC method uses a domain-specific sentiment dictionary to find hidden feelings while keeping all important linkages and data to make SA more accurate. It also contains a mechanism for repairing broken text so that all of its meanings may be comprehended. In the end, the TBoC result was better than state-of-the-art (SOTA) methods like NB, SVM, and NNs, especially for aspect-level SA.

The rise of e-commerce has increased online product reviews, so by analysing the sentiments expressed in online product reviews and correlating them with financial data, companies can gain valuable insights into customer preferences. This understanding allows businesses to make informed decisions and develop or modify their products and services accordingly. The strategy enables top

financial institutions in South Africa to stay competitive by aligning their offerings with the demands of their customers. The author in [28] demonstrates a procedure to discover text features called sentence-level features (SLF) and domain-sensitive features (DSF). These features explore the significance of words at both the sentence-level and domain-level of product reviews, then employ a word-sense disambiguation-based method to extract SLF. For every similarity used to generate SLF, the SentiCircle-based method was enhanced to generate DSF. Several MLA and feature selection methods (FSM) were used in WEKA[7] [29] using various MLA such as Bayesian Network (BN), NB, Naïve Bayes Multinomial (NBM), LR, Multilayer Perceptron (MP), J48, Random Forest (RF), and Random Tree (RT) to test how well the proposed features worked compared to baseline features. SLF favourably escalates the performance of the SA task by 6.2%, 6.1%, and 6.0% for precision (PR), recall (RC), and F-measure, respectively. Meanwhile, the combination of sentence-level features and domain-sensitive features boosted the performance of supervised sentiment analysis by 7.1%, 7.2%, and 7.4% for precision, recall, and F-measure, respectively.

Reference [23] used SA on the IMDb movie reviews dataset to show how valuable insights can be extracted from a large text collection gathered online. These valuable pieces of information are extracted by employing four Machine Learning Algorithms: Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT). The author used six different ways to rate the performance of these four algorithms: the confusion matrix, the accuracy, the precision, the recall, the F1 measure, and the Area Under the Curve (AUC). In conclusion, using TF-IDF and LR gave the best validation AUC of nearly 96% for the task. However, the method faces a common challenge known as 'Out of Vocabulary (OOV)', which occurs when it cannot generate a representation for a word that is not in the training data. Due to the likelihood of having more positive or negative reviews than neutral ones, sentiment analysis often deals with imbalanced data, where one class of data significantly outweighs the others. In such situations, approaches like resampling the data or utilizing alternative evaluation metrics can be implemented to tackle this imbalance. Reference [2] introduces a hybrid method that combines Support Vector Machine (SVM) with Particle Swarm Optimisation (PSO). This method also incorporates oversampling techniques like the Synthetic Minority Oversampling Technique (SMOTE), SVM-SMOTE, Adaptive Synthetic Sampling (ADASYN), and Borderline-SMOTE to address imbalanced data. AUC, accuracy, and the G-mean were used to measure the results of the experiment. The G-mean finds the balance of the classification by multiplying both recall-negative (RECN) and recall-positive (RECP) by the square root. This shows that the PSO-SVM method with SVM-PSO (borderlineSMOTE) is more accurate than typical MLA, with an accuracy of 89.70%, and is better at making accurate classifications across different versions of datasets. These findings significantly enhance SA tasks, particularly given the rising volume of online datasets.

The author in [30] proposed a methodology to improve SA using preprocessing stages such as normalisation, word representation to extract attributes from input text using the TF-IDF vectorizer to construct the embedding, and SMOTE to correct imbalances in the datasets. Finally, the writer tested the suggested framework using six MLAs: Random Forest Classifier (RF-C), Multinomial Naive Bayes (MNB-C), Support Vector Machine Classifier (SVM-C), Gradient Boost, XGB, and Decision Tree Classifier (DTC). The author performed a performance experiment on an X sentiment dataset that had tweets from six different airlines, including ''AmericanAir'', ''VirginAmerica'', ''United'', ''SouthwestAir'', and ''JetBlue'' [31]. As a consequence, the RF-C gave the best results for SA with the selected dataset, with an accuracy of 98.3% and an F1 score of 0.98. Similarly, SVM gave incredibly good results with the selected dataset for SA, with an accuracy of 97.8%. The findings show that the resampling procedure influences the results of each ML classifier. While under-sampling has a significant impact on accuracy due to under-fitting, which results from a reduction in the majority class of an already small dataset, oversampling may be advantageous but may also lead to overfitting.

Most existing models select the best classification model, resulting in overconfident decisions that ignore the inherent uncertainty of natural language because the textual data on the Web has grown tremendously and has created unique contents of massive dimensions, which makes the polarity classification of text very challenging. The author in [32] used ensemble learning to address this issue and produce a more precise polarity prediction. It is based on Bayesian model averaging (BMA), where both the uncertainty and reliability of each single model are taken into account. Finding the best set of models to combine with the ensemble model is the biggest problem with BMA. To choose which model to use, the author employs the discriminative marginal that each classifier makes to the ensemble model, and a vector space model based on TF-IDF was adopted to reduce the feature space for learning. The researcher experimented with the suggested models on dictionary, NB, SVM, Maximum Entropy (ME), and Conditional Random Fields (CRF). They used the Sentence Polarity Dataset v1.0, which has 10,662 positive and negative movie reviews taken from Rotten-Tomatoes[8] [33], the Fine-grained Sentiment Dataset,[9] which has product reviews from Amazon.com [34], and the Multi-Domain Sentiment[10] [35]. The second type of evaluation is based on social datasets collected from Twitter (now known as **X**). Based on dictionary and NB, the method achieves 75.53% accuracy compared to 70.31% accuracy by MV,

[7]https://www.cs.waikato.ac.nz/ml/weka/

[8]http://www.rottentomatoes.com/

[9]http://www.sics.se/people/oscar/datasets/

[10]https://www.cs.jhu.edu/ mdredze/datasets/sentiment/

72.76% accuracy by MAX, 72.76% accuracy by MEAN, and 72.76% accuracy by PRODUCT for compositions tagged by the same experts. Other approaches achieve accuracy ranging from 80.90% of MV, 82.15% of MAX, 83% of MEAN, and 82.45% of PRODUCT, following a bagging paradigm. Results from experiments show that the suggested solution works very well and quickly because it uses an accurate and fast-running heuristic to set up a strategic mix of different classifiers. However, an increasing number of classifiers are to be closed in the ensemble, together with a large dataset open to deeper considerations in terms of complexity. Also, the selection of the initial ensemble should consider different complexities for each single learner and the inference algorithm, leading to areas of trade-off between their contribution in terms of accuracy and the related computational time.

In [36], the author suggested a way to automatically pull out sentiment expressions from the informal text. It uses optimisation to pull out sentiment expressions for a certain target (like a movie or person) from a set of unlabeled tweets. In particular, the goal is to find a wider range of sentiment-bearing expressions in tweets, such as formal and slang words and phrases, rather than just pre-defined syntactic patterns. Then, each sentimental expression should be judged on its target-dependent polarity. The polarity of a sentiment expression in a given text can be found by creating a new way to assign polarity using SentParBreaker[11] to perform sentence splitting and parsing each sentence using Stanford Parser[12] to get the dependency relations of words to a sentiment expression as an optimisation problem over the tweet corpus that has certain limits. The researcher tested the method on two types of data: tweets about movies containing 168,005 tweets, and tweets about people containing 258,655 tweets. For each religious group, the researcher trains the SVM classifiers using LIBLINEAR[13] [37] and applies 10-fold cross-validation to its dataset. The author also represents each user as a vector of their friends, where each vector refers to whether the user follows on Twitter (now known as **X**) (1 if the user follows up and 0 otherwise). The researcher tested the method on two types of data: tweets about movies containing 168,005 tweets, and tweets about people containing 258,655 tweets. The approach achieves a macro average of an F-score of 70.97% compared to other methods, and the improvement gets stronger as the vocabulary size grows.

The researcher in [38] examined the sentiment of a TripAdvisor,[14] one of the main tourist review websites in South Africa that use a hybrid approach as an alternate solution for this problem, which combines two original methodologies, namely, the lexical-based method and the machine learning-based method. The researcher employed SenticNet, a lexical-based technique, to label each review comment with the appropriate representative word, utilizing the TF-IDF formula for the unnormalized weight of words in each document for the full corpus. Then, to find out how people were feeling, the researchers used group classifiers like bagged decision trees (BDT), logistic model trees (LMT), stochastic gradient boosting (SGB), and bagged multi-layer perceptron (BMLP) models on both old and new review data. The test results show that the homogeneously distributed ensemble RF method is the most accurate, with a score of 98.13% based on a scrap attraction review from TripAdvisor and 55 features. This suggests that the homogeneously distributed ensemble RF method, which achieves 95.26% and 97.98% of the predictions, does a better job of predicting scrap attraction reviews than the other classifier methods.

The author in [39] compares six MLAs to find out how people feel about customer reviews from an online store like Amazon. The algorithms they use are NB, SVMs, RF, Bagging, and Boosting over WEKA. The researcher employs unigram (with or without) stopword removal, bigram (with or without) stopword removal, and trigram (with or without) stopword removal. The dataset used in this paper is a customer review dataset collected from the Amazon website about electronic products such as the Kindle, Fire TV Stick, tablet, and laptop.[15] The dataset consists of 34,661 records and 21 features. The results reveal that the RF technique provides the highest accuracy (89.87%) in the case of utilizing a unigram and stopping word removal, but the voting algorithm performs better in other circumstances.

Recently, researchers used new methodologies, called deep learning, for SA. It is one of the most prevalent and powerful ML methods that has been extensively deployed in SA and has demonstrated substantial possibilities and implications for SA performance and other tasks such as text classification. One of the advantages of deep learning in SA is its innate capacity to automatically learn and extract complex features from textual data, which can capture subtle characteristics and improve the accuracy of sentiment classification. Additionally, deep learning models have shown promising results in handling large-scale datasets, making them suitable for SA tasks involving vast amounts of text data. Most of the deep learning models used word embedding, which is a type of word representation in which words are transformed into vectors. The author in [40], for example, used various NLP methods to conduct sentiment classification using binary and multiclass labels. Following several SVMs, RF, and logistic classifiers, the researcher used a BOW and word2vec with skip-gram for binary classification. The author aggregated word vectors into a single feature vector for each review using vector averaging and clustering. For the RF, SVM, and LR, the results of binary classification using Word2vec with averaging were 84.0%, 85.8%, and 86.6%, respectively. Word2vec with RF and clustering had an accuracy of 83.5% on a publicly available Kaggle[16] competition called "Bag

[11] http://text0.mib.man.ac.uk:8080/scottpiao/sentdetector
[12] http://nlp.stanford.edu/software/lex-parser.shtml
[13] https://github.com/cjlin1/liblinear
[14] https://www.tripadvisor.co.za/
[15] https://www.kaggle.com/datasets/bittlingmayer/amazonreviews
[16] https://t.ly/N0ohU

of Words Meets Bags of Popcorn.'' For the recursive neural tensor network (RNTN) for the multi-class case, the author achieved a loss of 0.25% after about 40 epochs using AdaGrad stochastic gradient descent with a mini-batch size of 30, L_2 regularisation with a strength of 10^{-6}, and a learning rate of 10^{-2}. One model, AraVec-Web [41], was made using the word2vec Skip-Gram technique [41], and the other is fastText Arabic Wikipedia word embeddings [42]. They were compared for Arabic aspect-based SA. The compared results showed that the performance of fastText Arabic Wikipedia word embeddings is slightly better than AraVec-Web. The author in [43] also uses FastText. It is used for both skip-gram and continuous BOW models in training and building sentiment-specific embeddings. After training and testing five classifiers on the three selected datasets, the results revealed that CBOW and skip-gram models perform well on syntactic and semantic analogies. The author in [44] adopted the word2vec model to analyze citation sentiment by constructing sentence embedding, which is formed on word embedding to obtain an average of the vectors of the words in a single phrase that has been taught using word2vec. The sentence embedding is evaluated using ACL embeddings (300 and 100 dimensions) from the ACL collection. ACL anthology reference corpus[17] containing 10,921 canonical computational linguistics papers, where 622,144 sentences were generated after filtering out sentences with lower quality. The results show that Word2Vec is successful and promising in distinguishing positive and negative citations; nonetheless, handcraft features outperform Word2Vec.

The researchers in [45] used the Word2Vec model's word embedding. In their tests, they used both Word2Vec models, skip-gram, and CBOW. The Word2Vec models were used to train four classifiers: Gaussian naíve Bayes (GNB), Bernoulli naíve Bayes (BNB), SVM, and LR. The SVM and LR employing the skip model exceed the naíve Bayes classifiers in terms of performance. Social media platforms have been used for various things outside of promoting goods and services based on customer feedback. The author in [46] used Word2Vec and weighted averages of Word2Vec to detect incited terrorism conversations in a given online text message known as a tweet. In addition, the feature vector is tested on MLAs such as SVM and RF, and then the proposed model is validated using cross-validation to show that the use of Word2Vec by weighted average with a 75% average for all the performance matrices (e.g., precision, recall, and F-measure) is slightly better than the Word2Vec method. The author in [47] implemented the Skip-Gram Model of Word2Vec to understand contextual information about words and minimize the high-dimensional space of word vectors to improve the sentiment classification accuracy of tweets relating to the U.S. Military Base in Ghana. A RF classifier is used for training and evaluation performance using accuracy, recall, precision, and F-measure metrics. The overall accuracy for the sentiment labels was 81%, which suggests that the word vector quality that the skip-gram model produces contributes to good results in sentiment polarity prediction.

The research in [48] proposed a new similarity distance model called the semantic orientation pointwise similarity distance (SO-SD) model. They used Word2Vec to make a sentiment dictionary based on their model. An emotional dictionary was created to determine the emotional tendencies of Weibo messages. The experiments showed good results using this approach. As the name suggests, this method uses Word2Vec and a brand-new technique called SVMperf to make multivariate performance measures better [49]. SVMperf trains faster and makes more accurate predictions than other SVM packages. It is used to improve the classification of people's emotions, as suggested in [50] on a set of Chinese review feedback on Amazon clothing products. The result showed that Word2Vec captured the semantic features of the Chinese language by grouping the features with similar input text (or contextual information). Next, Word2Vec and SVMperf were used to train and classify the comment texts again. The author's findings highlighted the superior performance of their method for sentiment classification. The author in [51] investigated the connection between sentiment categorization, emoticons, and the situations in which they are employed. To interpret emoticons in the context of tweets, they used Word2Vec to define the representation of the words in the dataset, which included emoticons. They clustered the words using the k-means technique so that the exact meaning of the emoticons could be deduced from the words that showed up in the same groups (or clusters).

To address SA in finance, market participants must constantly monitor financial and economic news and make every effort to ensure that all existing knowledge is reflected in stock prices and that new information is absorbed immediately in determining future stock prices. The author in [52] presented a platform for evaluating how well different SA methods are performing by combining different ways of representing text with machine-learning classifiers. The author performs more than one hundred experiments using publicly available datasets, labeled by financial experts. Subsequently, the authors evaluated the proposed method with specific lexicons for SA in finance. It was then expanded to include the newest transformer model and word and sentence encoders. The results show that contextual embeddings function better for SA than lexicons and fixed word and sentence encoders, even when large datasets are not available. Additionally, distilled versions of NLP transformers are effective just as well as their larger teacher models, which means they can be used in production settings. The researcher in [53] offered a fine-tuning of pre-trained BERT to recognize a text's sentimental inclination towards a certain element. To improve the performance of fine-tuning BERT, the concept is to use the last output layer of BERT and ignore the semantic knowledge in the intermediate layers. The authors add an extra pooling module to the already-trained BERT as a way to combine the multi-layer

[17] http://acl–arc.comp.nus.edu.sg/

representations of the classification token. The retrospective of the study looked at how well the model did on ABSA and ACL 14 Twitter. Experimental results showed good performance similar to the SOTA method, and the model could be used for other NLP tasks as well.

Chat-bots are another extension of NLP applications. Chat-bots are a type of AI computer that seeks to deliver proper responses to inquiries by simulating human communication processes via text or voice processing methods [54]. Chat-GPT (Chat Generative Pre-trained Transformer) is a chatbot that has exploded in popularity since its debut. On November 30, 2022, OpenAI made the AI chatbot ChatGPT available for public usage [55], [56]. ChatGPT is a supervised and reinforcement-learning NLP model that has been optimized. ChatGPT is one of the largest language models and contains the most parameters, with 175 billion [57]. ChatGPT is more than just an advanced question-answering (QA) robot; with its multi-language capabilities, it can write articles on a single topic in several languages [58], [59]. In the near term, the author in [60] instructed ChatGPT to construct research summaries based on the names and publications of research abstracts collected from five high-impact medical journals. ChatGPT was effective at producing scientific summaries from the collected abstracts, and plagiarism checkers did not recognize the summaries it produced. Also, the researcher in [61] investigates and uses ChatGPT in the realm of business customer SA. The use of ChatGPT helps identify opportunities that could be used to improve the products and services of the organization that have been pointed out by the analysis of customer reviews. Similarly, [62] conducted a research study that examined the performance of supervised and unsupervised MLA for SA. The preliminary findings demonstrated that supervised MLAs outperform unsupervised MLAs like Lexicon-based methods in terms of performance accuracy. Obtaining adequate annotated training data for supervised MLA is, however, time-consuming and costly. A valuable contribution is a discussion in [63], which emphasizes the vitality of using more advanced sentence models. Specifically, the author highlighted how sentences manipulated using GPT-3 can generate semantically incoherent outcomes that are promptly recognized by humans. An instruction was passed to GPT-3 to generate synonyms for negative words and then use these to reformulate sentences and evaluate the robustness of the models. Additionally, the authors underscore FinBERT's resilience against adverse attacks, especially when compared to traditional keyword-based methods (KBM).

While there's a surge in studies employing cutting-edge models for SA, there's no shortage of holistic overviews and thorough reviews that trace the evolution of this field over time. The researcher in [64] provides a comprehensive overview of SA, offering a comprehensive view of the subject, its methodologies, applications, and developments in the field. Also, the work in [65] encompassed both traditional methods and newer models, including BERT and GPT-2/3, spotlighting their roles and advancements in the domain of SA. GPT models, particularly the most recent GPT series, have emerged as favorable in recent investigations within the scope of SA. In [66], the author used the GPT-3.5 Turbo model to perform SA on social media posts. The author used the RoBERTA model for comparison analysis, benchmarking, and, in particular, to credit the mode's distinctive role to social scientists. The author in [19] deployed the GPT3.5 Turbo variant for SA on Amazon reviews. They revealed a major enhancement in accuracy. VADER and TextBlob were used as benchmark models for sentence classification. However, the specifics of the prompt they used were not specified in their publication. The author in [67] used GPT-1, GPT-2, GPT-3, GPT-4, and BERT to estimate the financial performance of a stock price based on the news segmentation in the newspaper or on the Internet. Interestingly, GPT-1, GPT-2, and BERT models are not particularly effective in accurately predicting a good return as a positive priority. The researcher in [68] attempted to figure out appropriate Twitter (now known as **X**) users within the financial community. The author found a correlation between a weighted sentiment measure using messages from these essential users and major financial market indices. While [69] addressed the interaction between Twitter (now known as **X**) sentiment and stock returns, they focused on expert users whose tweets predominantly revolved around financial topics. For SA, they utilized a dictionary-based method (DBM). The author in [70] explored the emotions expressed in 2.5 million Twitter (now X) messages about specific S&P 500 firms and their stock market performance. They discovered that untweeted tweets from individuals with fewer than 171 followers (which was themed) had a significant impact on the company's stock performance the next day, as well as 10 and 20 days later. The sentences in the tweets were analysed using the Harvard-IV dictionary.

The rise of social media platforms has heightened interest in discovering polarised viewpoints on specific topics. However, due to the complexities of human language and cultural and geographic variances, collecting sentiment from consumer speech in natural language is extremely challenging. The article uses ChatGPT-based zero-shot learning (ZSL) to overcome this challenge. This strategy uses the prediction model's awareness to discover causal relationships between words and concepts by assigning unique emotion scores to each dataset. The annotated data is then fed into BERT and BiLSTM to generate semantic instincts expressed over a SoftMax function to make predictions. Visual methods were employed to get insights into how sentiment aspects influence public perceptions of comprehending consumer comments and opinions in today's financial or online platforms. Table 1 provides an overview of the related research, categorised by model, feature extraction, and performance metrics. Metrics like recall (RC), accuracy, F-measure, and others are used to evaluate the sentiment of a given textual dataset or sentence. In this study, we used customer review messages posted

TABLE 1. Comparison of related studies based on common characteristics.

Reference	Features	Model	Metric	Dataset
[27]	TBoC	NB, SVM and NN	ACC, FM, PC and RC	SemEval
[25]	Frequency, Unigram	linear SVM	ACC & Cross-validation	Twitter (now X)
[26]	Frequency count, n-gram	BoW	Ratio	Hotel
[28]	Text feature with SLF and DSF	BN, NB, NBM, LR, MP, J48, RF, and RT.	PR, RC and FM	Amazon[18]
[23]	BoW	NB, LR, RF, and DT	ACC, PR, RC and FM	IMDb movie reviews[19]
[30]	*n*-gram, BoW and PSO	SVM, PSO-SVM, SVM-SMOTE and ADASYN	ACC, AUC and FM	Restaurant Reviews
[31]	TF-IDF	RF-C, MNB-C, SVM-C, GB, XGB, and DTC	ACC, and FM	Twitter (now **X**)
[33]	TF-IDF	DIC, NB, SVM, ME, and CRF	ACC	Twitter (now **X**) and Amazon
[37]	Language Parser	Liblinear SVM	FM	Twitter (now **X**)
[39]	SenticNet, TF-IDF	BDT, LMT and BMLP	FM	Twitter (now **X**)
[40]	n-gram model	Voting methods	ACC	Amazon website
[41]	BOW and word2vec	RNTN, RF, SVM, and LR	ACC	Kaggle
[74]	CBOW	SVM, LR and SGD	ACC	ASTD-ArTwitter,-QCRI, LABR, MPQA
[44]	Word2Ve with Skip-gram	Linear SVM	ACC, PR, RC and FM	Airline Tweets (now **X**)
[45]	Skip-gram,CBOW and FastText	SVM,RF,LR and SGD	FM	LABR, ASTD, MPQA
[46]	word2vec	SVM,	FM	ACL (100 and 300) and Brown100
[47]	CBOW	SVM,GNB, and LR	ACC, PR, RC, and FM	Airlines Tweets (now **X**)
[49]	skip-gram	RF	ACC, PR, RC and FM	Twitter (now **X**)
[50]	skip-gram and CBOW	SVMperf	ACC, PR, RC and FM	Sina Weibo
[53]	skip-gram and CBOW	NB	ACC, PR, RC and FM	Twitter (now **X**)
[54]	Lexicon, TF-IDF, Sentence embedding and Transformer	SVM, XGBoost, CNN, RNN	ACC, PR, RC and FM	FinPhrase and SemEval
[55]	BPE[20]	BERT-LSTM and BERT-Attention	ACC and FM	ABSA[21] and SNLI[22]
Proposed method	**OpenAI's GPT zero-shot learning**	**BERT-BiLSTM**	**ACC, FM and visualization**	**Hellopeter [3] & FinData**

on an online platform like HelloPeter interchangeably to denote input text or data related to consumer feedback for financial services. Looking at Table 1, which summarises previous research methods and datasets, it is clear many of these datasets are publicly accessible and annotated. On the other hand, our dataset, employed to validate our proposed model in Figure 1, is a newly curated dataset sourced from customer feedback on financial services in South Africa. In this research, we amalgamate ChatGPT with a zero-shot learning (ZSL) model to understand the subjective sentiment in text data using deep learning for accurate sentiment predictions. Additionally, the study adapts the Synthetic Minority Oversampling Technique (SMOTE) by creating samples from minority classes by using samples that are similar to those in the minority class to address the bias as a result of an imbalance in the dataset, which is often a limitation in most existing studies. This makes the model have access to more classes while it is being trained [71]. For consistency, a text, textual, or sentence refers to the text data submitted by a customer or user on the HelloPerter website and retrieved through the platform's API. A document or writing sample is used interchangeably to refer to the minimum unit of text data to be analysed or annotated. The text dataset may include emails, customer reviews, social media posts, or any other form of written content.

III. METHODOLOGY

In this paper, the primary focus of the study is to explore the use of ChatGPT as an annotating technique for SA tasks and evaluate it on distinct sentiment datasets with varying purposes. The architecture is made up of different methods, such as OpenAI's ChatGPT, which is a zero-shot learning method for making high-quality human-like responses and understanding how different customer questions and feedback are when annotating the dataset (see Figure 1). Next, we use bidirectional encoder representations from the transformer (BERT) and bidirectional long short-term memory (BiLSTM) to capture the contextual semantic representation. We then employ a SoftMax layer to determine the

[18]http://jmcauley.ucsd.edu/data/amazon/

[19]Kaggle Bag of Words Meet Bag of Popcorn Challenge https://www.kaggle.com/code/yagli18/bag-of-words-meets-bags-of-popcorn

[20]byte-pair-encoding.

[21]AspectBasedSentimentAnalysis.

[22]NaturalLanguageInference(NLI).

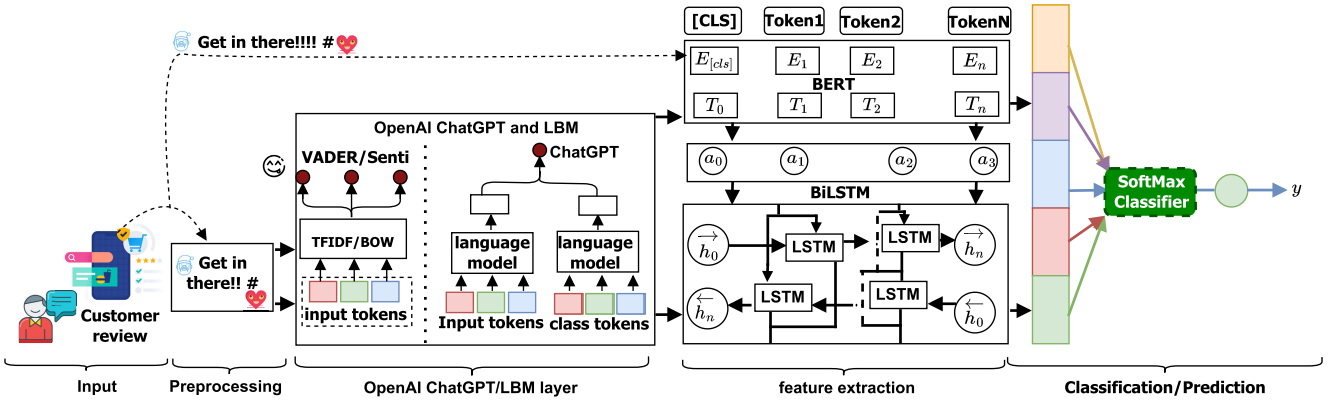


FIGURE 1. Structure of OpenAI ChatGPT BERT-BiLSTM network for sentiment analysis.

sentiment orientation of the dataset and, with high-precision visualisation capability, interpret the feature contribution. The proposed model demonstrates superior accuracy and efficiency compared to the SOTA models.

A. INPUT LAYER

The input layer uses datasets from the HelloPeter website to train the proposed architecture shown in Figure 1. HelloPeter is a South African-based website that focuses on consumer reviews. Unlike most websites, HelloPeter focuses completely on customer reviews. HelloPeter's website allows customers to browse businesses, read reviews, and offer feedback on their services. The website covers a variety of businesses, such as banking, insurance, telecommunications, fast food, and autos. HelloPeter offers a business tool that enables businesses to view and manage reviews. Companies that join HelloPeter receive consumer engagement data based on their platform assessments. HelloPeter allows users to link their evaluations to other platforms, such as websites and social media, to improve visibility and SEO. The textual data, which includes consumer feedback and comments, is entered into the system for analysis and reporting.

B. PREPROCESSING

The preprocessing layer is crucial for improving machine learning models' performance by providing high-quality input data. The preprocessing layer involves specific actions such as anonymizing data, splitting it into training and testing sets, replacing names to protect privacy, and removing personally identifiable information. Sensitive information, such as email addresses and phone numbers, is also removed. HTML elements are removed to reduce unnecessary styling. Accent characters are removed for readability, and contractions like *"can't"* are expanded to *"cannot"* for better understanding. The textual dataset was standardised by converting all uppercase letters to lowercase, removing special characters, and replacing them with corresponding English alphabet characters. SpaCy from the Python library is used for lemmatization, which is the process of extracting basic word forms from a set of data. This step ensures consistency and accuracy in the results obtained [73], [74]. This step ensures consistency and accuracy in the results obtained, ultimately improving the overall quality of the analysis.

C. OpenAI ChatGPT/LBM LAYER

The OpenAI ChatGPT allows for advanced natural language processing capabilities with the use of artificial intelligence to capture the full range of human sentiment from a given instruction in a natural language with human feedback to fine-tune the required annotation of the text compared to a lexical-based model (LBM), whose reliance is upon a pre-defined sentiment lexicon. In our quest to understand the subjectivity of the comments, feedback of the user, and textual content we extracted from HelloPeter, we use OpenAI ChatGPT built on the GPT3.5 architecture with "*text-davinci-003*" embedding [75] trained on extensive text data sets to analyse the sentiment and tone of the reviews efficiently to generate human-like responses and comprehend diverse subjects comprehensively [76]. With such impressive capabilities, it becomes imperative to explore the possibility of this approach, e.g., OpenAI ChatGPT, called the ZSL model, to understand and develop advanced annotation tools and improve automated content creation. The aim is to investigate the efficacy of OpenAI ChatGPT as a text annotation tool for sentiment analysis and natural language understanding tasks. The primary objective in this layer is to use OpenAI ChatGPT to generate responses that sound human-like and are helpful as text annotation tools for labelling or annotating customer service interactions in various domains and text sources, including customer reviews, social media posts, or new articles. As a result, in this layer, we use the OpenAI ChatGPT pre-trained language model (LLM) to get around the problems with traditional methods like LBM, which need a lot of labelled data to make more accurate and nuanced sentiment analysis results by incorporating human feedback to label or annotate customer service interactions. Given an input text or sentence x from

a set of training datasets, we instructed OpenAI ChatGPT to generate the sentiment score of the content in the x sentence, and we used the score to annotate the entire sentence in the dataset. Then, we investigate the performance of OpenAI ChatGPT against other cutting-edge sentiment analysis techniques such as AFINN, SentiWord, and VADER, as discussed in Section IV. Then, we integrated the annotated categories into the dataset and fed it to the BERT-BiLSTM layer to extract feature vectors and the Softmax layer to get an accurate reading of the text's sentiment polarity.

IV. LEXICON-BASED MODELS

A lexicon-based model requires a predefined lexicon, e.g., a stock of terms that belong to a particular subject or language. The approach uses a sentiment lexicon with information about which words and phrases are positive and which are negative [77]. To look at sentiment, we cleaned the data and then used three lexicon-based models: AFINN-lexicon [78], SentiWordNet [79], and Valence Aware Dictionary and Sentiment Reasoner (VADER) [13]. We did this to see how well they worked as follows and to compare them:

A. AFINN-LEXICON MODEL

AFINN is a rule-based process that uses statistical modeling to develop a hybrid approach to sentiment classification [80]. It is based on comparing a sample of each review in the dataset with a list of weights of positive or negative keywords derived from the affective norms for English words in the dataset [81], [82]. The AFINN is a list of manually labeled English words with integer values ranging from 5 (very negative) to +5 (extremely positive). Using the lexicon, a value is assigned to each word in a tweet. The values are averaged to create the sentiment score for the entire textual data or messages in the dataset, with computing speed being one of its significant features [83]. The AFINN lexicon has been widely used in SA tasks due to its simplicity and effectiveness. It provides a quick and efficient way to determine the overall sentiment of a text by assigning scores to individual words. This makes it particularly useful for analyzing large datasets with limited computational resources.

B. SentiWordNet MODEL

Our exploration into SentiWordNet involved associating sentiment scores with each word in the reviews. By analyzing the parts of speech (POS) and using SentiWordNet's synsets, we calculated positive, negative, and objective scores for each review. This fine-grained analysis allowed us to derive sentiment and objectivity measures for the entire text [84], [85]. We defined a sentiment as 'positive' when the normalised sentiment score was 'positive' and 'negative' when it was negative. The SentiWordNet model offered in-depth insight into the sentiment composition of the reviews [86]. By utilizing SentiWordNet's synsets, we were able to obtain a comprehensive understanding of the sentiment and objectivity levels within each review. This detailed analysis provided valuable information on the overall sentiment composition of the reviews, enabling us to accurately classify them as either positive or negative. The utilization of the SentiWordNet model greatly enhanced our ability to delve into the intricacies of sentiment expressed in the text.

C. VADER

It is a lexical database and rule-based SA tool that is appropriate for each of the customer evaluations in our dataset. It employs a wide range of methodologies, such as gathering lexical features (e.g., words) that are rated as positive or negative depending on their sentiment polarity. It displays not just the positivity and negativity scores but also the degree to which a sentiment is positive or negative [87]. By combining grammatical rules and syntactical patterns, VADER can precisely identify sentiments for each customer review data point across the whole dataset. The approach gives a more inclusive assessment of sentiment by assessing the degree of positive or negative emotions represented in the text [88], [89]. In this paper, we used VADER to establish a compound sentiment, e.g., a score that took into account both positive and negative sentiments, assisting in determining the overall sentiment of a review. We used 0.4 as a threshold value to categorize reviews as 'positive' or 'negative' [84], [85].

V. DATASET AND EXPLORATORY ANALYSIS

In this section, we analyse the basic peculiarities in the dataset. This includes customer review feedback from the HelloPeter website and financial data from the top financial institutions listed on the JSE. We aim to understand how customer feedback influences financial indicators. We use a statistical model to study how customer reviews are structured. This help us assess overall customer satisfaction by analysing the frequency of terms or phrases in each sentence of the dataset.

A. DATASET

The dataset in the experiment is based on customer review feedback collected from Hellopeter and the financial statement data of the top financial institutions in the JSE for five years.

1) CUSTOMER REVIEWS DATA

Customers post short messages on the Hellopeter platform, a leading consumer review network connecting South African consumers to businesses, which serves as a channel for consumers to interact with businesses directly. These messages contain customer experiences, thoughts, opinions, commentary on market trends, and insights into market stocks for five top South African financial institutions: ABSA (ABG), Standard Bank (SBK), Capital Bank (CPI), Nedbank (NED), and First National Bank (FSR). The data were collected from January 2018 to December 2022, using

TABLE 2. A comprehensive summary of the dataset and tokenisation of the dataset.

Dataset	Size	Positive	Negative	Token	Features
ABG	20,286	8466	11820	2780995	35942
SBK	22,776	9530	13244	3022224	35359
CPI	16,456	6485	9971	2084806	28960
FSR	54,900	24338	30561	6865630	54547
NED	14,923	6026	8897	2013942	29343

Hellopeter's publicly available API[23] to crawl attributes related to the author (user), author display name, author ID, date, review rating, review content, business name, and other information. Table 2 provides a comprehensive summary of the datasets collected for each institution, including their corresponding attributes, different versions of tokens, and the total number of features.

In addition to the customer reviews that were collected from the HelloPeter website, we went ahead and stored each review according to class type, such as ABSA, Capitec, Standard, Nedbank, and FNB. Then, we used the ChatGPT Open AI zero-shot learning model to annotate each sentence in the review based on how subjective it is to learn in natural language with human feedback to help them understand the sentiment. Two options were available under each sentence of the review: negative and positive. Consequently, the class label of each sentence in the reviews was assigned based on the capabilities of ChatGPT's pre-trained language model.

Following that, each review was saved in its file, and all files containing reviews were collected and kept according to the class type, which is the name of the financial institution. A CSV file containing all of the reviews, together with their context in one column and the associated class designation in the other, and so the total size of the positive and negative reviews, is shown in Table 2. After annotating, each of the datasets was preprocessed by removing the stopword, duplication, and non-English letters. It reduces the overall number of features to enhance feature selection, where irrelevant features without meaning are eliminated. Then, bag-of-word (BOW) as a feature extraction method is applied for text tokenization from all the datasets collected via Hellopeter (see Table 2). As observed in Table 2, we observed different features for each of the datasets with tokens based on their sentiment and total sizes after preprocessing and stemming.

2) FINANCIAL DATA

Stock markets are volatile, making it challenging to predict future stock prices. To gather data, we used Yahoo Finance APIs[24] [90] and Finchat APIs[25] with Stratosphere to identify unique company IDs from top South African banks listed on the JSE.[26] We also obtained the consumer price index (CPI) and financial stability review (FSR). Our data collection spanned 2018–2022, with an extension into 2023 for specific institutions. Some institutions have already disclosed their 2023 financial results, such as FSR and CPI, which contribute to the quality of our analysis. The data was verified for accuracy and reliability by comparing it to official bank financial statements [91], [92], [93], [94], [95].

3) SELECTION OF FINANCIAL INDICATORS

We provide insight into the careful selection of specific financial indicators that are pivotal to our research objectives. The importance of the financial indicators in evaluating the financial performance and health of the chosen institutions guided our decision. The chosen metrics—Interest Income, Total Interest Expense, Total Net Interest Income, Non-Interest Income, Total Revenues Before Provision For Loan Losses, Provision For Loan Losses, and Total Revenues—are important for understanding different parts of the institution's financial health, profitability, and risk management. By focusing on these metrics, we aim to gain a holistic understanding of the financial landscape within which these institutions operate [96]. We improve our analysis by including daily, monthly, and yearly stock market data, as shown in Figure 2 (a,b). This allows us to look into how market dynamics and financial performance interfere with each other [97], [98]. The financial indicator selection helps align the proposed model (see Figure 1) with industry best practices and enables a rigorous evaluation of financial resilience for growth prospects.

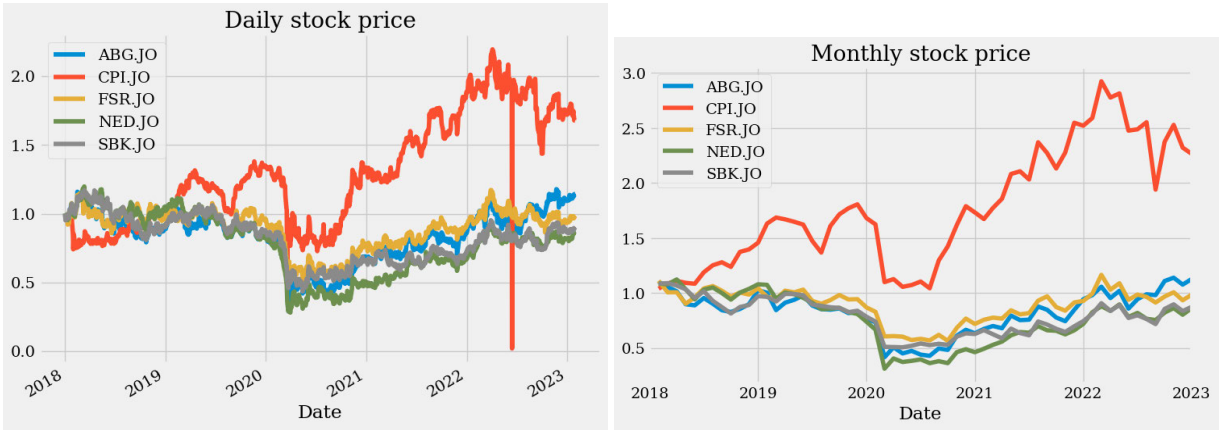
B. VISUALISE INTERPRETATION OF THE DATASET

This section (IV-B) uses statistical methods to evaluate, interpret, and demonstrate the distribution of textual evidence offered by HelloPeter users. This section also contains visual representations of the data to help you understand the patterns and trends.

1) UNIVARIATE DISTRIBUTIONS

To compare the different univariate distributions, we look at the approach of probability binning (PB), which involves dividing the distributions into a relatively small number of bins. The number of events falling into these bins is compared for a test with training samples, and a chi-squared computation is performed on the counts (i.e., the square of the differences divided by the sum). Rather than the standard binning algorithm, which selects bins of equal width, the binning algorithm is selected with each bin containing the same number of events. The result is a randomly selected occurrence from the training sample with an equal probability of falling into any of the bins. This process results in bins of unequal width (see Figure 3), with the property that each bin carries equal weighting when used for further statistical tests. Figure 3 demonstrates a plot of sentiment polarity count, indicating the sentence length at each 0.05-interval interval.

[23] https://business.hellopeter.com/docs/api/v5
[24] https://developer.yahoo.com/api/
[25] https://finchat.io/api/docs/
[26] https://www.jse.co.za/



(a) Daily Stock Price Distribution (b) Monthly Stock Price Distribution

FIGURE 2. Distribution of financial indicator from financial statements.

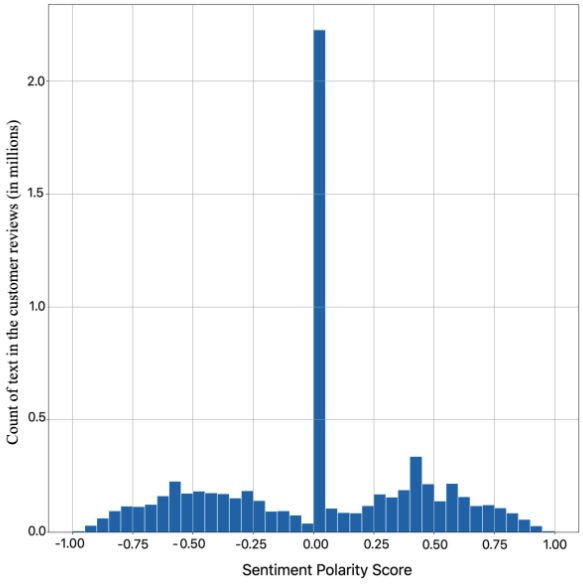


FIGURE 3. Polarity count.

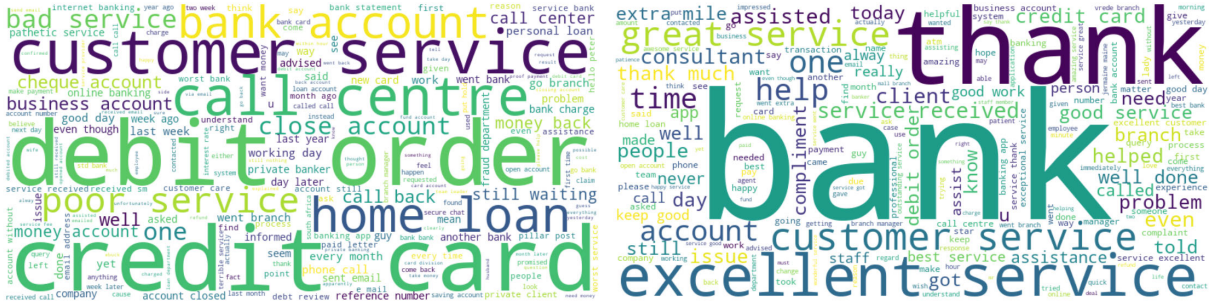
Each set of training data is subjected to PB using either 10, 20, 30, or 40 bins, and the mean and standard deviation of $1,000_{\chi'^2}$ values are calculated for sentiment polarity scores. The resulting χ'^2 distributions for each case are nearly normally distributed (data not shown), making the standard deviation of the distributison an appropriate measure of the variance of χ'^2. Therefore, the positive sentiment messages from the customer reviews posted on Hellopeters show polarity scores that are in the middle range (approximately between $+0.25$ and $+0.75$). Similarly, most negative sentiment tweets have polarity scores that are in the middle range (approximately between -0.25 and -0.75). In other words, text conversations in customer reviews with extreme negative or positive polarity scores are uncommon.

2) WORD CLOUD

We use word clouds, which are visual representations of each sentence in the text, to understand customer sentiments behind words and determine whether a message indicates a positive, negative, or neutral reaction. Here, we use a maximum of 300 words in the word cloud, and the minimum frequency is 50, which means any word that appears 50 times is included in Figures 4. Furthermore, we change the scale so that the highest frequency word is 5 and the lowest frequency word is 0.3. The word cloud in Figures 4 (a, b) provides a visual representation of the most common words used by consumers when expressing their satisfaction or dissatisfaction with a service. This analysis allows businesses to identify the key areas that customers appreciate or find lacking, helping them improve their overall service quality.

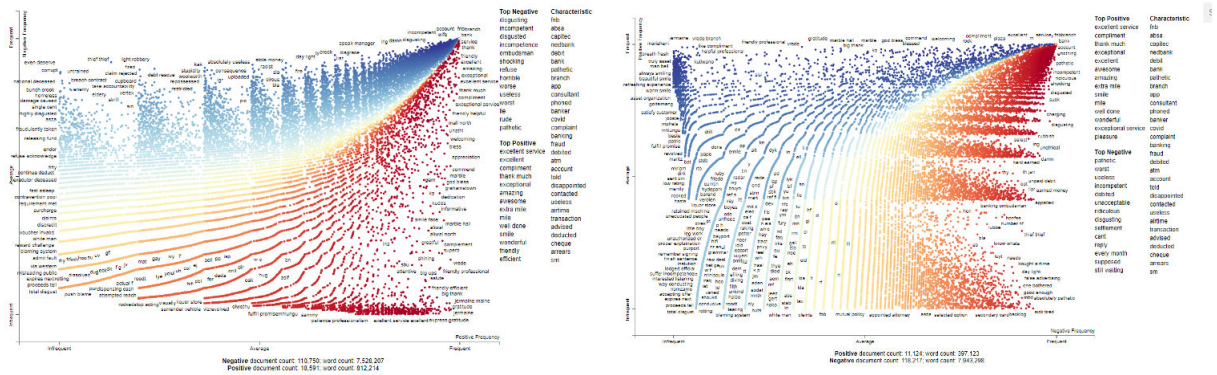
3) WORDS AND PHRASE VISUALISATION

In this section, we utilize Scattertext [99] to evaluate the customer review data to discover the overall trends and characteristics, as well as terms significantly connected with consumer demands and preferences. The scattertext is an interactive tool that identifies or distinguishes phrases whose frequency occurs in each sentence of a text and displays them in a scatter plot with non-overlapping term labels. Scattertext is effective in identifying phrases that reflect two opposing concepts, i.e., comprehending what customers are saying or identifying nuanced patterns in text data. In doing this, we construct a model using Python's ''numpy'', ''pandas'', and ''collections'' features with a certain number of texts and frequency of words and phrases used to understand the flow or changes in subjects using scattertext and a scaled F-score. Figure 5 (a) and (b) show the words and phrases in each sentence of the customer reviews from HelloPeter. These figures show noun phrases using BOW and n-gram features for customer review comment posts on Hellopeter with positive and negative sentiment, respectively. Figures 5 (a) and (b) show that the features used



(a) Negative (b) Positive

FIGURE 4. A word cloud comprising negative and positive customer reviewer keywords.



(a) The visualization of empath categories shows red categories indicating negative sentiment feedback and blue categories indicating positive sentiment tweets, with the intensity of each color indicating its strength.

(b) Phrase visualization indicates negative sentiment feedback in red and positive sentiment feedback in blue, with the intensity of a phrase's color indicating its category link.

FIGURE 5. A word cloud comprising negative and positive customer review keywords.

more frequently by positive sentiment posted on Hellopeter appear higher on the y-axis and closer to the upper left part of the chart. On the other hand, features used more frequently by negative sentiment posted on Hellopeter are further right on the x-axis and closer to the lower right part of the chart. Features used frequently in both categories are closer to the upper right part of the chart, while features that are used infrequently in both categories are closer to the lower left part of the chart automatically generated using scatter-text [99].

Figures 5 (a) shows that the most frequent terms in negative sentiment customer feedback posted on Hellopeter include terms that are disgusting, shocking, lying, rude, and pathetic. By contrast, the most frequent terms in positive sentiment posted on Hellopeter conveyed excellent, awesome, amazing, friendly, efficient, and compliment using BoW features on the scattertext graph. Terms that frequently appeared in both categories were related to appreciation, exceptional service, banking, kudos, and salute. Terms that infrequently appeared in both categories were related to fifty, highly disguising, bunch crook, corrupt, total disguising, and discredit. Terms that were used with average frequency in both categories are related to stolen money, huge, fulfilling promises, and useless. Figures 5 (b) show that the most frequent phrases associated with negative sentiment posted on Hellopeter that reflected customers' concerns about issues using *n*-grams features include threat, market link, priority pas, etc. On the other hand, positive sentiment was posted that reflected users' interest in issues such as incompetence, excellent services, and horrible. This indicates a misinterpretation of words, which could cause an imbalance in the dataset.

VI. EXPERIMENTS AND RESULTS

This section analyses the experiments conducted on the proposed OpenAI ChatGPT BERT-BiLSTM system in Figure 1. We describe the parameters, assessment metrics, oversampling, comparisons with various sentiment algorithms, and methods used to interpret the feature vector to provide a clear understanding of the outcome.

A. EXPERIMENTS SETTING

In this paper, the proposed method was conducted on a PC running Ubuntu 20.04 with a 4.0 GHz Intel Core i7 and 64G DDR4 memory. Further, the sci-kit-learn library and Python 3.7 Keras with the TensorFlow 2.0 backend were used to run tests.

B. EVALUATION METRICS

The proposed OpenAI ChatGPT BERT-BiLSTM system in Figure 1 was evaluated using accuracy, F-measure, and AUC. We further measure the performance of the proposed model using evaluation classification models, looking at the number of true negatives divided by the number of predicted positives and false positives, and then construct a confusion matrix using the model's predictions on a held-out test set in Figure 9 and 10. The mathematical formulation is as follows:

$$\text{Accuracy (ACC)} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

In contrast, the F-measure takes into account both precision and recall and gives a balanced estimate of the model's performance. It is computed by dividing twice the product of precision and recall by their sum, as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F - measure = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

AUC (area under the curve) is another commonly used metric for evaluating the performance of classification models. It finds the area under the receiver's operating characteristic curve and gives a single value that shows how well the OpenAI ChatGPT BERT-BiLSTM system shown in Figure 1 differentiates between classes that are positive and classes that are negative. This metric is particularly useful when datasets are imbalanced, such as the one used in this study, or when the cost of false positives and false negatives is not equal. It measures the model's ability to distinguish between positive and negative instances across different probability thresholds. However, random selection or classification of AUC equals 0.5, whereas a perfect classifier will have an AUC equal to 1 using the following equation:

$$\text{AUC} = \int_0^1 \left(\frac{\text{TP}}{P}\right) \text{d}\left(\frac{\text{FP}}{N}\right) \tag{5}$$

where:

$$P = TP + FN \quad \text{(total number of positives)}$$
$$N = TN + FP \quad \text{(total number of negatives)}$$

In other words, the AUC is the integral of the true positive rate (TPR) with respect to the false positive rate (FPR):

$$\text{AUC} = \int_0^1 \text{TPR}\, \text{dFPR} \tag{6}$$

where:

$$\text{TPR} = \frac{\text{TP}}{P}$$
$$\text{FPR} = \frac{\text{FP}}{N}$$

Another equivalent expression for the AUC, emphasizing the integration with respect to the false positive counts, is:

$$\text{AUC} = \frac{1}{P \times N} \int_0^N \text{TP}\, \text{dFP} \tag{7}$$

This integral represents the area under the ROC curve, indicating the performance of a binary classifier.

C. EVALUATION

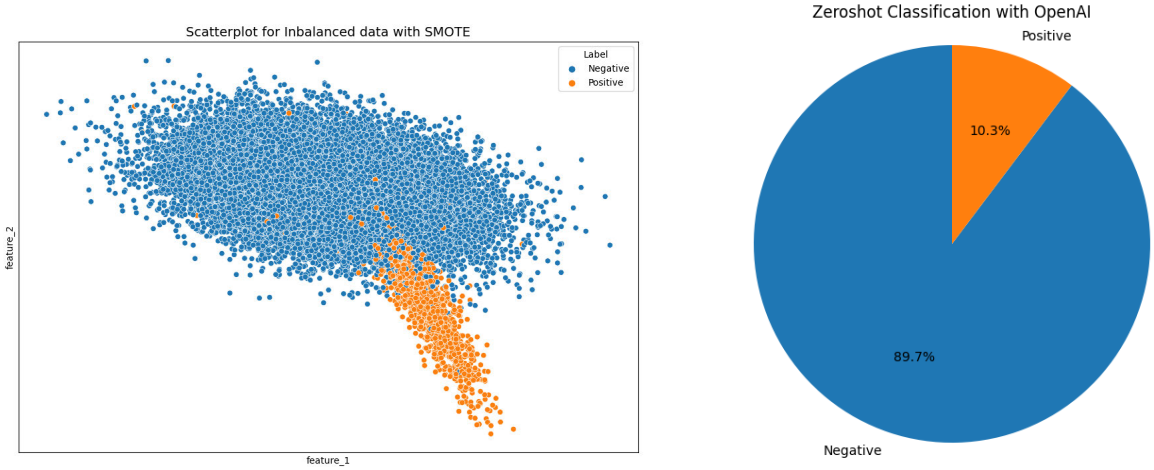
In this section, we report the experimental findings of the suggested model in Figure 1. The datasets obtained from the Hellopeter website are severely unbalanced, as seen in Figure 6. We used the Lexicon-based methods presented in Section IV and ChatGPT to generate features and train ML classifiers to understand and analyze the distribution of each sentence in the text with the sentiment label in Table 2. The representation in Table 3 illustrates a clear connection between how a sentence is put together and the quantifiable inherent subjectivity of the text in natural language. We implement this by integrating Lexicon-based methods like AFINN and SentiWordNet with rule-based algorithms like VADER and the new feature ChatGPT. The dataset is completely unbalanced and biased towards certain classes, and to address this, we employ SMOTE to artificially raise the minority class.

1) SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE (SMOTE)

We employ SMOTE to address the class imbalance in the datasets. SMOTE is a vital component of our research, as it plays a pivotal role in ensuring balanced and unbiased training data for our deep learning models. We then call the negative majority class, while the positive class hereinafter will be called the minority class. In our experiment, we implemented diverse proportional oversampling. Then we do classification and matrices. We implement classification using the sklearn ML libraries built in python[27] [100]. In this paper, we investigate the use of SMOTE methods for resampling the minority class with the majority class. This acts as a key part of making sure that the results of our training data are fair and balanced, which, in consequence, could be used to improve the overall performance and dependability of our SA models. After SMOTE techniques were used on the dataset, Figure 7 (a,b) shows the label distribution along with the data distribution with the class label in the 2-D feature space. Figure 7 (a) provides a visualized diverse resampling proportion and an oversampling method using the SMOTE technique by resampling 100% of the majority class. The graph depicts the successful generation of synthetic samples to balance the class distribution.

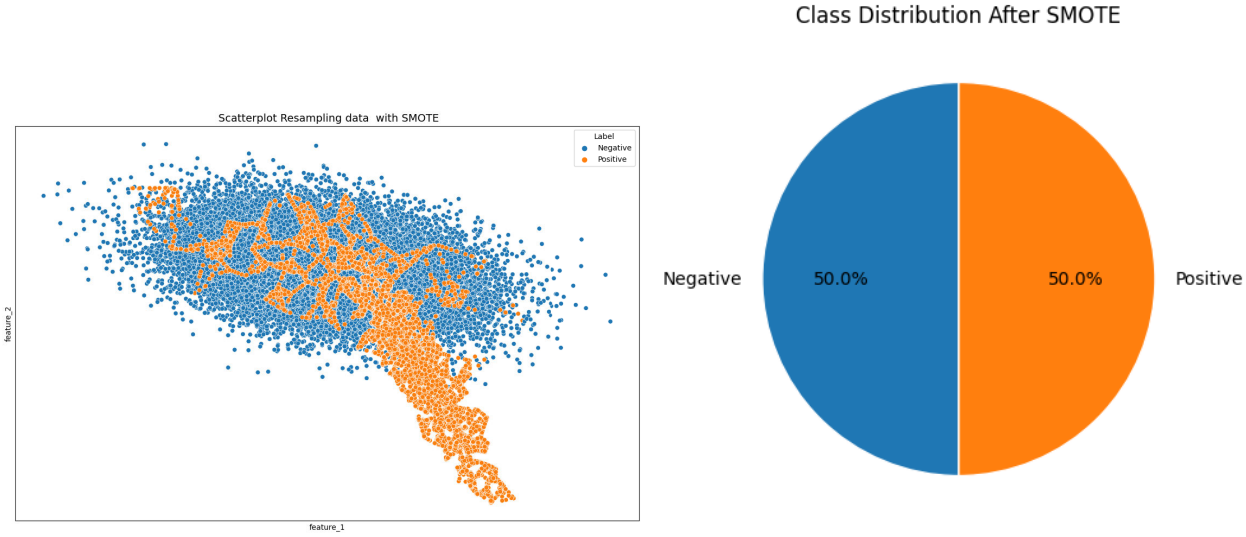
On the newly generated dataset from SMOTE, we went ahead to evaluate the efficacy of the methods in improving the accuracy of our SA models by ensuring the datasets were well-optimized. Since our dataset is a binary classification

[27] https://scikit-learn.org/stable/



(a) 2-D feature without SMOTE (b) ChatGPT OpenAI without SMOTE

FIGURE 6. **Imbalanced class label and features on the dataset from the Hellopeter website.**



(a) 2-D feature with SMOTE (b) ChatGPT OpenAI SMOTE

FIGURE 7. **Distribution of the dataset in 2-D feature space using SMOTE to balance an imbalanced dataset with class labels.**

dataset, we first fit and evaluated it using an RF algorithm. We use the default hyperparameters that come with Sklearn Python libraries, and then we use a repeated stratified k-fold cross-validation (i.e., k = 5 in our case) to evaluate the distribution pattern of the dataset for SA tasks. Hence, Figure 8 shows the performance of metrics using the F-score and recall using the SMOTE techniques in Algorithm 1. In this experiment, we comparatively analyzed the results of the model with and without SMOTE techniques and showed that the recall and F-measure improved significantly with the number of balanced classes. Figure 8 provides the graphic that shows the performance starting from the original data until balanced data with the ChatGPT OpenAI as a feature (e.g., 50:50). The graphic clearly illustrates the positive impact of balancing classes on model performance, with recall values consistent in the last three experiments amounting to 0.845, 0.833, and 0.845, respectively. Similarly, the highest F-measure was in the proportion of the data (10:89), which amounts to 0.853 with a balanced 50:50 distribution but decreases to 0.849. in the proportion of the data (0.624) with an imbalanced class label (90:10).

This is because when the data is evaluated using the SMOTE technique, it balances the ratio of the two classes. Hence, with more training data, the proposed OpenAI ChatGPT BERT-BiLSTM system shown in Figure 1 accuracy in predicting the correct class improves the result in Table 4. The analysis indicates that the improvement in

Algorithm 1 Pseudo-Code of SMOTE Technique

Data: T: Number of minority class samples, N: Amount of SMOTE (percentage), k: Number of nearest neighbors
Result: Synthetic minority class samples
if $N < 100$ **then**
  Randomize the T minority class samples;
  $T \leftarrow \left(\frac{N}{100}\right) \times T$;
  $N \leftarrow 100$;
end
$N \leftarrow$ (int) $\left(\frac{N}{100}\right)$;
numattrs $\leftarrow$ Number of attributes;
Sample[][] $\leftarrow$ Array for original minority class samples;
newindex $\leftarrow 0$;
Synthetic[][] $\leftarrow$ Array for synthetic samples;
for $i \leftarrow 1$ **to** T **do**
  Compute k nearest neighbors for i, and save the indices in *nnarray*;
  Populate(N, i, *nnarray*);
end
Function `Populate`(N, i, *nnarray*) **:**
  while $N \neq 0$ **do**
    Choose a random number between 1 and k, call it *nn*;
    for *attr* $\leftarrow 1$ **to** *numattrs* **do**
      Compute: dif $\leftarrow$ *Sample*[*nnarray*[*nn*]][*attr*] $-$ *Sample*[*i*][*attr*];
      Compute: gap $\leftarrow$ random number between 0 and 1;
      Synthetic[*newindex*][*attr*] $\leftarrow$ *Sample*[*i*][*attr*] $+$ gap $\times$ dif;
    end
    newindex $\leftarrow$ *newindex* $+ 1$;
    $N \leftarrow N - 1$;
  end
return Synthetic;

TABLE 3. A comprehensive summary of the class distributions for the combined dataset.

Model	Total Size	Positive	Negative
AFINN	129,341	34,275	95,066
SentiWordNet	129,341	61,049	68,292
VADER	129,341	54,849	74,492
OpenAI	129,341	13,315	116,026

class balance through SMOTE positively affects the average accuracy of the model, along with metrics like F-measure and recall. The continuous enhancement of various aspects demonstrates the stability and effectiveness of the proposed solution in addressing imbalanced datasets. These results suggest that balancing datasets improves the performance of sentiment analysis algorithms. This study identifies the impact of data imbalance and shows that using SMOTE to balance data significantly boosts the efficiency of sentiment analysis algorithms. Furthermore, it emphasises the need to use suitable evaluation metrics to precisely evaluate the performance of these algorithms.

D. RESULTS

This section provides information on the performance of the proposed model in Figure 1 with and without oversampling strategies, allowing for a thorough review and comparison. Table 4 displays the dataset outcomes in terms of accuracy, F-measure, and AUC.

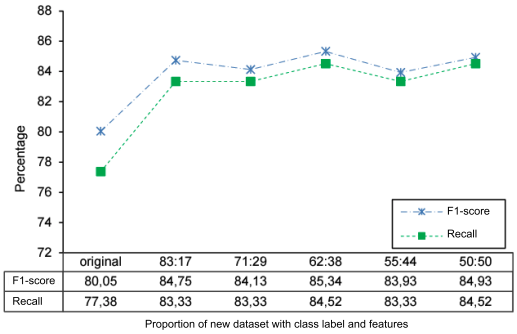


FIGURE 8. Metric performance on balanced dataset with class labels and feature space using SMOTE techniques.

Table 4 first part shows ChatGPT OpenAI used as a feature selection method along with SVM classifiers from customer reviews collected for ABSA on Hellopeter without SMOTE had an accuracy of 0.970 and an AUC of 0.976. This is better than the LR classifier, which had the same accuracy score but a different AUC score of 0.971. If we consider the

TABLE 4. Accuracy, F-measure, and AUC results for the proposed model.

ABSA															
Model/Matrix	KNeighbors			SVM			Logistic Regression			Multinomial NB			Random Forest		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
AFINN	0.651	0.587	0.696	0.659	0.628	0.733	0.0.731	0.618	0.626	0.706	0.589	0.696	0.709	0.588	0.697
SentiWordNet	0.686	0.683	0.502	0.693	0.691	0.508	0.552	0.624	0.695	0.685	0.640	0.645	0.524	0.760	0.694
VADER	0.551	0.541	0.518	0.507	0.502	0.692	0.563	0.615	0.510	0.571	0.643	0.697	0.581	0.628	0.500
ChatGPT OpenAI	0.908	0.871	0.731	0.970	0.873	0.976	0.971	0.912	0.971	0.961	0.871	0.610	0.887	0.877	0.806
ABSA with SMOTE															
Model/Matrix	KNeighbors			SVM			Logistic Regression			Multinomial NB			Random Forest		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
AFINN	0.501	0.734	0.510	0.819	0.817	0.976	0.933	0.932	0.747	0.754	0.640	0.900	0.650	0.737	0.696
SentiWordNet	0.714	0.767	0.738	0.825	0.923	0.951	0.817	0.808	0.849	0.826	0.722	0.715	0.812	0.611	0.614
VADER	0.703	0.840	0.914	0.909	0.906	0.733	0.864	0.958	0.818	0.921	0.760	0.885	0.966	0.794	0.758
ChatGPT OpenAI	0.502	0.737	0.527	0.956	0.956	0.939	0.949	0.723	0.821	0.920	0.920	0.993	0.916	0.915	0.967
Capitec															
Model/Matrix	KNeighbors			SVM			Logistic Regression			Multinomial NB			Random Forest		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
AFINN	0.681	0.640	0.570	0.674	0.620	0.556	0.695	0.571	0.512	0.727	0.612	0.521	0.714	0.595	0.699
SentiWordNet	0.678	0.675	0.673	0.693	0.693	0.546	0.567	0.547	0.529	0.535	0.683	0.503	0.518	0.754	0.688
VADER	0.525	0.505	0.688	0.507	0.699	0.602	0.559	0.609	0.695	0.582	0.647	0.687	0.581	0.627	0.525
ChatGPT OpenAI	0.911	0.896	0.928	0.981	0.930	0.906	0.974	0.924	0.922	0.962	0.561	0.798	0.889	0.777	0.608
Capitec with SMOTE															
Model/Matrix	KNeighbors			SVM			Logistic Regression			Multinomial NB			Random Forest		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
AFINN	0.602	0.519	0.525	0.826	0.824	0.932	0.855	0.849	0.931	0.792	0.787	0.711	0.659	0.650	0.531
SentiWordNet	0.501	0.735	0.545	0.520	0.519	0.576	0.514	0.698	0.638	0.584	0.582	0.507	0.503	0.502	0.525
VADER	0.693	0.672	0.525	0.615	0.612	0.672	0.613	0.602	0.720	0.642	0.638	0.597	0.568	0.562	0.931
ChatGPT OpenAI	0.501	0.736	0.531	0.974	0.964	0.992	0.920	0.910	0.807	0.910	0.727	0.845	0.901	0.835	0.876
Standard															
Model/Matrix	KNeighbors			SVM			Logistic Regression			Multinomial NB			Random Forest		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
AFINN	0.628	0.600	0.698	0.661	0.620	0.667	0.733	0.621	0.693	0.708	0.588	0.504	0.709	0.588	0.547
SentiWordNet	0.504	0.502	0.698	0.507	0.506	0.634	0.532	0.787	0.530	0.544	0.513	0.692	0.524	0.760	0.660
VADER	0.693	0.684	0.660	0.694	0.689	0.641	0.559	0.612	0.688	0.569	0.648	0.504	0.581	0.628	0.511
ChatGPT OpenAI	0.915	0.877	0.547	0.970	0.874	0.694	0.975	0.918	0.665	0.962	0.871	0.698	0.887	0.877	0.698
Standard with SMOTE															
Model/Matrix	KNeighbors			SVM			Logistic Regression			Multinomial NB			Random Forest		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
AFINN	0.543	0.657	0.511	0.811	0.809	0.974	0.910	0.909	0.899	0.764	0.759	0.706	0.652	0.645	0.530
SentiWordNet	0.514	0.764	0.541	0.543	0.541	0.667	0.600	0.592	0.627	0.595	0.592	0.508	0.507	0.506	0.683
VADER	0.595	0.647	0.683	0.602	0.593	0.707	0.657	0.598	0.694	0.620	0.647	0.594	0.568	0.616	0.561
ChatGPT OpenAI	0.522	0.735	0.530	0.967	0.957	0.992	0.951	0.992	0.809	0.921	0.910	0.910	0.913	0.730	0.764
Nedbank															
Model/Matrix	KNeighbors			SVM			Logistic Regression			Multinomial NB			Random Forest		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
AFINN	0.657	0.611	0.532	0.654	0.605	0.678	0.708	0.587	0.654	0.714	0.594	0.511	0.709	0.588	0.630
SentiWordNet	0.515	0.509	0.696	0.697	0.696	0.509	0.515	0.505	0.695	0.505	0.667	0.694	0.517	0.755	0.678
VADER	0.501	0.693	0.678	0.523	0.518	0.611	0.548	0.605	0.511	0.587	0.655	0.681	0.580	0.626	0.506
ChatGPT OpenAI	0.907	0.868	0.630	0.982	0.873	0.740	0.977	0.921	0.635	0.969	0.865	0.676	0.887	0.878	0.506
Nedbank with SMOTE															
Model/Matrix	KNeighbors			SVM			Logistic Regression			Multinomial NB			Random Forest		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
AFINN	0.593	0.652	0.500	0.821	0.820	0.877	0.882	0.877	0.970	0.739	0.729	0.916	0.739	0.729	0.916
SentiWordNet	0.696	0.733	0.522	0.964	0.964	0.715	0.908	0.907	0.996	0.744	0.740	0.824	0.558	0.652	0.601
VADER	0.500	0.518	0.580	0.624	0.622	0.744	0.607	0.505	0.689	0.633	0.626	0.731	0.576	0.568	0.606
ChatGPT OpenAI	0.525	0.905	0.922	0.974	0.964	0.739	0.956	0.876	0.744	0.928	0.907	0.996	0.904	0.740	0.824
FNB															
Model/Matrix	KNeighbors			SVM			Logistic Regression			Multinomial NB			Random Forest		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
AFINN	0.653	0.598	0.692	0.649	0.605	0.723	0.735	0.623	0.503	0.709	0.589	0.538	0.709	0.588	0.699
SentiWordNet	0.546	0.542	0.553	0.685	0.685	0.613	0.542	0.611	0.671	0.516	0.673	0.503	0.524	0.760	0.697
VADER	0.560	0.552	0.535	0.511	0.506	0.561	0.547	0.601	0.618	0.567	0.640	0.672	0.581	0.628	0.509
ChatGPT OpenAI	0.914	0.874	0.694	0.972	0.874	0.511	0.974	0.917	0.613	0.968	0.872	0.673	0.887	0.877	0.505
FNB with SMOTE															
Model/Matrix	KNeighbors			SVM			Logistic Regression			Multinomial NB			Random Forest		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
AFINN	0.876	0.737	0.516	0.818	0.817	0.648	0.917	0.915	0.921	0.778	0.773	0.918	0.651	0.651	0.641
SentiWordNet	0.538	0.753	0.529	0.526	0.525	0.917	0.603	0.584	0.648	0.567	0.564	0.604	0.508	0.508	0.514
VADER	0.697	0.734	0.693	0.615	0.612	0.918	0.602	0.595	0.694	0.607	0.603	0.677	0.570	0.563	0.606
ChatGPT OpenAI	0.505	0.745	0.534	0.994	0.954	0.729	0.987	0.917	0.993	0.913	0.729	0.805	0.905	0.999	0.985

TABLE 5. Accuracy, F-measure, and AUC results for the proposed model with cross-validation.

Model	ABSA			CAP			STD			NED			FNB		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
ChatGPT OpenAI+BERT-BiLSTM (S[28])	0.884	0.869	0.890	0.885	0.858	0.883	0.904	0.897	0.887	0.922	0.878	0.880	0.931	0.928	0.923
ChatGPT OpenAI+BERT-BiLSTM (NSS[29])	0.505	0.497	0.531	0.660	0.676	0.485	0.788	0.716	0.514	0.833	0.831	0.723	0.787	0.787	0.546

second part in Table 4, the SVM model had an accuracy of 0.956 and an AUC of 0.940. This suggests that the SVM model is not as effective as the RF classifier, with an AUC of 0.970 and an accuracy of 0.920 with SMOTE techniques. This shows there is a consistency of 0.5 in the AUC score for imbalanced datasets and a variance of less than 1 in the AUC over accuracy for unbalanced datasets.

Table 4 second part shows that the SVM model had the best accuracy, F1, and AUC scores for the Capitec dataset from Hellopeter without SMOTE. These scores were 0.981 for accuracy, 0.93 for F1, and 0.906 for AUC. The SVM model also outperformed the RF in terms of accuracy, F1, and AUC, while being less superior to LR in terms of accuracy and AUC. This is also a similar situation with SMOTE techniques over the Capitec dataset in terms of accuracy, F1, and AUC, which were 0.97, 0.96, and 0.99, respectively. However, the SVM model had a higher F1 score than the LR or RF models.

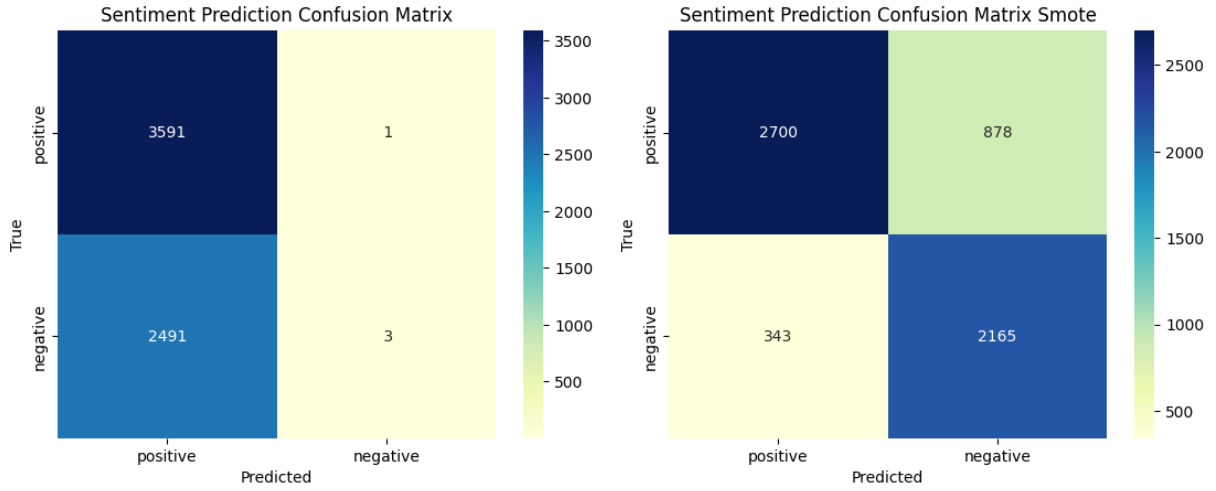
Further, the SVM classifier with the SMOTE technique performed better than LR, K-nearest neighbor, multinomial NB, and RF for the standard dataset (see Table 4) in terms of accuracy, F1, and AUC with 0.967, 0.957, and 0.99, respectively. Whereas the ChatGPT achieves an accuracy of 0.970 for each sentence in the standard dataset, it is classified without SMOTE. In terms of accuracy, F1, and AUC, ChatGPT OpenAI with LR got 0.975, 0.92, and 0.665. On the other hand, AFINN, SentiWordNet, and VADER achieved an accuracy of 0.810, 0.543, and 0.602 using SVM classifiers when using the SMOTE technique on the standard dataset. This highlights the effectiveness of ChatGPT Open AI for sentences containing emoji symbols because they were never removed from the dataset with SVM classifiers and SMOTE for handling unbalanced datasets.

The SVM classifier got scores of 0.974 for accuracy, 0.964 for F1, and 0.738 for AUC with SMOTE for the Nedbank dataset. This demonstrated that ChatGPT was the most effective. ChatGPT obtained 0.981, 0.873, and 0.740 accuracy, F1, and AUC without using SMOTE methods. On the other hand, AFINN, SentiWordNet, and VADER achieved an accuracy of 0.821, 0.964, and 0.624 using SVM classifiers when using the SMOTE technique on the Nedback dataset. Comparatively, K-nearest neighbour, LR, multinomial NB, and RF achieved an accuracy rating of 0.594, 0.882, 0.738, and 0.738 for the AFINN model, and the results consistently remain unsatisfactory to classifier sentiment text for the rest of the standard ML model (see Table 4).

[28]SMOTE.
[29]NON-SMOTE.

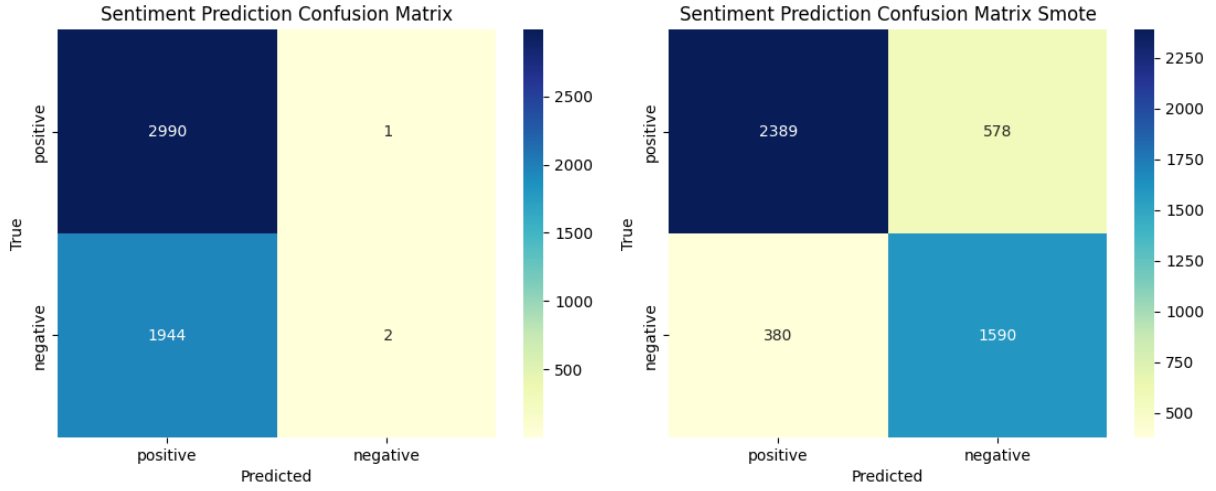
The SVM classifier using the SMOTE technique ranks the best on the FNB dataset, with scores of 0.994 for accuracy and 0.954 for F1, and is less superior with an AUC of 0.730 when compared with an AUC of 0.993 for LR. RF, multinomial NB, SVM, and K-nearest neighbour are placed in the third, fourth, and fifth spots of 0.985, 0.804, 0.730, and 0.534 using ChatGPT, respectively. As for F1, RF obtained the best results of 0.998, followed by SVM with 0.954, and LR came in third place with 0.916, K-nearest neighbour, and multinomial NB with 0.744 and 0.730 fourth and fifth, respectively. Meanwhile, LR obtained the best F1 score without SMOTE, yielding values of 0.917; RF ranked second with 0.876; SVM ranked third with 0.874; and multinomial NB and K-nearest neighbour shared the fourth and fifth positions with 0.872 and 0.874, respectively. AFINN has the highest accuracy for LR, which yields 0.917 with the SMOTE technique, followed by K-nearest neighbour with 0.876, SVM with 0.818, multinomial NB with 0.778, and lastly, RF yield 0.651. At the same time, SentiWordNet with K-nearest neighbour has the highest value of 0.697 for VADER, followed by SVM with 0.612 when the SMOTE technique is used. The SVM classifier that used the FNB dataset with SMOTE and ChatGPT had the best overall accuracy (0.994), but its AUC values were lower than those of the ABSA dataset (0.975), as shown in Table 4. Compared with the ML model using SMOTE, AFINN outperforms the existing Lexicon-based methods model in terms of accuracy, with 0.876 for the K-nearest neighbour and multinomial NB with 0.996, which is superior to all the other classifiers for AUC.

Additionally, we improve the reliability of the proposed model, e.g., ChatGPT OpenAI for text annotation tools and BERT-BiLSTM to improve SA tasks. We implement cross-validation to fit the model and estimate the prediction accuracy of the unseen sets of data drawn from the training samples. In doing this, we partition the data into two independent sets, one for training and one for testing, and then estimate the accuracy of the events and tasks. The first set is used to train the model, while the subsequent set of data (the test) is used to evaluate its performance using a 10-fold cross-validation (CV) [101]. Cross-validation is a kind of resampling process that assesses the model and trains it roundly many times using test data from samples drawn from the original. In this paper, we used 5-fold stratified cross-validation with shuffling and a random seed for reproducibility. The results are shown in Table 5, which provide valuable information into the proposed model's Figure in 1 performance and accuracy. This process enables us to measure the model's generalizability to new data and



(a) ABSA without SMOTE (b) ABSA SMOTE

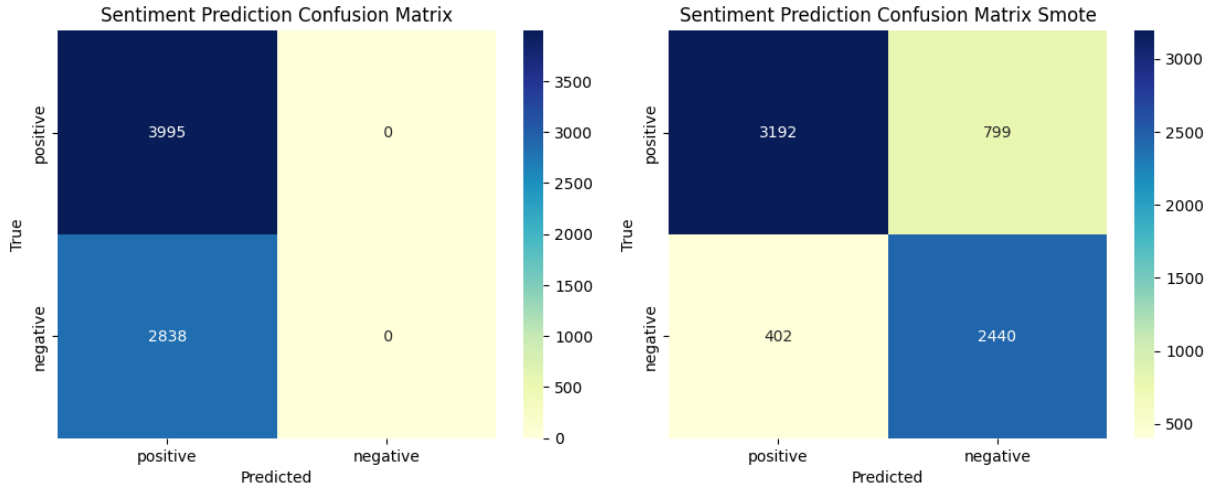
FIGURE 9. Confusion matrix performance for ABSA dataset.



(a) CPT without SMOTE (b) CPT SMOTE

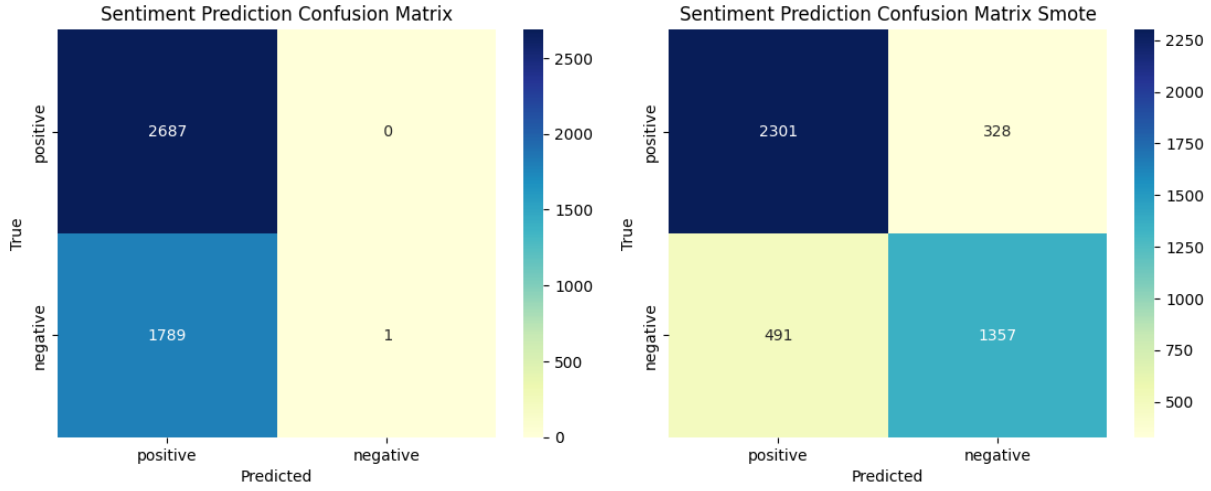
FIGURE 10. Confusion matrix performance for CPT dataset.

identify any potential overfitting issues, which ensures that the proposed model is robustly constructed and reliable in predicting sentiment in unseen text data. Our test shows that ChatGPT OpenAI with the BERT-BiLSTM model and SMOTE technique is more accurate than the same model without the SMOTE technique. It reached scores of 0.932 for the FNB dataset, 0.922 for NED, 0.902 for STD, 0.885 for CAP, and 0.884 for ABSA. The results highlight the effectiveness of using SMOTE techniques in improving model performance on imbalanced text datasets. In contrast, the result in terms of the AUC score also indicated that the FNB dataset outperformed the other datasets with an AUC score of 0.923, while the ABSA dataset ranked second with an AUC score of 0.890, and the STD, CAP, and NED datasets came in third, fourth, and fifth, respectively, with an AUC score of 0.887, 0.883, and 0.880. Whereas when the ChapGPT OpenAI BERT-BiLSTM model was used without SMOTE technique, it did not perform at all as planned. For the ABSA, CAP, STD, FNB, and NED datasets, the accuracy and AUC scores were 0.505, 0.659, 0.786, 0.788, and 0.88, respectively. We add SoftMax as an activation function in the layer of the proposed OpenAI ChatGPT BERT-BiLSTM system in Figure 1 to normalise the raw model outputs to ensure that the probability scores for each sentiment polarity sum up to one. This helps the OpenAI ChatGPT BERT-BiLSTM system shown in Figure 1 sort sentiments into groups based on their orientation. The high accuracy and AUC scores in the evaluation section show how this approach helps to accurately classify sentiment orientations across datasets. This demonstrates the effectiveness of our model



(a) STD without SMOTE (b) STD SMOTE

FIGURE 11. **Confusion matrix performance for STD dataset.**



(a) NED without SMOTE (b) NED SMOTE

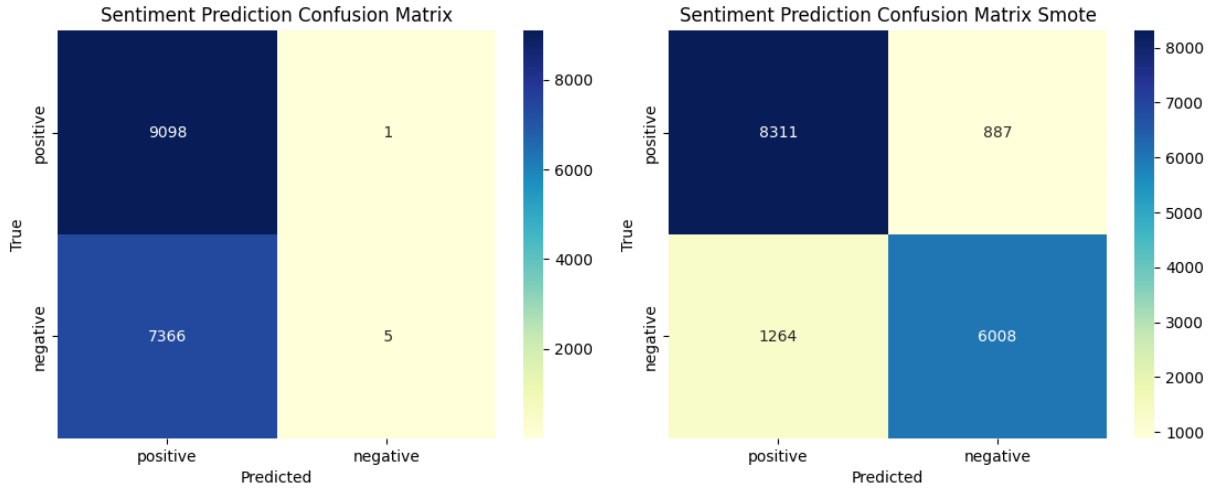
FIGURE 12. **Confusion matrix performance for NED dataset.**

in predicting sentiment in unseen text data. This technique enables our model to offer probabilities that can be readily interpreted as confidence levels for each sentiment category.

To summarise, the proposed model, ChatGPT OpenAI BERT-BiLSTM with the SMOTE technique, outperformed the standard models using zero-shot learning in terms of accuracy and AUC for the FNB dataset, followed by NED, STD, CAP, and ABSA when precision is taken into account. This highlights the value of using pre-trained language models to enhance SA tasks. With an accuracy of 0.94 on the NED, only SentiwordNet SVM plus SMOTE does better than regular machine learning models that use Lexicon-based methods or rule-based algorithms. VADER, combined with LR and SMOTE, comes in second with 0.957 on the ABSA dataset. ChatGPT as a text annotation from a pre-trained language model performs better for SA tasks and offers a user-friendly interactive interface. VADER, however, is more accurate for aspect-based SA tasks. VADER, however, is more accurate for SA tasks, achieving an F1 score of 0.699 on the CAP dataset.

E. VISUALISED DESIGN

In this section, we make it simple for non-expert users to understand and be able to interpret the information extracted from the OpenAI ChatGPT BERT-BiLSTM system in Figure 1 for sentiment analysis tasks on the Hellopeter website and various customer reviews comments on the social media platform or Internet related to the product or service. We experimented with all the traditional methods, with or without SMOTE, and compared them with the proposed



(a) FNB without SMOTE

(b) FNB SMOTE

FIGURE 13. Confusion matrix performance for FNB dataset.

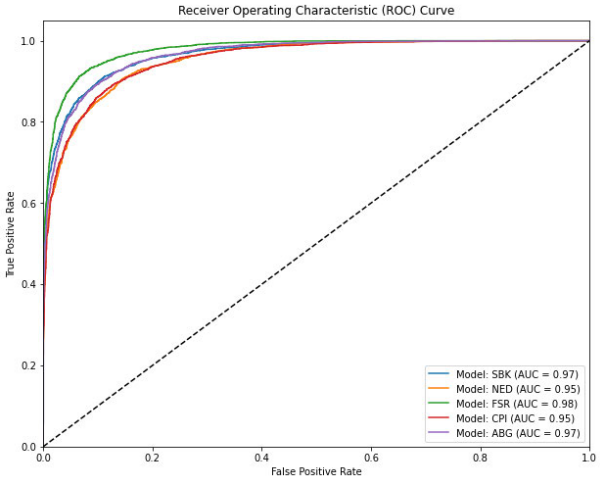


FIGURE 14. Area under the ROC Curve with SMOTE techniques on the datasets.

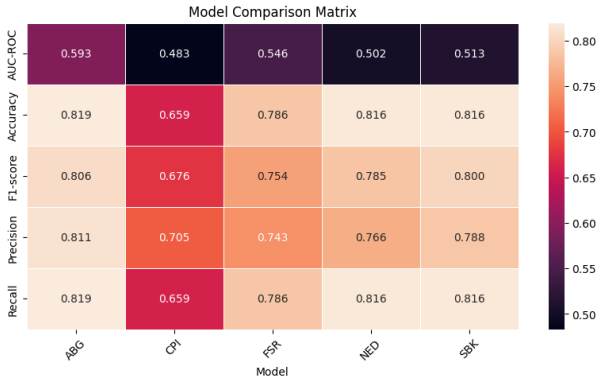


FIGURE 15. Proposed model comparison matrix on the datasets.

model. The proposed model outperformed all other methods in terms of accuracy and efficiency.

1) CONFUSION MATRIX

We look at the confusion matrix (CM) to understand and analyze the number of correctly and incorrectly classified (e.g., discrepancies) subjects in each sentence in the dataset with SMOTE and non-STOME techniques to improve the prediction of the proposed model for SA tasks. The Figures from 9–13 show the correct and incorrect confusion matrix. This shows how many mistakes there were compared to the true positive predictions for each sentence class across all datasets. These figures (from 9–13) detail inter-class confusion that indicates the 2 classes (positive and negative) that produce most FP and FN and group errors for all other classes by the true class of the sentence. It also indicates the potential biases in the dataset, e.g., high confusion between classes yielding correlated but unrepresentative subject pairs in the results. Therefore, such information or datasets lost with the omission of the true class label can be critical for downstream applications. As demonstrated in the figures from 9–13, the majority of the predictions end up on the diagonal (predicted class = actual class), which is what we want, but the discrepancy without SMOTE significantly impacts the performance of MLA. Figure 14 shows the average AUC for ChatGPT OpenAI BERT-BiLSTM using SMOTE approaches across all datasets. Notably, all of the datasets performed exceptionally well with the proposed model, but the FNB dataset had the highest AUC at 98%. Similarly, Figure 15 provides us with a summary result of the ChatGPT OpenAI BERT-BiLSTM model with STOME technique in terms of recall, precision, F-measure, accuracy, and ROC AUC for all the datasets based on the customer reviews.

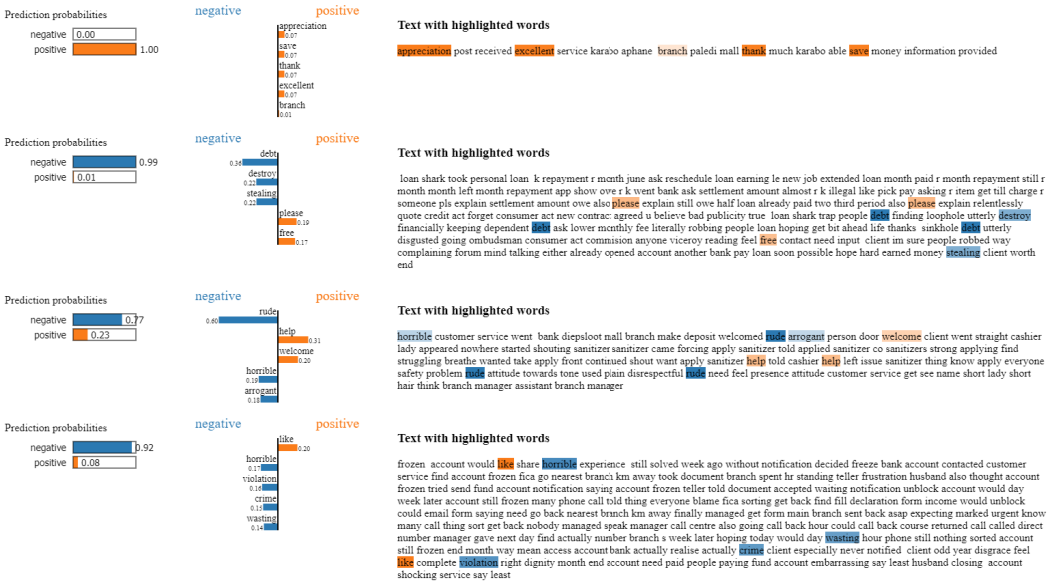


FIGURE 16. LIME model interpretability on FNB customer review based on prediction.

F. INTERPRETATION

This section discusses the results and their utility in demonstrating that the feature vectors obtained from the proposed method can assist trained professionals in understanding and analyzing the inherent subjectivity in customer reviews to improve decision-making for SA tasks. We used two different approaches to explain the results: Shapley additive explanations (SHAP), which is a game-theoretic way to explain the output of any machine learning model [102] and local interpretable model-agnostic explanations (LIME), an agnostic framework for interpretability [103].

Figure 16 shows the predictions from randomly selected sampled documents in the FNB datasets. As shown in Figure 16, the proposed method puts more weight on the positive label, with a significant prediction probability on words such as appreciation at 7%, save at 7%, and excellent at 7%, with a phrase such as ''appreciation post received excellent service karabo aphane branch paled mall thank much karabo able to save money information provided'' This shows that 100% of the text in this phrase is coming from a positive class label. Equally, with another phrase within the same FNB dataset, we have a significant prediction probability on words such as debt at 36%, destroy at 22%, and steal at 22%, and within the same phrase, words like please and free are labeled as positive with 19% and 17% predicted within the sample of the text, while the entire perception of the phrase reaches 99% negative and 1% positive, respectively.

Another scenario shows the proposed model through LIME highlights words such as ''rude'' at 60% compared to the words ''horrible'' and ''arrogant'' at 19% and 18% (as a feature now) to classify the text sample as 77% negative and 23% positive sentiment, which is the confidence the proposed model has in the text. As figure 16 demonstrated, it provides non-experts with a great way of elucidating what is going on when training the ChatGPT Open BERT-BiLSTM with SMOTE techniques on FNB datasets for non-technical folks rather than just being a black box that can not be interpreted. Additionally, we also applied SHAP [103] to the feature vectors to get a good estimate of the SHAP values for the extracted features. The permutation was accomplished by turning one word on and off. For example, if a word is missing from a particular post, enabling it could make a significant difference in revealing the inherent subjectivity of each sentence in the customer review in natural language, improving interpretation and explainability to non-experts in the field about what the proposed model is learning for effective sentiment classification of a given text. The document's nonexistent feature could obtain a high SHAP score. Before SoftMax, the collected features were subject to permutations to avoid this. By doing so, the proposed model was able to better understand the importance of each feature in the sentiment classification process. This approach helped ensure that the model's predictions were based on meaningful and relevant information.

Figure 17 depicts the first random sample from the FNB databases. The SHAP scores for the contributing characteristics picked during the training set of the proposed model are also displayed. The features that push the prediction higher are shown in red, representing the corrected prediction by the model used in predicting and classifying a sample text as positive, and those that push the prediction lower (predicting the incorrect as negative) are shown in blue, along with the confidence level. The SHAP local explanation takes only one occurrence at a time and creates an explanation by indicating which feature values make decisions about the position and

(a) FNB (b) FNB

FIGURE 17. SHAP plot for local explanations of an instance on the FNB dataset.

which are negative. Figure 17 demonstrates the robustness of using SHAP to explain the complexity of the proposed model with the customer reviews dataset to understand and analyse the subject in each sentence and make it explainable. We observe in Figure 17 that the probability of the proposed model confidently predicting the sample text based on the characteristics highlighted in red is 48.22% as a positive sentiment, which includes words like ''take care'', loyalty'', and the likelihood of using features in blue as negative is 1.44% using the words like ''pay debt'', ''including debt'', and ''get punished kid'' using the FNB dataset, where based probability confident of using words like ''assign incident'', ''frustrated'' are tailored towards a negative in blue at 0.12% and positive sentiment at 8.61%. These wholeheartedly correlate to the same explainable interpretation using LIME in Figure 16.

We will also look into how the SHAP model can be used to enumerate explainable components of specific sentences from FNB customer review documents. The idea here is that SHAP takes the input features, sums up the difference between the baseline (expected or value) model output, and uses the results as the current model output. The easiest way to implement this in SHAP is to use a waterfall plot that starts with our background expectations of the model as a function to find the inherent subjective value of a given sentence as $\mathbb{E}[f(X)]$, and then you can start adding features one at a time until we reach the current model output.

VII. CONCLUSION

This study has explored the dynamic interplay between customer sentiment, financial performance, and the South African economy, with a particular focus on the country's top financial institutions listed on the JSE. The investigation reveals essential insights into the intricate relationship between customer reviews on Hellopeter, the financial performance of these institutions, and their broader economic implications. Our research establishes a notable correlation between customer sentiment, as reflected in HelloPeter reviews, and the financial performance of South Africa's major financial institutions. Specifically, we find a substantial relationship between customer sentiment and the total revenues of these institutions. This linkage underscores the significance of customer satisfaction in driving financial success in the South African financial sector. Through rigorous analysis, this study highlights the potential of advanced SA models, including BERT, LSTM, SVM, and LR. These models exhibit remarkable accuracy in predicting customer sentiment and review scores based on Hellopeter data. This predictive capability empowers financial institutions to proactively address customer concerns and enhance satisfaction. Customer sentiment plays a pivotal role in shaping the South African economy. Changes in sentiment are associated with noticeable fluctuations in stock prices, market capitalization, and other market-related indicators. These findings are of particular interest to policymakers and regulatory bodies, like the South African Reserve Bank (SARB), in their pursuit of economic stability. This research carries significant implications for the South African financial sector, policymakers, and regulatory authorities. In future work, we intend to expand our research to encompass a broader spectrum of financial institutions listed on the JSE, focusing on the top 30 institutions. Furthermore, we aim to delve deeper into the comparative analysis of the South African Reserve Bank (SARB) against other central banks worldwide. This expansion will enable us to predict financial metric changes such as interest rates, the Consumer Price Index (CPI), and more. Additionally, we intend to use restaurant and movie review datasets with our proposed model to evaluate the efficiency of the model with the publicly available dataset. This expansion will enable us to predict financial metric changes such as interest rates, the Consumer Price Index (CPI), and more. The study is not without limitations. First, financial data was only available for a limited period of five years, potentially limiting the depth of our analysis. A more extended dataset would provide a broader perspective. Additionally, data collection constraints prevented us from obtaining WhatsApp reviews due to personal information protection laws. Furthermore, we were unable to access the total number of customers for each financial institution over the five years. The lack of WhatsApp reviews and total customer headcounts can be considered a limitation of the research. In future research, we will consider the legal means of accessing WhatsApp data and attempt to obtain JSE data for a period longer than five years. Also, obtaining extensive customer headcount details from financial institutions will be essential. Ensuring the collection of more complete and extensive datasets will enhance the robustness and generalizability of future studies. Moreover, the study's SA models, including BERT, LSTM, SVM, and LR, while effective, may have limitations in certain contexts. These models' performance could vary based on the specific nature of customer reviews and language nuances, necessitating careful consideration in real-world applications. This research provides a comprehensive understanding of the relationship between customer sentiment, financial performance, and the broader economic landscape in South Africa.

By harnessing the power of machine learning and SA, financial institutions, regulatory authorities, and policymakers can navigate the complex dynamics of customer satisfaction and its far-reaching implications effectively. In future work, we plan to explore how the proposed method can be used for sentiment analysis tasks. This includes analyzing user reviews from social media related to restaurants and movies to assess the framework's functionality using a publicly available dataset. Finally, we will also explore how this method can be applied in industries outside the realm of restaurants and movies.

ACKNOWLEDGMENT

The authors would like to thank the ABSA (through their Chair in Data Science) for their steadfast support and funding of the Data Science for Social Impact (DSFSI) Research Group at the Department of Computer Science. We also acknowledge the support of Google, Google Tensorflow and OpenAI whose support made this research possible.

REFERENCES

[1] R. Murphy. (May 12, 2024). *Local Consumer Review Survey 2020*. [Online]. Available: https://www.brightlocal.com/research/local-consumer-review-survey-2020/

[2] R. Obiedat, R. Qaddoura, A. M. Al-Zoubi, L. Al-Qaisi, O. Harfoushi, M. Alrefai, and H. Faris, "Sentiment analysis of customers' reviews using a hybrid evolutionary SVM-based approach in an imbalanced data distribution," *IEEE Access*, vol. 10, pp. 22260–22273, 2022.

[3] A. Alpaslan, "Hellopeter! On: How are they doing?" *Perceptions, Experiences Recommendations Social Work Students Customers Open Distance Learn. Univ., Unpublished Inaugural Lecture. Pretoria, Univ. South Africa*, 2012. Accessed: Jul. 15, 2024. [Online]. Available: https://uir.unisa.ac.za/bitstream/handle/10500/6061/Inaugural%20Lecture%20%20-%20Prof%20AH%20Alpaslan%20%28Department%20of%20Social%20Work%29%2018%20July%202012%20-%20Final%20edited%20version.pdf?sequence=1&isAllowed=y

[4] S. Mahmoud and A. El-Masry, "Exploring the evolving landscape of social media marketing: Opportunities and limitations in the digital age," *J. Intell. Connectivity Emerg. Technol.*, vol. 8, no. 1, pp. 1–15, 2023.

[5] A. Modupe, T. Celik, V. Marivate, and O. Olugbara, "Post-authorship attribution using regularized deep neural network," *Appl. Sci.*, vol. 12, no. 15, p. 7518, Jul. 2022.

[6] E. Asani, H. Vahdat-Nejad, and J. Sadri, "Restaurant recommender system based on sentiment analysis," *Mach. Learn. Appl.*, vol. 6, Dec. 2021, Art. no. 100114.

[7] A. M. Rajeswari, M. Mahalakshmi, R. Nithyashree, and G. Nalini, "Sentiment analysis for predicting customer reviews using a hybrid approach," in *Proc. Adv. Comput. Commun. Technol. High Perform. Appl. (ACCTHPA)*, Jul. 2020, pp. 200–205.

[8] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. 7th Int. Conf. Lang. Resour. Eval.*, 2010, pp. 2200–2204.

[9] A. Cernian, V. Sgarciu, and B. Martin, "Sentiment analysis from product reviews using SentiWordNet as lexical resource," in *Proc. 7th Int. Conf. Electron., Comput. Artif. Intell. (ECAI)*, Jun. 2015, pp. WE-15–WE-18.

[10] N. Medagoda, S. Shanmuganathan, and J. Whalley, "Sentiment lexicon construction using SentiWordNet 3.0," in *Proc. 11th Int. Conf. Natural Comput. (ICNC)*, Aug. 2015, pp. 802–807.

[11] A. Park, "Enhancing health information-gathering experiences in online health communities," Ph.D. dissertation, Dept. Biomed. Inform. Med. Educ., Univ. Washington, Seattle, WA, USA, 2015.

[12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2022, pp. 27730–27744.

[13] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 8, May 2014, pp. 216–225.

[14] S. Vosoughi, P. Vijayaraghavan, and D. Roy, "Tweet2 Vec: Learning tweet embeddings using character-level CNN-LSTM encoder–decoder," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2016, pp. 1041–1044.

[15] S. H. H. Ding, B. C. M. Fung, F. Iqbal, and W. K. Cheung, "Learning stylometric representations for authorship analysis," *IEEE Trans. Cybern.*, vol. 49, no. 1, pp. 107–121, Jan. 2019.

[16] S. Al-Natour and O. Turetken, "A comparative assessment of sentiment analysis and star ratings for consumer reviews," *Int. J. Inf. Manage.*, vol. 54, Oct. 2020, Art. no. 102132.

[17] D. Difallah, E. Filatova, and P. Ipeirotis, "Demographics and dynamics of mechanical Turk workers," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, Feb. 2018, pp. 135–143.

[18] X.-R. Gong, J.-X. Jin, and T. Zhang, "Sentiment analysis using autoregressive language modeling and broad learning system," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 1130–1134.

[19] M. Belal, J. She, and S. Wong, "Leveraging ChatGPT as text annotation tool for sentiment analysis," 2023, *arXiv:2306.17177*.

[20] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," *Proc. Nat. Acad. Sci. USA*, vol. 120, no. 30, Jul. 2023, Art. no. e230501612.

[21] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is chatgpt a general-purpose natural language processing task solver?" in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 1–46.

[22] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," 2016, *arXiv:1605.08900*.

[23] S. Tripathi, R. Mehrotra, V. Bansal, and S. Upadhyay, "Analyzing sentiment using IMDb dataset," in *Proc. 12th Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Sep. 2020, pp. 30–33.

[24] M. Jooshaki, A. Nad, and S. Michaux, "A systematic review on the application of machine learning in exploiting mineralogical data in mining and mineral industry," *Minerals*, vol. 11, no. 8, p. 816, Jul. 2021.

[25] S.-A. Bahrainian and A. Dengel, "Sentiment analysis using sentiment features," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI) Intell. Agent Technol. (IAT)*, vol. 3, Nov. 2013, pp. 26–29.

[26] K. V. Raju and M. Sridhar, "Based sentiment prediction of rating using natural language processing sentence-level sentiment analysis with bag-of-words approach," in *Proc. 1st Int. Conf. Sustain. Technol. Comput. Intell. (ICTSCI)*. Cham, Switzerland: Springer, 2020, pp. 807–821.

[27] Y. S. Mehanna and M. B. Mahmuddin, "A semantic conceptualization using tagged bag-of-concepts for sentiment analysis," *IEEE Access*, vol. 9, pp. 118736–118756, 2021.

[28] B. S. Rintyarna, R. Sarno, and C. Fatichah, "Evaluating the performance of sentence level features and domain sensitive features of product reviews on supervised sentiment analysis tasks," *J. Big Data*, vol. 6, no. 1, pp. 1–19, Dec. 2019.

[29] M. Alshammari and M. Mezher, "A comparative analysis of data mining techniques on breast cancer diagnosis data using weka toolbox," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, 2020, pp. 224–229, 2020.

[30] S. N. Almuayqil, M. Humayun, N. Z. Jhanjhi, M. F. Almufareh, and D. Javed, "Framework for improved sentiment analysis via random minority oversampling for user tweet review classification," *Electronics*, vol. 11, no. 19, p. 3058, Sep. 2022.

[31] M. V. Valero, "Thousands of scientists are cutting back on Twitter," *Nature*, vol. 620, pp. 4–482, Aug. 2023.

[32] E. Fersini, E. Messina, and F. A. Pozzi, "Sentiment analysis: Bayesian ensemble learning," *Decis. Support Syst.*, vol. 68, pp. 26–38, Dec. 2014.

[33] A. B. Goldberg and X. Zhu, "Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization," in *Proc. 1st Workshop Graph Methods Natural Lang. Process.*, 2006, pp. 45–52.

[34] O. Täckström and R. McDonald, "Semi-supervised latent variable models for sentence-level sentiment analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 569–574.

[35] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification," in *Proc. 45th Annu. Meet. Assoc. Comput. Linguist.*, 2007, pp. 440–447.

[36] R. Campos, S. Canuto, T. Salles, C. C. A. de Sá, and M. A. Gonçalves, "Stacking bagged and boosted forests for effective automated classification," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, vol. 15, Aug. 2017, pp. 105–114.

[37] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.
[38] Murni, T. Handhika, A. Fahrurozi, I. Sari, D. P. Lestari, and R. I. M. Zen, "Hybrid method for sentiment analysis using homogeneous ensemble classifier," in *Proc. 2nd Int. Conf. Comput. Informat. Eng. (IC IE)*, Sep. 2019, pp. 232–236.
[39] A. Alrehili and K. Albalawi, "Sentiment analysis of customer reviews using ensemble method," in *Proc. Int. Conf. Comput. Inf. Sci. (ICCIS)*, Apr. 2019, pp. 1–6.
[40] H. Pouransari and S. Ghili, "Deep learning for sentiment analysis of movie reviews," *CS224N Proj*, pp. 1–8, 2014. [Online]. Available: https://cs224d.stanford.edu/reports/PouransariHadi.pdf
[41] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "AraVec: A set of Arabic word embedding models for use in Arabic NLP," *Proc. Comput. Sci.*, vol. 117, pp. 256–265, Jan. 2017.
[42] M. M. Ashi, M. A. Siddiqui, and F. Nadeem, "Pre-trained word embeddings for Arabic aspect-based sentiment analysis of airline tweets," in *Proc. Int. Conf. Adv. Intell. Syst. Inform.* Cham, Switzerland: Springer, 2019, pp. 241–251.
[43] A. A. Altowayan and A. Elnagar, "Improving Arabic sentiment analysis with sentiment-specific embeddings," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 4314–4320.
[44] H. Liu, "Sentiment analysis of citations using word2vec," 2017, *arXiv:1704.00177*.
[45] J. Acosta, N. Lamaute, M. Luo, E. Finkelstein, and C. Andreea, "Sentiment analysis of Twitter messages using word2vec," in *Proc. Student-Faculty Res. Day, CSIS*, vol. 7, 2017, pp. 1–7.
[46] K. A. Djaballah, K. Boukhalfa, and O. Boussaid, "Sentiment analysis of Twitter messages using word2vec by weighted average," in *Proc. 6th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, Oct. 2019, pp. 223–228.
[47] B. O. Deho, A. W. Agangiba, L. F. Aryeh, and A. J. Ansah, "Sentiment analysis with word embedding," in *Proc. IEEE 7th Int. Conf. Adapt. Sci. Technol. (ICAST)*, Aug. 2018, pp. 1–4.
[48] B. Xue, C. Fu, and Z. Shaobin, "A study on sentiment computing and classification of Sina weibo with word2vec," in *Proc. IEEE Int. Congr. Big Data*, Jun. 2014, pp. 358–363.
[49] T. Joachims, "A support vector method for multivariate performance measures," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 377–384.
[50] D. Zhang, H. Xu, Z. Su, and Y. Xu, "Chinese comments sentiment classification based on word2vec and SVMperf," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 1857–1863, Mar. 2015.
[51] H. Wang and J. A. Castanon, "Sentiment expression via emoticons on social media," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Nov. 2015, pp. 2404–2408.
[52] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of sentiment analysis in finance: From lexicons to transformers," *IEEE Access*, vol. 8, pp. 131662–131682, 2020.
[53] Y. Song, J. Wang, Z. Liang, Z. Liu, and T. Jiang, "Utilizing BERT intermediate layers for aspect based sentiment analysis and natural language inference," 2020, *arXiv:2002.04815*.
[54] M. A. Al-Sharafi, M. Al-Emran, M. Iranmanesh, N. Al-Qaysi, N. A. Iahad, and I. Arpaci, "Understanding the impact of knowledge management factors on the sustainable use of AI-based chatbots for educational purposes using a hybrid SEM-ANN approach," *Interact. Learn. Environ.*, vol. 31, no. 10, pp. 7491–7510, Dec. 2023.
[55] K. I. Roumeliotis and N. D. Tselikas, "ChatGPT and open-AI models: A preliminary review," *Future Internet*, vol. 15, no. 6, p. 192, May 2023.
[56] T. Talan and Y. Kalinkara, "The role of artificial intelligence in higher education: Chatgpt assessment for anatomy course," *Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi*, vol. 7, no. 1, pp. 33–40, 2023.
[57] J. G. Meyer, R. J. Urbanowicz, P. C. N. Martin, K. O'Connor, R. Li, P.-C. Peng, T. J. Bright, N. Tatonetti, K. J. Won, G. Gonzalez-Hernandez, and J. H. Moore, "ChatGPT and large language models in academia: Opportunities and challenges," *BioData Mining*, vol. 16, no. 1, p. 20, Jul. 2023.
[58] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 121–154, Jan. 2023.
[59] A. Korkmaz, C. Aktürk, and T. Talan, "Analyzing the user's sentiments of ChatGPT using Twitter data," *Iraqi J. Comput. Sci. Math.*, vol. 4, no. 2, pp. 202–214, May 2023.
[60] C. A. Gao, F. M. Howard, N. S. Markov, E. C. Dyer, S. Ramesh, Y. Luo, and A. T. Pearson, "Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers," *NPJ Digit. Med.*, vol. 6, no. 1, p. 75, Apr. 2023.
[61] F. Sudirjo, K. Diantoro, J. A. Al-Gasawneh, H. K. Azzaakiyyah, and A. M. A. Ausat, "Application of ChatGPT in improving customer sentiment analysis for businesses," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 5, no. 3, pp. 283–288, Jul. 2023.
[62] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowl. Inf. Syst.*, vol. 60, pp. 617–663, Jul. 2019.
[63] M. Leippold, "Sentiment spin: Attacking financial sentiment with GPT-3," *Finance Res. Lett.*, vol. 55, Jul. 2023, Art. no. 103957.
[64] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022.
[65] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Syst. Appl.*, vol. 223, Aug. 2023, Art. no. 119862.
[66] K. Kheiri and H. Karimi, "SentimentGPT: Exploiting GPT for advanced sentiment analysis and its departure from current machine learning," 2023, *arXiv:2307.10234*.
[67] A. Lopez-Lira and Y. Tang, "Can ChatGPT forecast stock price movements? Return predictability and large language models," 2023, *arXiv:2304.07619*.
[68] S. Y. Yang, S. Y. K. Mo, and A. Liu, "Twitter financial community sentiment and its predictive relationship to stock market movement," *Quant. Finance*, vol. 15, no. 10, pp. 1637–1656, Oct. 2015.
[69] A. Groß-Klußmann, S. König, and M. Ebner, "Buzzwords build momentum: Global financial Twitter sentiment and the aggregate stock market," *Expert Syst. Appl.*, vol. 136, pp. 171–186, Dec. 2019.
[70] H. K. Sul, A. R. Dennis, and L. Yuan, "Trading on Twitter: Using social media sentiment to predict stock returns," *Decis. Sci.*, vol. 48, no. 3, pp. 454–488, Jun. 2017.
[71] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance," *Inf. Sci.*, vol. 505, pp. 32–64, Dec. 2019.
[72] A. A. Altowayan and L. Tao, "Word embeddings for Arabic sentiment analysis," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 3820–3825.
[73] F. Nausheen and S. H. Begum, "Sentiment analysis to predict election results using Python," in *Proc. 2nd Int. Conf. Inventive Syst. Control (ICISC)*, Jan. 2018, pp. 1259–1262.
[74] C. Kaur and A. Sharma, "Social issues sentiment analysis using Python," in *Proc. 5th Int. Conf. Comput., Commun. Secur. (ICCCS)*, Oct. 2020, pp. 1–6.
[75] Y. Heryanto and A. Triayudi, "Evaluating text quality of GPT engine davinci-003 and GPT engine Davinci generation using BLEU score," *J. Technol. Inf. Syst.*, vol. 1, no. 4, pp. 121–129, Dec. 2023.
[76] J. Nay, "Large language models as corporate lobbyists," 2023, *arXiv:2301.01181*.
[77] Z. Nanli, Z. Ping, L. Weiguo, and C. Meng, "Sentiment analysis: A literature review," in *Proc. Int. Symp. Manage. Technol. (ISMOT)*, Nov. 2012, pp. 572–576.
[78] S. Ahuja and G. Dubey, "Clustering and sentiment analysis on Twitter data," in *Proc. 2nd Int. Conf. Telecommun. Netw. (TEL-NET)*, Aug. 2017, pp. 1–5.
[79] A. Esuli and F. Sebastiani, "SENTIWORDNET: A publicly available lexical resource for opinion mining," in *Proc. 5th Int. Conf. Lang. Resour. Eval.*, 2006, pp. 417–422.
[80] P. Bo and L. Lee, "Opinion mining and sentiment analysis foundations and trends in information retrieval," *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, p. 1135, 2008.
[81] F. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," 2011, *arXiv:1103.2903*.
[82] N. Rathee, N. Joshi, and J. Kaur, "Sentiment analysis using machine learning techniques on Python," in *Proc. 2nd Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, Jun. 2018, pp. 779–785.
[83] B. Gupta, M. Negi, K. Vishwakarma, G. Rawat, P. Badhani, and B. Tech, "Study of Twitter sentiment analysis using machine learning algorithms on Python," *Int. J. Adv. Res. Sci., Commun. Technol.*, vol. 165, no. 9, pp. 448–454, Oct. 2022.
[84] I. G. S. M. Diyasa, N. M. I. M. Mandenni, M. I. Fachrurrozi, S. I. Pradika, K. R. N. Manab, and N. R. Sasmita, "Twitter sentiment analysis as an evaluation and service base on Python textblob," in *Proc. Conf., Mater. Sci. Eng.*, vol. 1125, 2021, Art. no. 012034.

[85] S. Elbagir and J. Yang, "Sentiment analysis on Twitter with Python's natural language toolkit and VADER sentiment analyzer," in *Proc. IAENG Trans. Eng. Sci.*, Jan. 2020, pp. 63–80.
[86] I. Hemalatha, G. P. S. Varma, and A. Govardhan, "Preprocessing the informal text for efficient sentiment analysis," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 1, no. 2, pp. 58–61, 2012.
[87] V. Bonta, N. Kumaresh, and N. Janardhan, "A comprehensive study on lexicon based approaches for sentiment analysis," *Asian J. Comput. Sci. Technol.*, vol. 8, no. S2, pp. 1–6, Mar. 2019.
[88] S. Elbagir and J. Yang, "Twitter sentiment analysis using natural language toolkit and vader sentiment," in *Proc. Int. Multiconference Eng. Comput. Scientists*, vol. 122, 2019, p. 16.
[89] A. Borg and M. Boldt, "Using VADER sentiment and SVM for predicting customer response sentiment," *Expert Syst. Appl.*, vol. 162, Dec. 2020, Art. no. 113746.
[90] G. Bathla, R. Rani, and H. Aggarwal, "Stocks of year 2020: Prediction of high variations in stock prices using LSTM," *Multimedia Tools Appl.*, vol. 82, no. 7, pp. 9727–9743, Mar. 2023.
[91] Absa. (2023). *Absa Investor Relations*. Accessed: Aug. 30, 2023. [Online]. Available: https://www.absa.africa/investor-relations/financial-results/
[92] Capitec Bank. (2023). *Capitec Bank Financial Results 2023*. Accessed: Sep. 5, 2023. [Online]. Available: https://www.capitecbank.co.za/financial-results/2023/
[93] Nedbank. (2023). *Nedbank Investor Relations*. Accessed: Aug. 10, 2023. [Online]. Available: https://www.nedbank.co.za/content/nedbank/desktop/gt/en/investor-relations/information-hub/financial-results.html
[94] Standard Bank. (2023). *Standard Bank Financial Results*. Accessed: Jul. 22, 2023. [Online]. Available: https://reporting.standardbank.com/results-reports/financial-results/
[95] FirstRand. (2023). *Firstrand Integrated Reporting Hub*. Accessed: Jul. 15, 2023. [Online]. Available: https://www.firstrand.co.za/investors/integrated-reporting-hub/financial-reporting/
[96] D. W. Arner, J. Barberis, and R. P. Buckey, "FinTech, RegTech, and the reconceptualization of financial regulation," *Nortwestern J. Int. Law Bus.*, vol. 37, p. 371, May 2016.
[97] Z. K. Chomba, *The Impact of the COVID-19 Pandemic on the South African Equity Market: An Event Study Analysis*. Johannesburg, South Africa: Univ. Johannesburg, 2021.
[98] R. Harrisberg, "An analysis of the low-volatility anomaly on the Johannesburg stock exchange," M.S. thesis, Fac. Commerce ,Dept. Finance Tax, Univ. Cape Town, South Africa, 2019. [Online]. Available: https://open.uct.ac.za/items/b2758f2c-3339-4dcc-a6fd-21624599cb00
[99] J. Kessler, "Scattertext: A browser-based tool for visualizing how corpora differ," in *Proc. ACL, Syst. Demonstrations*, 2017, pp. 85–90.
[100] F. Pedregosa, S. Varoquaux, A. Gramfort, V. Michel, and B. Thirion, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Dec. 2011.
[101] S. Bates, T. Hastie, and R. Tibshirani, "Cross-validation: What does it estimate and how well does it do it?" *J. Amer. Stat. Assoc.*, vol. 119, pp. 1434–1445, Apr. 2023.
[102] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4765–4774.
[103] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

MIEHLEKETO MATHEBULA (Member, IEEE) is currently pursuing the Master of Science degree in computer sciences with the University of Pretoria, South Africa, specializing in machine learning, NLP, and algorithm optimization. He is an Active Member of the data science for social impact research group with the University of Pretoria, he contributes to research spanning data science for society and local language NLP, with expertise in machine learning, NLP, social media analysis, and the examination of South African financial institutions.

ABIODUN MODUPE (Member, IEEE) received the Ph.D. degree in computer science from the University of the Witwatersrand. He is currently a Lecturer and a Coordinator of the data science program with the University of Pretoria. His research interests include computational linguistics, AI, and data science, aiming to develop unique language processing, and comprehension solutions.

VUKOSI MARIVATE (Senior Member, IEEE) is currently an Associate Professor with the University of Pretoria and holds the ABSA UP Chair of Data Science. He specializes in machine learning (ML) and AI, focusing on NLP and local or low-resource languages. As a Leader of the Data Science for Social Impact (DSFSI) Group, he investigates improving NLP for African languages and better AI models for societal settings. He co-founded Lelapa AI, an AI startup focusing on AI for Africans, and co-founded the Masakhane NLP research foundation. He is a Co-Founder of the Deep Learning Indaba.