Cloud Computing and Data Science
https://ojs.wiserpub.com/index.php/CCDS/

Research Article

Challenges in Detecting Nuanced Sentiment with Advanced Models

Edgar Ceh-Varela*, Sarbagya Ratna Shakya, Essa Imhmed

Department of Mathematical Sciences, Eastern New Mexico University, USA
E-mail: Eduardo.Ceh@enmu.edu

Received: 28 December 2024; **Revised:** 20 February 2025; **Accepted:** 26 February 2025

Abstract: Sentiment analysis, an essential task in Natural Language Processing (NLP), determines the sentiment expressed in texts. This paper compares six different sentiment analysis models, categorized into three groups based on their underlying techniques: lexicon-based, machine learning-based, and zero-shot learning. The models are evaluated on four publicly available datasets (Movie Reviews, Amazon, Yelp, and Financial), each varying in complexity. The main objective is to assess the efficiency of these models in both binary (positive and negative) and ternary (positive, neutral, and negative) sentiment classification scenarios. Our results indicate that for binary classification, pre-trained large-scale NLP state-of-the-art models outperform other approaches, demonstrating superior results across all evaluated metrics. On average, across all datasets, these models achieved 94% accuracy, 96% precision, 94% recall, and 94% F1-score. However, these pre-trained NLP models face significant challenges in three-class classification tasks, where their performance noticeably declines, achieving on average, across datasets, 60% accuracy, 66% precision, 60% recall, and 56% F1-score. This study highlights the limitations of current state-of-the-art models in handling more subtle sentiment distinctions. It emphasizes the need for further advancements in sentiment analysis techniques to effectively manage multi-class sentiment categorization that captures and interprets specialized jargon, technical terminology, and nuanced language.

***Keywords*:** sentiment classification, natural language processing, zero-shot learning

1. Introduction

Sentiment analysis is an essential task in Natural Language Processing (NLP) [1-4]. Its goal is to classify text into predefined categories of sentiments, such as positive, negative, or neutral [5, 6]. This classification allows for measuring public opinion and sentiment trends effectively. The task has become increasingly important in different applications, such as social media and customer feedback.

Different models have been developed to work on sentiment analysis problems. These models fall into different categories [5-7]. In this study, we focus on lexicon-based, machine learning, and zero-shot models. Each type of model has its own strengths and weaknesses, making the choice of model dependent on the specific requirements of the sentiment analysis task at hand.

Lexicon-based models depend on predefined dictionaries of words and their sentiment [8-10]. These models are straightforward and efficient, particularly for analyzing social media text where slang and informal language are prevalent. However, they often struggle with context and tone, which can lead to inaccuracies in sentiment classification.

DOI: https://doi.org/10.37256/ccds.6220256316

Machine Learning (ML) models rely on statistical methods to classify sentiments [11, 12]. These models are typically trained on labeled datasets where they learn to identify sentiment patterns based on word frequencies and other textual features. While they are generally faster to train and deploy, their performance can be limited by the quality and quantity of the training data and their ability to capture the context and subtleties of human language.

Zero-shot learning sentiment models offer a novel approach by enabling sentiment classification without the need for labeled training data specific to the target task [13]. These models use large language models and transfer learning to predict sentiment labels in scenarios where annotated data is small or unavailable [14]. By understanding general language patterns and semantics, zero-shot models can infer sentiment categories from text based on minimal examples or descriptions of the categories. This approach is particularly advantageous in rapidly evolving domains or low-resource languages, but it may struggle with domain-specific nuances and fine-grained sentiment distinctions.

One of the key challenges in sentiment analysis is accurately classifying neutral sentiments. While binary classifiers effectively distinguish between positive and negative sentiments, their performance declines when a neutral category is introduced [15, 16]. This difficulty arises because neutral sentiments are inherently more ambiguous and context- dependent, making them harder to classify accurately. Even pre-trained models, despite their overall robustness, struggle with this added complexity, leading to a noticeable drop in performance for ternary classification tasks (i.e., positive, neutral, and negative). Accurately classifying neutral sentiments is crucial for reliable sentiment analysis across various real-world applications. In customer feedback analysis, misclassifying neutral reviews can lead to misguided business strategies, while in social media monitoring, it can result in a biased understanding of public opinion. Financial markets rely on precise sentiment analysis to predict trends, and misclassification can cause significant financial losses. In healthcare, accurate sentiment analysis of patient feedback ensures better service improvements, and in political analysis, it helps tailor effective campaigns by understanding undecided voters. Addressing these challenges is essential for obtaining balanced insights and making informed decisions [15, 17-19].

To address this challenge, our study evaluates sentiment analysis models using four publicly available datasets (Movie Reviews, Amazon, Yelp, and Financial) chosen for their diverse textual styles and domain complexities. This comprehensive approach ensures a robust and generalizable assessment of model performance across varied real-world scenarios.

This study makes the following important contributions:

(a) We compare six different sentiment analysis models. This comprehensive comparison provides a broader understanding of model performance across different scenarios.

(b) We assess the models' efficiency in both binary and ternary sentiment classification scenarios across four publicly available datasets. This dual approach addresses the limitation of previous works that typically analyze only binary or ternary classification, providing a more complete understanding of sentiment analysis.

(c) Our results reveal the need for advancing sentiment analysis techniques to improve multi-class categorization, particularly for nuanced language, technical terminology, and specialized jargon. This insight addresses the limitation of existing models that struggle with complex and domain-specific language, highlighting areas for future research and development.

Our findings have important implications for developing more accurate and efficient sentiment analysis models. Additionally, our research contributes to the growing body of knowledge on sentiment prediction techniques. Beyond model comparisons, we provide new insights into the limitations of zero-shot learning models in nuanced sentiment classification, particularly in domain-specific contexts such as financial texts.

The remainder of this paper is organized as follows: Section 2 presents the literature related to our research. Section 3 details the proposed methodology. The results and discussion of using the sentiment analysis models with the different datasets are presented in Sections 4 and 5, respectively. Finally, we present the conclusions in Section 6.

2. Literature review

The field of sentiment analysis has seen significant advancements driven by the development of diverse methodologies and the increasing availability of textual data. For this study, we focus on three approaches to sentiment analysis: lexicon-based, machine learning, and zero-shot learning models. These approaches encompass a range of

techniques, each with its own strengths and applications. The following subsections detail these approaches, highlighting key methods and findings in each area.

2.1 *Lexicon-based models*

Lexicon-based sentiment models operate on the premise that the semantic orientation of a text is directly linked to the polarity of the words and phrases it contains. These models utilize sentiment lexicons to assign sentiment scores to the text, considering the presence and intensity of various parts of speech such as words, adjectives, adverbs, nouns, verbs, as well as the phrases and sentences that include them [8-10].

Researchers used Valence Aware Dictionary for sentiment Reasoning (VADER) [20] to analyze the average daily sentiment of 86,581,237 tweets to identify emerging themes about COVID-19 and observe sentiment changes in response to the pandemic [21]. Al Mansoori et al. [22], assessed criminal behavior on Facebook and Twitter by performing sentiment analysis with VADER to classify data as negative, positive, or neutral, ultimately aiding in suspect identification. VADER was also employed by Scholz and Jeznik [23] to conduct an integrated semantic analysis of tweets from 2008 to 2018, aiming to detect tourism flows in the province of Styria, Austria. Isnan et al. [24] explored sentiment analysis in TikTok reviews using both VADER and Support Vector Machine (SVM) models, where VADER was utilized to label reviews as positive, neutral, or negative.

Al Mansoori et al. [22], used TextBlob [25, 26], a sentiment analysis tool that uses a prebuilt rule-based model, yielded higher classification accuracy compared to original annotations and outperformed VADER. Likewise, Diyasa et al. [27] utilized TextBlob to analyze 3,324 tweets about TELKOM's products and services. Furthermore, Chandrasekaran and Hemant [28] combined TextBlob with ML and deep learning (DL) techniques, to assess public sentiment during the peak of the Coronavirus pandemic.

Thus, lexicon-based sentiment models analyze text using predefined sentiment lexicons, making them effective for large-scale studies and adaptable for integration with other techniques. However, while they excel in scalability and simplicity, they face challenges when applied to more complex datasets for sentiment classification tasks. Their reliance on predefined lexicons limits their ability to capture nuanced or domain-specific sentiments, such as sarcasm or irony, which can hinder their performance.

2.2 *Machine learning models*

Machine learning models address sentiment analysis by employing ML algorithms to classify text [11, 12, 29]. Sentiments are commonly categorized as positive or negative, although texts can also fall into neutral categories. Training an ML model on a large dataset of effectively annotated texts allows it to learn both the affective valence of specific keywords and consider other contextual factors like punctuation and word co-occurrence frequencies [30].

Text embeddings [31-33] have revolutionized NLP. Embedding algorithms generate vector representations for text tokens, capturing context and aiding word prediction. Moreover, word embeddings offer learned representations where semantically related words share similar vectors or numeric mappings [34, 35].

Ahmad et al. [12], introduced a sentiment analysis model that leverages Naïve Bayes and SVM for analyzing 100,000 tweets. Likewise, Afifah et al. [36] suggested a sentiment analysis approach using Extreme Gradient Boosting (XGBoost), which combines weak learners into a powerful ensemble classifier. The study specifically examines reviews of Indonesia's widely used telemedicine app, Halodoc, especially during the COVID-19 pandemic.

Sentiment analysis was used by Shakya et al. [37] to improve stock market predictions. The researchers used Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM) and Convolutional Neural Network + Bidirectional Long Short-Term Memory (CNN + BiLSTM) models combining stock values with sentiment from news and tweets. The results indicate that using the sentiment improved considerably the stock predictions. Ahuja et al. [38], explored the effect of Term Frequency-Inverse Document Frequency (TF-IDF) and N-gram features on sentiment analysis using a tweet dataset. The findings reveal that TF-IDF enhances sentiment analysis performance by 3-4% compared to N-gram features. The study employs six classification algorithms: Decision. Trees, SVM, K-Nearest Neighbors (K-NN), Random Forest, Logistic Regression, and Naïve Bayes. Furthermore, Rana et al. [39] employed the Bidirectional Encoder Representations from Transformers (BERT) model to directly convert plain text into numerical representations. These representations, combined with SVM and Naïve Bayes algorithms, analyze sentiment patterns in movie reviews.

Overall, machine learning-based sentiment analysis models leverage large annotated datasets, text embeddings, and various feature extraction methods to effectively analyze sentiment, considering both keyword meanings and contextual factors. These models offer significant advancements in sentiment classification, but challenges remain, particularly with neutral sentiments and detecting nuanced emotions like sarcasm or irony, which can be difficult to capture. While embedding techniques and deep learning models provide more refined text representations, their ability to generalize across diverse datasets requires further evaluation.

2.3 *Zero-shot learning models*

The requirement for labeled data is one of the primary challenges in developing robust NLP systems. Consequently, recent research has focused on models capable of operating in zero-shot learning methods, where they are not explicitly trained on data from a specific target language or domain. Zero-Shot Learning (ZSL) has gained significant interest because it allows algorithms to scale across unseen classes and be applied to diverse datasets [13, 40, 41].

ZSL using Natural Language Inference (NLI) enables models to classify previously unseen text by leveraging semantic knowledge [42, 43]. This technique has become widely used in NLP research, with various studies applying zero-shot models to tasks such as sentiment analysis.

For example, Pelicon et al. [44] focused on sentiment analysis in news articles using a zero-shot cross-lingual setting, aiming to train models in one language that could also perform well on data in another language. Likewise, Shu et al. [45] aimed to train a unified model for zero-shot Aspect-Based Sentiment Analysis (ABSA) without using any annotated data from a new domain. Kumar et al. [46] proposed a novel method for sentiment analysis in Sanskrit using zero-shot cross-lingual sentiment analysis and machine translation, utilizing cross-lingual mapping and Generative Adversarial Networks (GANs). Lossio-Ventura et al. [3], presented a comparison of different sentiment analysis models applied to health-related free-text survey data in the context of COVID-19. The goal was the prediction of sentiments for two independent COVID-19 survey data sets. Similarly, Tesfagergish et al. [47] proposed a hybrid semi-supervised model for sentiment analysis that integrates emotion detection. They used a zero-shot transformer model to identify emotions, transforming the results into one-hot encoded feature vectors. These vectors are then used to train a supervised machine learning classifier for sentiment classification. Wang et al. [40], assessed ChatGPT's efficiency across seven representative sentiment analysis tasks. Their findings highlight ChatGPT's impressive zero-shot capabilities in sentiment classification.

In conclusion, zero-shot learning has become a key approach in NLP, enabling models to classify unseen text across various domains and languages without the need for labeled data. While these models show promise in sentiment analysis by eliminating the need for labeled data and enabling cross-lingual applications, they may face challenges when applied to diverse and complex datasets. Additionally, their ability to generalize across domain-specific language and contexts require further investigation.

3. Methodology

We adapted our methodology inspired by other works [48, 49]. This methodology is formed by five phases (1) data loading, (2) data pre-processing, (3) model execution, (4) model evaluation, and (5) performance analysis. Figure 1 shows the sequential order of these phases in a schematic manner.

The five phases in our methodology can be described as follows:

(1) Collect data from four distinct datasets related to sentiment analysis prediction.

(2) Preprocess the data and resampling it, if needed, to mitigate class imbalance.

(3) Execute the different models to predict the sentiments for each dataset instance.

(4) Evaluate each model by quantifying performance metrics, such as accuracy, precision, recall, and F1-score.

(5) Conduct a comparative analysis to understand the effectiveness of each type of sentiment analysis model.

Our experiments were executed inside Google Colab [https://colab.research.google.com], a platform used for machine learning projects.

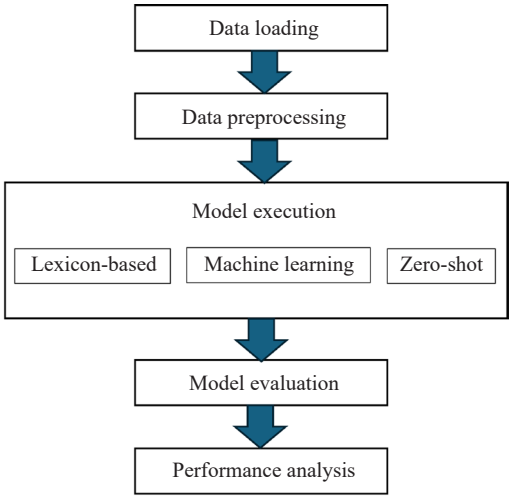

Figure 1. Methodology for this study

3.1 *Datasets*

In this section, we describe the datasets used in the study. We selected four publicly available datasets that support both binary (positive and negative) and ternary (positive, neutral, and negative) classification. These datasets were chosen for their diverse textual styles and domain complexities, making them suitable for evaluating sentiment analysis challenges. Table 1 summarizes the datasets, with further details provided in the following sections.

Table 1. Summary of the datasets Instances

	Instances	% Negative	% Neutral	% Positive
Movie reviews	4,008	26.74	41.3	31.96
Amazon	1,000	50	-	50
Yelp	1,000	50	-	50
Financial	2,254	13.38	61.4	25.22

3.1.1 *Movie reviews dataset*

The Rotten Tomatoes movie review dataset [https://www.kaggle.com/competitions/sentiment-analysis-on-movie-reviews/data], initially collected by [50], is utilized for sentiment analysis. This dataset allows for benchmarking sentiment analysis models by labeling phrases on a five-point scale: negative, somewhat negative, neutral, somewhat positive, and positive. It presents challenges such as sentence negation, sarcasm, terseness, and language ambiguity. The sentiment labels range from 0 (negative) to 4 (positive). This dataset contains a total of 156,060 instances. For our study, we focus exclusively on the negative, neutral, and positive classes. Specifically, there are 1,072 instances in the negative class, 1,655 instances in the neutral class, and 1,281 instances in the positive class. Figures 2a and 3a show the two- and three-class distribution for this dataset.

3.1.2 *Amazon dataset*

The Amazon dataset [https://archive.ics.uci.edu/dataset/331/sentiment+labelled+sentences] is a public dataset

created by [51]. It contains reviews and scores for products sold on amazon.com in the cell phones and accessories category. There are 500 clearly positive (i.e., class 1) and 500 clearly negative (i.e., class 0) sentences. Those were selected randomly for a larger dataset collected by [52]. Figure 2b shows the class distribution for this dataset.

3.1.3 *Yelp dataset*

The Yelp [https://archive.ics.uci.edu/dataset/331/sentiment+labelled+sentences] dataset contains restaurant reviews and scores extracted from the Yelp dataset challenge. In total there are 1,000 instances, 500 clearly positive as class 1 and 500 clearly negative as class 0. This dataset was created by [51]. Figure 2c shows the class distribution for this dataset.

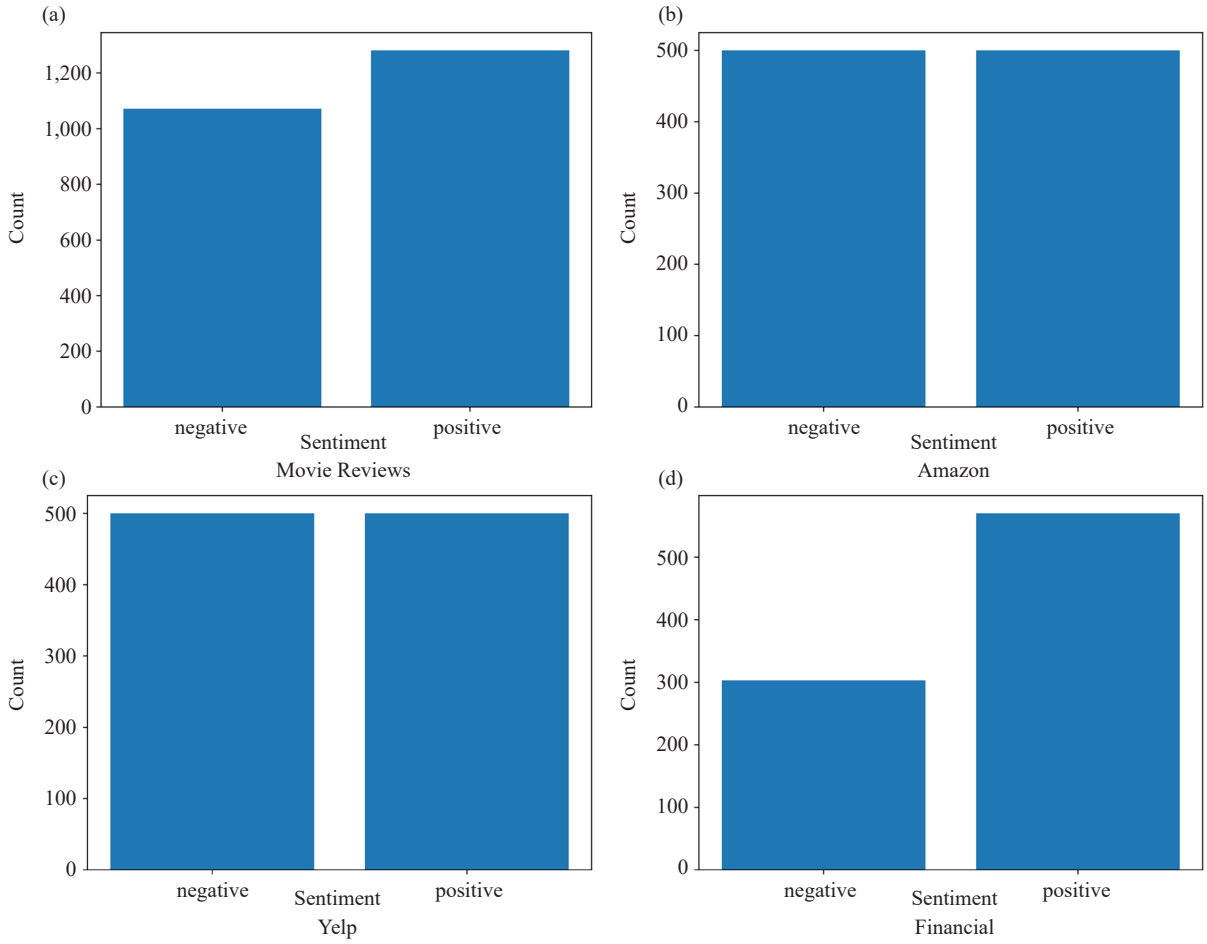

Figure 2. Two-class distribution for the datasets

3.1.4 *Financial dataset*

The dataset is part of the HuggingFace repository [https://huggingface.co/datasets/takala/financial_phrasebank] and consists of sentences from English- language financial news categorized by sentiment. Originally presented in [53], this polar sentiment dataset includes a total of 4,840 sentences. In this release, the financial phrase bank comprises 2,264 sentences, annotated by 16 individuals with sufficient financial market expertise. The dataset contains 303 negative instances, 1,391 neutral instances, and 570 positive instances. Figures 2d and 3b show the two- and three-class distribution for this dataset.

3.2 *Models*

This section describes the different models used for this study. We were interested in using sentiment analysis models of different characteristics. As mentioned in Section 2, three different types of models were used: lexicon-based, machine learning, and zero-shot learning.

3.2.1 *Lexicon-based*

Lexicon-based sentiment analysis is a technique in natural language processing that detects sentiment by using predefined word lists (lexicons) and rules associated with specific sentiment scores. These lexicons help evaluate the overall sentiment of text based on the presence and intensity of these words.

VADER. We utilize Valence Aware Dictionary for Sentiment Reasoning (VADER) [20], a widely used rule-based sentiment analysis model specifically designed for short, informal text such as social media posts. VADER assigns sentiment scores by leveraging a gold-standard sentiment lexicon and applying five grammatical and syntactical rules to enhance sentiment intensity detection. Given its effectiveness in microblogging contexts, VADER is well-suited for analyzing our dataset.

TextBlob. We incorporate TextBlob [25, 26], a Python-based NLP library offering various text-processing functions, including sentiment analysis. TextBlob computes polarity, which ranges from -1 (negative) to +1 (positive), and subjectivity, which ranges from 0 (objective) to 1 (subjective). This dual evaluation allows for a nuanced understanding of sentiment within our dataset.

3.2.2 *Machine learning*

Machine learning models for sentiment classification learn patterns from large labeled text datasets. During training, the model adjusts its internal parameters to enhance pattern recognition. Once trained, it can classify new, unseen text based on its learned knowledge.

Base classifier. For sentiment classification, we employ XGBoost [54], a high-performance gradient boosting algorithm known for its efficiency and scalability. XGBoost constructs an ensemble of decision trees, iteratively refining predictions by correcting errors from previous iterations. It incorporates built-in regularization to mitigate overfitting, handles missing data automatically, and applies tree pruning techniques to enhance generalization. In our study, we configure XGBoost to balance model complexity and predictive performance. We set max_depth = 10 to capture intricate patterns in textual data, n_estimators = 100 to ensure sufficient boosting iterations, and learning_rate = 0.01 to stabilize convergence while preventing overfitting. These hyperparameters allow XGBoost to effectively classify sentiment in large and complex datasets. The implementation is based on the official XGBoost library [https://xgboost.readthedocs.io/en/stable/index.html].

Text representation. In NLP tasks, text representation converts raw text into a format that machines can understand. Simple techniques, such as Bag of Words (BoW) and n-grams, count word frequencies and sequences but lack semantic context. More advanced methods, like Term Frequency-Inverse Document Frequency (TF-IDF), emphasize word importance within documents. Topic modeling, such as Latent Dirichlet Allocation (LDA), identifies underlying themes across a corpus. The most sophisticated approaches involve embeddings like Word2Vec, GloVe, and BERT, which create dense vector representations capturing both syntactic and semantic meanings, providing rich and effective text representations. For our study, we selected TF-IDF and SentenceBERT as our two text representation techniques. These techniques have been used in other works [38, 55-58].

TF-IDF [59] is a widely used technique for representing text, particularly in traditional NLP tasks. TF-IDF assesses a word's importance within a document relative to a larger collection of documents (corpus). It combines two key metrics: Term Frequency (TF), which counts how often a word appears in a document, and Inverse Document Frequency (IDF), which downweights common words and emphasizes unique terms. The product of TF and IDF yields a score that reflects a word's significance in a specific document while mitigating the impact of frequently occurring words across the entire corpus. TF-IDF is especially valuable for tasks like document classification and information retrieval, as it highlights the most informative words that differentiate one document from another. We use the Scikit-learn's TF-IDF vectorizer [https://scikit-learn.org/], keeping the top 5,000 most frequent terms.

Sentence-BERT (SBERT) [60] is an advanced technique for representing text that extends the BERT [61] model to create embeddings for entire sentences, rather than individual words. SBERT fine-tunes BERT using Siamese and triplet network structures, resulting in sentence embeddings that can be compared using cosine similarity. This approach captures both syntactic and semantic relationships between sentences, making it ideal for tasks involving context and sentence meaning, such as semantic textual similarity, clustering, and retrieval. Unlike traditional BERT, which generates context-aware word embeddings, SBERT produces fixed-size vector representations for sentences, significantly enhancing efficiency in similarity comparisons and other downstream NLP tasks. We use the HuggingFace sentence-transformer with the all-MiniLM-L6-v2 model [https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2].

For this study, we create two machine learning models: one combines XGBoost with TF-IDF and another uses XGBoost with SBERT. To ensure a fair comparison with the pretrained models, SBERT was used with its default settings, and for XGBoost-based models, we applied the default hyperparameters from the official XGBoost library without additional optimization.

3.2.3 *Zero-shot learning*

Zero-shot learning sentiment classification models leverage pre-trained language models with extensive linguistic knowledge to analyze and categorize text without relying on domain-specific training data. While this approach offers flexibility across different tasks, it may be less precise than fine-tuned models tailored for specific domains or nuanced sentiment analysis.

NLI. The NLI-based zero-shot text classification model, proposed by [62], is an innovative approach that repurposes pre-trained Natural Language Inference (NLI) models for text classification without requiring task-specific training data. This method treats the text to be classified as the premise and constructs hypotheses from candidate labels. The model then evaluates the likelihood of entailment between the premise and each hypothesis, converting entailment and contradiction probabilities into label probabilities.

This approach offers remarkable flexibility, allowing new classes to be added by simply formulating new hypotheses without retraining. It has shown promising results across various domains and tasks, particularly when used with larger pre-trained models like Bidirectional and Auto-Regressive Transformer (BART) [63]. The effectiveness of this method stems from the rich semantic understanding captured by NLI models, enabling informed decisions about text classification even for previously unseen categories, making it a powerful tool for zero-shot classification scenarios. We use the HuggingFace pretrained bart-large-mnli [https://huggingface.co/facebook/bart-large-mnli] model created by Facebook (i.e., Meta).

Fast Language-Agnostic In-context Representations-Task-aware (FLAIR-Task-aware) Representation of Sentences (TARS). The zero-shot FLAIR-TARS [64] model is an advanced text classification system within the FLAIR natural language processing library, which focuses on sentiment analysis tasks and leverages deep learning techniques. This model utilizes a Task-aware Representation of Sentences (TARS) approach, enabling it to perform classification tasks on unseen labels or classes without requiring specific training data for those classes.

At its core, FLAIR [65] employs sequence labeling to train models that predict sentiment labels for individual words or tokens within a text. This granular approach allows the model to capture sentiment expressed by specific words or phrases within a sentence or document. The FLAIR library provides pre-trained models in various languages that can be utilized with minimal initial training, and these models can be fine-tuned using unique datasets relevant to particular applications or domains. The key advantage of FLAIR, and by extension the FLAIR-TARS model, is its ability to extract contextual information, considering neighboring words and sentence structure to comprehend the sentiment conveyed in complex and ambiguous texts more accurately. This contextual understanding is particularly valuable in zero-shot scenarios, where the model must classify text into new categories without prior examples, relying solely on its pre-existing knowledge and the semantics of the class labels. For our study, we used the model provided by [64].

3.3 *Preprocessing*

Text preprocessing plays a vital role in NLP and sentiment analysis. It involves preparing raw textual data for analysis by applying techniques like tokenization, stemming, lemmatization, stop word removal, and special character

handling. Effective text preprocessing enhances analysis quality and efficiency by reducing noise, ensuring data consistency, and facilitating the extraction of meaningful patterns. For this study, we perform a simple preprocessing step. For all the datasets we remove special characters and numbers and transform the text to lowercase characters.

In sentiment analysis, including stop words is critical. Stop words, such as 'not', 'never', and 'without', can significantly impact the sentiment of a sentence. Removing them may lead to inaccurate interpretations and sentiment analysis results. Therefore, retaining stop words is essential to preserve the true sentiment expressed in the text [66]. Therefore, we do not remove them as part of our preprocessing.

Only for the model using TF-IDF we apply lemmatization. TF-IDF benefits from lemmatization by ensuring consistency and relevance in the analyzed textual data. Lemmatization reduces words to their base or root form, treating different variations as a single term. This enhances term frequency calculations and improves understanding of word importance, benefiting text mining and information retrieval tasks [67].

An imbalanced dataset can bias a prediction model toward the majority class, leading to poor generalization, low recall for minority classes, and misleading performance metrics. The Financial dataset exhibits a significant imbalance when three sentiment classes are used (see Figure 3b). To mitigate this, we applied an upsampling strategy to increase the instances of negative and positive classes to match the neutral class. Balancing the dataset improves model fairness, refines decision boundaries, enhances recall and precision, and reduces overfitting to the dominant class. This results in more reliable sentiment classification and ensures meaningful insights across all sentiment categories.

We split the datasets as 80% of the instances for training and 20% for testing. The training set was used to train the machine learning models. The other models do not need a training step. Therefore, all models were tested with the same test set for fairness.

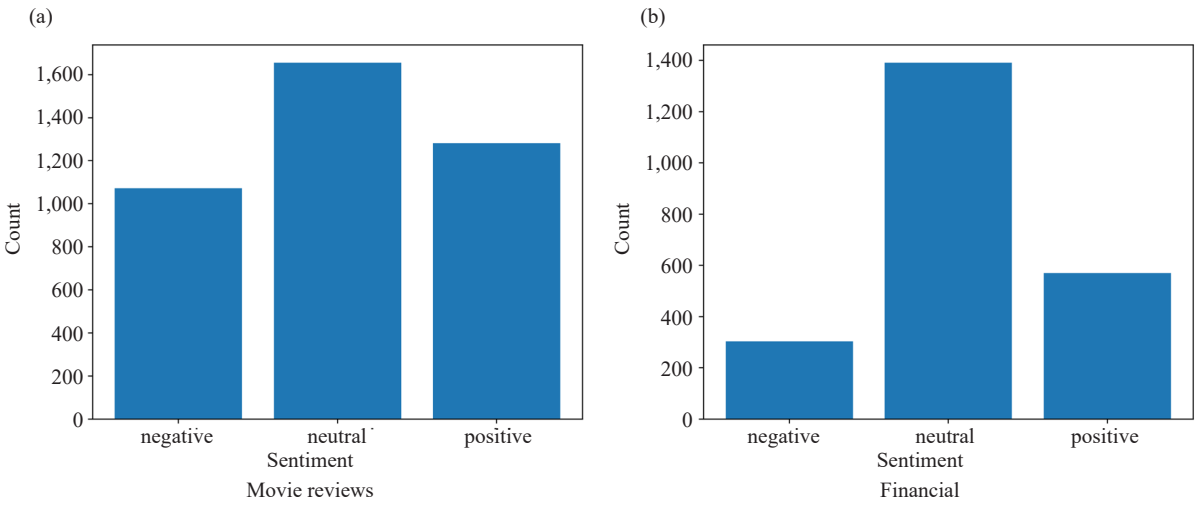

Figure 3. Three-class distribution for the datasets

3.4 *Performance metrics*

Different performing metrics were used to evaluate the sentiment analysis models. These metrics are widely used in other works [3, 29, 49] for sentiment classification problems. Table 2 presents a description of the metrics we used in this study. Each of these metrics offers a unique view of the model's performance, making it crucial to consider them all when assessing a machine learning model. In certain situations, high accuracy alone may not be sufficient to judge a model's effectiveness, as metrics like precision, recall, and others could hold greater significance.

Table 2. Metrics used for evaluating the models [49]

Metric	Description
Accuracy	Calculated by dividing the number of accurate predictions by the total number of predictions, it represents the model's percentage of accurate predictions.
Precision	Divides the number of true positive forecasts by the sum of the true positive and false positive predictions. It determines the percentage of positive predictions that are genuinely accurate.
Recall	This metric, which is determined by dividing the number of true positive forecasts by the total of the true positive and false negative predictions, indicates the percentage of real positive cases that the model properly detects.
F1-score	The harmonic mean of accuracy and recall is used to compute this metric, which shows how well the two are balanced. A higher score denotes a better balance between the two.

4. Results

In this section, we present the results obtained from using the different datasets with all the models. We analyzed the results for binary (i.e., negative and positive) and ternary (i.e., negative, neutral, and positive) sentiment classification. The results are presented using the performance metrics (accuracy, precision, recall, and F1-score) for each model.

4.1 *Binary sentiment classification*

Table 3 shows the results obtained for the Movie Reviews dataset. The results demonstrate a clear superiority of zero-shot models in this task. FLAIR achieved the highest performance across all metrics, with an accuracy of 0.972, followed closely by the NLI-based model with 0.964. These zero-shot approaches significantly outperformed all other models, showcasing their ability to generalize well to unseen data without task-specific training.

BERT + XGBoost showed the best performance among the machine learning models, with an accuracy of 0.76, surpassing the lexicon-based models and the TFIDF + XGBoost approach. This highlights the effectiveness of BERT's contextual embeddings in capturing sentiment nuances. Interestingly, the lexicon-based models, VADER and TextBlob, outperformed the TFIDF + XGBoost model. TextBlob achieved an accuracy of 0.743, slightly higher than VADER's 0.726, indicating that these simple, rule-based approaches can be surprisingly effective for sentiment analysis tasks. The TFIDF + XGBoost model showed the lowest performance among all tested approaches, with an accuracy of 0.628. This suggests that the bag-of-words representation might not capture the complexities of sentiment in movie reviews as effectively as other methods.

Table 3. Results for the Movie Reviews dataset for 2 classes

	Accuracy	Precision	Recall	F1-score
Vader	0.726	0.742	0.726	0.716
TextBlob	0.743	0.773	0.743	0.73
TFIDF + XGBoost	0.628	0.645	0.628	0.599
BERT + XGBoost	0.76	0.762	0.76	0.757
NLI	0.964	0.965	0.964	0.964
FLAIR	0.972	0.974	0.972	0.972

We evaluated the models on the Amazon dataset for binary classification. Table 4 shows that the zero-shot models outperformed both the lexicon-based and machine learning approaches. Notably, the NLI-based model achieved the highest performance across all metrics, with a 0.97 for accuracy, precision, recall, and F1-score. This consistent performance indicates that the NLI model excels in both correctly identifying positive instances (i.e., precision) and

capturing a high proportion of actual positive instances (i.e., recall). The FLAIR model, showed the second-best performance with scores of 0.955, closely following the NLI model. This further underscores the effectiveness of zero-shot learning techniques in this classification task.

Among the machine learning models, BERT + XGBoost slightly outperformed TFIDF + XGBoost, achieving accuracy and F1-score of 0.8 compared to 0.735 and 0.733, respectively. This suggests that the contextual embeddings provided by BERT offer some advantages over the traditional TFIDF approach. Interestingly, the lexicon-based models showed competitive performance, with TextBlob achieving slightly better results (i.e., 0.795 accuracy, 0.794 F1-score) compared to VADER (0.775 accuracy, 0.765 F1-score). This indicates that even simple lexicon-based approaches can provide reasonable results for this particular classification task.

Table 4. Results for the Amazon dataset for 2 classes

	Accuracy	Precision	Recall	F1-score
Vader	0.775	0.809	0.775	0.765
TextBlob	0.795	0.796	0.795	0.794
TFIDF + XGBoost	0.735	0.755	0.735	0.733
BERT + XGBoost	0.8	0.804	0.8	0.8
NLI	0.97	0.97	0.97	0.97
FLAIR	0.955	0.956	0.955	0.955

The results for the Yelp dataset are presented in Table 5. The results demonstrate a clear superiority of the zero-shot models over both lexicon-based and machine learning approaches. The NLI-based model achieved the highest performance across all metrics, with an accuracy of 0.98, precision of 0.981, recall of 0.98, and F1-score of 0.98. This performance indicates the model's strong ability to generalize and accurately classify sentiments without task-specific training data. Following closely, the FLAIR model, showed robust performance with an accuracy of 0.95 and consistent scores across all metrics. This further reinforces the effectiveness of zero-shot learning techniques in sentiment analysis tasks.

Among the machine learning models, BERT + XGBoost outperformed TFIDF + XGBoost, achieving an accuracy of 0.76 and an F1-score of 0.76. This suggests that the contextual embeddings provided by BERT offer a significant advantage over traditional TFIDF features in capturing sentiment nuances. Interestingly, the lexicon-based models showed competitive performance, with TextBlob slightly outperforming VADER. TextBlob achieved an accuracy of 0.75 and a balanced F1-score of 0.75, while VADER reached an accuracy of 0.725 with a slightly lower F1-score of 0.706.

Table 5. Results for the Yelp dataset for 2 classes

	Accuracy	Precision	Recall	F1-score
Vader	0.725	0.785	0.725	0.706
TextBlob	0.75	0.751	0.75	0.75
TFIDF + XGBoost	0.675	0.72	0.675	0.662
BERT + XGBoost	0.76	0.764	0.76	0.76
NLI	0.98	0.981	0.98	0.98
FLAIR	0.95	0.952	0.95	0.95

Finally, Table 6 presents the results for the Financial dataset. The lexicon-based models, VADER and TextBlob, show the lowest performance across all metrics, with accuracies of 0.619 and 0.61 respectively. This suggests that simple sentiment lexicons are insufficient for capturing the nuances of financial text classification. Machine learning

models demonstrate significantly better performance. The TFIDF + XGBoost model achieves an accuracy of 0.917 and consistent scores across all metrics. However, the BERT + XGBoost model outperforms all other approaches, achieving the highest scores in every metric with accuracy, precision, recall, and F1-score of 0.978.

Interestingly, the zero-shot models were not the models presenting the highest scores as happened with the other datasets. The NLI-based model performs well, with an accuracy of 0.971, only slightly behind the best-performing BERT + XGBoost model. The FLAIR model, while not matching the performance of NLI or the machine learning models, still outperforms the lexicon-based approaches with an accuracy of 0.803.

Overall, the different results highlight the remarkable advancements in zero-shot learning techniques for sentiment analysis, significantly outperforming both traditional lexicon-based methods and sophisticated machine learning approaches. We need to recall that the sentences from the Amazon and Yelp datasets present clear sentiments (i.e., clearly positive and negative), as described in Section 3. The Movie reviews dataset presents some challenges such as sentence negation, sarcasm, terseness, and language ambiguity. The Financial dataset presents sentences where the sentiments are not clearly positive or negative. Similarly, as presented in Tables 9 and 10, in Section 5, the text in the Financial dataset is more complex than the other datasets. These features seem to influence the performance of the zero-shot models.

Table 6. Results for the Financial dataset for 2 classes

	Accuracy	Precision	Recall	F1-score
Vader	0.619	0.672	0.619	0.566
TextBlob	0.61	0.659	0.61	0.594
TFIDF + XGBoost	0.917	0.922	0.917	0.917
BERT + XGBoost	0.978	0.979	0.978	0.978
NLI	0.971	0.972	0.971	0.971
FLAIR	0.803	0.91	0.803	0.803

As we see from Table 6, the zero-shot learning models did not perform as well on the Financial dataset compared to the other datasets. Thus, we further analyzed the confusion matrices for the BERT + XGBoost and FLAIR models to gain deeper insights into their classification performance. By examining the distribution of true positives, false positives, and false negatives, we can identify specific classes where the models struggled, detect potential biases, and understand misclassification patterns that may inform further model improvements.

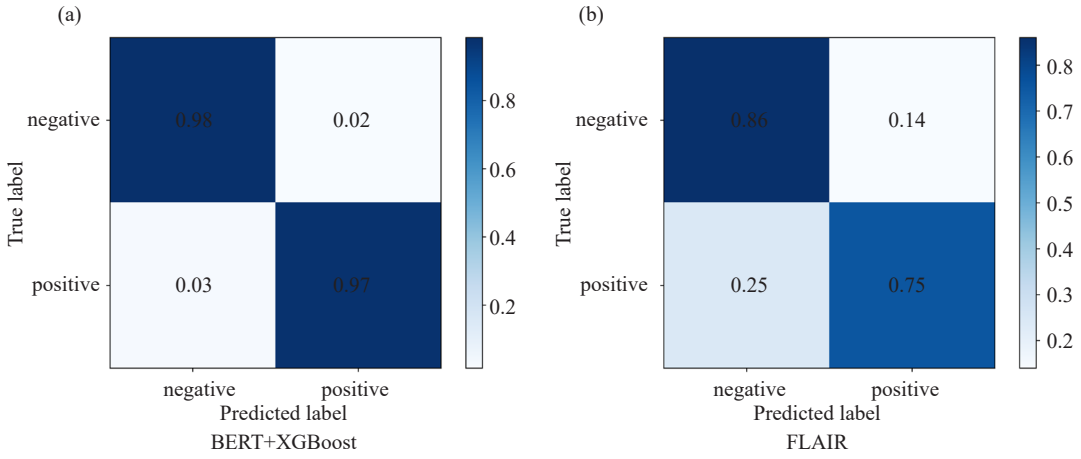

Figure 4. Two-class normalized confusion matrices for the Financial dataset

The comparison of confusion matrices for BERT + XGBoost and FLAIR in the Financial dataset reveals notable performance differences. Figure 4a shows BERT + XGBoost achieving higher accuracy, with strong true positive and true negative rates and minimal false classifications, reflecting its high precision and recall. In contrast, Figure 4b shows FLAIR with higher error rates, including more false positives and false negatives, indicating weaker sentiment differentiation and lower overall precision and recall. Thus, BERT + XGBoost emerges as the more reliable and accurate model for sentiment analysis in this dataset.

4.2 *Ternary sentiment classification*

Table 7 shows the performance metrics for the Movie Reviews dataset for 3-class sentiment classification. From the lexicon-based models, TextBlob slightly outperformed VADER across all metrics, with an accuracy of 0.516 compared to VADER's 0.489. This suggests that TextBlob's more nuanced approach to sentiment analysis provides a modest advantage over VADER's rule-based system for this particular dataset.

The machine learning models demonstrated similar performance to each other, with BERT + XGBoost marginally outperforming TFIDF + XGBoost in terms of accuracy and F1-score. This suggests that for this particular task, the advanced contextual embeddings provided by BERT did not offer a significant advantage over the simpler TF-IDF representation when combined with XGBoost.

Notably, the zero-shot models showed the highest performance among all tested approaches. The NLI-based model achieved an accuracy of 0.59 and the highest precision of 0.681, indicating its strength in correctly identifying positive classifications. However, its lower F1-score of 0.483 suggests some imbalance in its predictions. The FLAIR model demonstrated the best overall performance across all metrics, significantly outperforming all other models. It achieved the highest accuracy of 0.71, precision of 0.71, recall of 0.706, and F1-score of 0.698. This superior performance highlights the effectiveness of FLAIR's task-aware representation approach in capturing the nuances of sentiment in movie reviews, even without task-specific training data.

Table 7. Results for the Movie Reviews dataset for 3 classes

	Accuracy	Precision	Recall	F1-score
Vader	0.489	0.489	0.433	0.433
TextBlob	0.516	0.544	0.516	0.491
TFIDF + XGBoost	0.536	0.543	0.536	0.517
BERT + XGBoost	0.537	0.534	0.537	0.533
NLI	0.59	0.681	0.59	0.483
FLAIR	0.71	0.71	0.706	0.698

Table 8. Results for the Financial dataset for 3 classes

	Accuracy	Precision	Recall	F1-score
Vader	0.384	0.316	0.384	0.249
TextBlob	0.483	0.502	0.483	0.485
TFIDF + XGBoost	0.866	0.876	0.866	0.866
BERT + XGBoost	0.963	0.963	0.963	0.963
NLI	0.666	0.74	0.666	0.601
FLAIR	0.45	0.516	0.45	0.454

Table 8 presents the performance of the sentiment analysis models on the Financial dataset for a 3-class classification task. The lexicon-based models, VADER and TextBlob, showed the lowest performance among all tested models. VADER achieved an accuracy of 0.384 and an F1-score of 0.249, while TextBlob performed slightly better with an accuracy of 0.483 and an F1-score of 0.485. These results suggest that simple lexicon-based approaches may not be sufficient for capturing the nuances of sentiment in financial texts.

Machine learning models demonstrated significantly better performance. The TFIDF + XGBoost model achieved an accuracy of 0.866 and consistent scores across all metrics. However, the BERT + XGBoost model outperformed all other approaches, achieving the highest scores across all metrics with 0.963 for accuracy, precision, recall, and F1-score. This superior performance highlights the effectiveness of combining pre-trained language models like BERT with powerful classifiers such as XGBoost for financial sentiment analysis. Interestingly, the zero-shot models showed mixed results. The NLI-based model achieved moderate performance with an accuracy of 0.666 and an F1-score of 0.601, outperforming the lexicon-based approaches but falling short of the machine learning models. On the other hand, the FLAIR zero-shot model performed poorly, with an accuracy of 0.45 and an F1-score of 0.454, only slightly better than the lexical-based models.

We can see from Table 8 that the zero-shot learning models do not perform as well on the Financial dataset compared to the other datasets. We need to recall that, as presented in Table 10, the text in the Financial dataset is more complex than the other datasets. Therefore, the complexity of the text seems to influence the performance of the zero-shot models.

We further analyzed the confusion matrices for the BERT + XGBoost and FLAIR models to gain deeper insights into their classification performance.

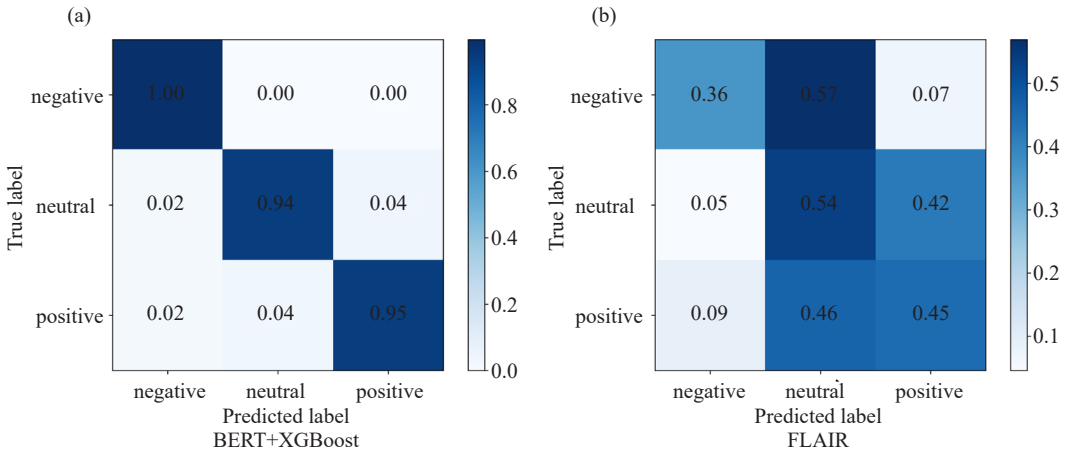

Figure 5. Three-class normalized confusion matrices for the Financial dataset

Figure 5a shows BERT + XGBoost's exceptional accuracy in identifying negative sentiment, with near-perfect clas- sification and minimal errors. Neutral sentiment is also classified with high precision, though some confusion with positive sentiment occurs. The positive class performs well but has the most room for improvement, with occasional misclassifications as negative or neutral. Notably, the model rarely confuses negative and positive sentiments, a valuable trait in sentiment analysis. Overall accuracy exceeds 96%, suggesting strong real-world applicability, especially for detecting negative sentiment.

In contrast, Figure 5b reveals FLAIR's bias toward the neutral class, with frequent misclassifications of negative and positive sentiments as neutral. This indicates the model struggles to differentiate neutral sentiment from positive and negative emotions effectively. This is exemplified by the following example, which is labeled as negative in the dataset but misclassified as neutral: *'Coca-Cola was the market leader of manufacturers with a market share of 36.9%,*

down 2.2% from the corresponding period in 2004-2005'. The misclassification likely stems from the model's confusion between the positive connotation of 'market leader' and the negative indication of the market share decline, leading it to default to a neutral classification.

Particularly striking is the high rate of confusion between neutral and positive sentiments, indicating a potential difficulty in differentiating subtle positive cues from neutral language. Interestingly, the model shows the least confusion between negative and positive sentiments, suggesting it can more easily distinguish these polar opposites. These findings highlight the challenges in sentiment analysis, especially in capturing the gradations between sentiment categories.

From these results, we can see that zero-shot models perform well in datasets like Yelp and Amazon because these datasets contain clear, distinct sentiment labels (positive/negative). However, they struggle with the Financial dataset due to its nuanced language, domain-specific terminology, and the presence of neutral statements, making classification more ambiguous. Similar challenges are observed in the Movie Reviews dataset, where sarcasm and negation reduce performance. BERT + XGBoost outperforms zero-shot models in ternary classification due to its ability to fine-tune domain-specific labeled data. While zero-shot approaches rely on broad, pre-trained knowledge, BERT + XGBoost learns dataset-specific patterns, improving performance where sentiment is subtle or multi-dimensional. Its strength in handling complex sentence structures and contextual dependencies makes it particularly effective in distinguishing among three classes.

5. Discussion

Several insights can be drawn regarding the efficiency of sentiment analysis models in both binary and ternary sentiment classification scenarios. The study reveals significant differences in performance across various models, including lexicon-based, machine learning, and zero-shot learning models.

In binary sentiment classification tasks, zero-shot learning models (i.e., NLI and FLAIR) generally outperformed the other models, demonstrating their potential to classify sentiment accurately without the need for domain-specific training. This capability is particularly valuable as it highlights the adaptability and broad applicability of zero-shot models across different domains. However, for the Financial dataset, the BERT + XGBoost model exhibited superior performance across all metrics. This underscores the robustness and effectiveness of combining pre-trained language models with powerful classifiers for handling domain-specific texts. In contrast, lexicon-based models like VADER and TextBlob showed significantly lower performance, indicating that these simpler models may not capture the nuances and complexities of sentiment in different texts as effectively as more advanced machine learning and zero-shot models.

In ternary sentiment classification scenarios, the performance of the models varied significantly. The BERT + XGBoost model maintained high performance and consistent scores across all metrics for the Financial dataset, highlighting its effectiveness even in more complex classification tasks involving three sentiment classes. However, the performance of zero-shot models like FLAIR significantly declined in these scenarios, achieving low accuracy and F1-score for the financial texts. This drop in performance suggests that while zero-shot models are versatile and generally perform well in binary sentiment classification, they face challenges when dealing with more nuanced and complex classification tasks involving multiple sentiment classes. Despite this, in non-Financial datasets, zero-shot learning models continued to outperform other models, demonstrating their versatility and generalization capabilities across various domains.

We conducted an analysis of text complexity across four datasets, which is essential for understanding how these complexities might correlate with model performance, particularly in sentiment classification tasks. By examining these factors, we can gain insights into how text complexity influences the effectiveness of the models. Table 9 summarizes the complexity of four datasets by analyzing the average number of words in texts with negative, neutral, and positive sentiments, as well as the average number of digits in the texts. The Financial dataset stands out for having the highest average word count across all sentiments, particularly with neutral texts averaging 24.88 words, suggesting that financial texts may be more verbose and complex. In contrast, the Amazon and Yelp datasets are simpler, with average word counts below 12 and no neutral text values reported. Interestingly, the Financial dataset also has a notably high average number of digits. This highlights a clear distinction in the nature and complexity of the financial dataset compared to consumer review datasets.

Table 9. Summary of the complexity of each dataset

	Negative text avg. words	Neutral text avg. words	Positive text avg. words	Avg. digits in text
Movie reviews	19.07	18.09	18.78	0.119
Amazon	10.58	-	9.91	0.219
Yelp	11.49	-	10.29	0.125
Financial	20.93	24.88	24.79	6.612

Table 10. Readability for the datasets

	Flesch reading ease	ARI
Movie reviews	62.23	9.39
Amazon	79.175	4.865
Yelp	81.251	4.544
Financial	57.839	10.52

Table 10 shows the readability level of the preprocessed texts from the test datasets using two commonly used metrics: Flesch Reading Ease and the Automated Readability Index.

The Flesch Reading Ease metric [68] is a well-established measure used to evaluate the readability of English texts. This metric calculates a readability score based on the average number of syllables per word and the average number of words per sentence. Scores range from 0 to 100, with higher scores indicating easier readability. Texts with scores between 60 and 70 are considered easily understandable by 8th to 9th graders, whereas lower scores signify more complex and harder-to-read texts. In our study, we applied the Flesch Reading Ease metric to all datasets to compare their readability levels.

Similarly, the Automated Readability Index (ARI) [69] is another widely-used readability metric that assesses the understandability of texts. Unlike the Flesch Reading Ease metric, ARI is calculated based on the number of characters per word and the number of words per sentence. The resulting score corresponds to the U.S. grade level required to comprehend the text. For example, an ARI score of 10 suggests that a 10th grader would be able to understand the content. Therefore, higher scores represent more complex texts.

The scores we obtained for the datasets indicate that financial texts are generally more difficult to read, as evidenced by their lower Flesch and higher ARI scores compared to the other datasets. These results are important as text complexity can impact the models' ability to understand the text and make accurate sentiment classifications.

The complexity of the financial dataset appears to significantly impact the performance of zero-shot learning models. Financial texts often contain specialized jargon, technical terminology, and nuanced language that are challenging to interpret without domain-specific knowledge. Market indicators like "bearish" or "bullish" carry specific sentiment connotations that may not align with their general language usage. Percentage changes and numerical comparisons (e.g., "down 2.2%", "increased by 1.5%") require contextual understanding (e.g., a small percentage decrease might signal significant negative sentiment in one market context but be relatively neutral in another). Additionally, financial texts frequently combine factual corporate achievements ("market leader", "largest provider") with performance metrics that may contradict the apparent sentiment. The combination of domain-specific terminology, context-dependent numerical interpretations, and mixed sentiment signals makes financial texts particularly challenging.

This increased complexity can lead to difficulties in sentiment classification. Lexicon-based models, for example, struggle with nuanced sentiments since they rely on predefined word lists and lack the ability to understand context, sarcasm, or subtle emotional cues. These models often misinterpret domain-specific terminology, as sentiment polarity

can vary across fields. Additionally, mixed sentiment signals within a sentence pose a challenge, as lexicon-based models typically assign a single sentiment score without accounting for opposing sentiments.

Machine learning models, on the other hand, may underperform on simpler tasks due to their over-complicated architecture. The use of high-dimensional embeddings or static features like TF-IDF can struggle to capture nuances such as sarcasm or negation. Our results show that zero-shot models, while efficient for binary tasks, often struggle with complex sentiment analysis, particularly in ternary classification tasks. This is due to their reliance on generalized understanding without specific training on domain-specific data, such as financial language.

While zero-shot models perform well in binary sentiment classification across various domains, their effectiveness diminishes in ternary classification scenarios, especially with complex texts like those in financial datasets. In contrast, hybrid models, such as BERT + XGBoost, demonstrate strong performance in both binary and ternary sentiment classification tasks by combining pre-trained language models with powerful classifiers.

Despite zero-shot models having limitations in nuanced sentiment classification, future research could explore fine-tuning these models with domain-specific data or incorporating external knowledge sources to improve their contextual understanding. Additionally, hybrid approaches that combine zero-shot learning with task-specific classifiers may further enhance performance in multi-class sentiment analysis.

However, model selection is not solely determined by performance considerations. Machine learning models require substantial computational resources for training due to their reliance on labeled datasets and iterative optimization processes. While zero-shot models eliminate the need for task-specific training, their large parameter sizes (e.g., billions of weights) and high inference-time computational demands can restrict their deployment in resource-constrained environments. Lexicon-based models, although less accurate for nuanced tasks, remain highly relevant for lightweight, real-time applications where speed, simplicity, and minimal infrastructure are prioritized over state-of-the-art accuracy. These trade-offs underscore the importance of aligning model selection with use-case constraints, balancing computational costs, latency, and performance requirements.

6. Conclusions

Our study presents the challenges modern NLP techniques face in sentiment classification, especially when dealing with complex tasks involving multiple sentiment categories. We compared six different sentiment analysis models, across binary and ternary sentiment classification tasks, including simple lexicon-based methods, machine learning approaches, and advanced zero-shot learning models that can work without specific training. Our findings show that zero-shot models (NLI, FLAIR) excel in binary classification due to their adaptability without domain-specific training. However, BERT + XGBoost outperformed them on financial data, highlighting the strength of combining pre-trained models with specialized classifiers.

In ternary classification scenarios, the performance of zero-shot models declined significantly, particularly with complex texts like those found in financial datasets, indicating the challenges these models face in more nuanced classification tasks. This performance drop in more complex tasks reveals a key weakness in current NLP methods. Despite recent progress, these models still have trouble understanding subtle differences in sentiment and dealing with specialized language in specific fields. This gap between simple and complex sentiment analysis tasks points to the need for further improvements in NLP techniques. Although these limitations, zero-shot models are valuable across industries like customer service, market research, healthcare, finance, and more, offering quick sentiment analysis and content categorization. They enable businesses to analyze data without extensive domain-specific training, making them ideal for dynamic environments and large datasets.

Our findings also underscore broader implications for NLP model development beyond sentiment analysis, particularly in areas such as text summarization. Sentiment analysis enhances summarization by improving relevance, focus, and emotional context. It prioritizes strong opinions, balances biased language, and prevents misleading interpretations of sarcasm. However, over-reliance on sentiment labels can exclude important neutral content, such as factual details in news. When accurately applied, sentiment-aware summarization creates more actionable insights for tasks like reputation management and market analysis while maintaining a balance between emotional tone and factual completeness.

In future work, we aim to enhance zero-shot learning models to better handle complex, domain-specific texts, ensuring more accurate and reliable sentiment analysis across diverse datasets. This will involve developing specialized methods for capturing and interpreting jargon, technical terminology, and nuanced language, potentially through fine-tuning existing models with domain-specific data or creating custom embeddings tailored to specific fields. Additionally, we plan to explore hybrid approaches that combine the strengths of zero-shot learning with domain-specific training, such as integrating pre-trained models like BERT with domain-specific classifiers (e.g., XGBoost or domain-adapted neural networks), which could yield more robust models for complex tasks. We also intend to focus on augmenting domain-specific data using techniques like data augmentation and synthetic data generation to simulate various sentiment expressions in specialized domains. Further, we will improve context modeling by leveraging advanced transformer-based architectures with attention mechanisms to better capture contextual cues for nuanced sentiment analysis. Finally, we aim to create new evaluation metrics that assess models' sensitivity to subtle sentiment changes and their ability to correctly classify ambiguous or domain-specific sentiments. These advancements will improve sentiment analysis in fields like finance, healthcare, and legal domains, where precise interpretation of complex texts is essential.

In conclusion, while sentiment analysis has advanced in handling basic tasks, its effectiveness in complex scenarios remains limited, particularly in managing ambiguity, domain-specific language, and nuanced emotions. Addressing these challenges requires models that adapt to specialized contexts, integrating domain-aware training and robust interpretability. Enhancing adaptability will be key to developing sentiment analysis tools that not only perform well across diverse fields, such as finance, healthcare, and social media but also provide deeper, context-sensitive insights essential for high-stakes decision-making.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Wankhade M, Rao ACS, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*. 2022; 55(7): 5731-5780.

[2] Arya V, Mishra AKM, González Briones A. *Analysis of Sentiments on The Onset of COVID-19 using Machine Learning Techniques*. Ediciones Universidad de Salamanca; 2022.

[3] Lossio-Ventura JA, Weger R, Lee AY, Guinee EP, Chung J, Atlas L, et al. A comparison of ChatGPT and fine-tuned Open Pre-Trained Transformers (OPT) against widely used sentiment analysis tools: sentiment analysis of COVID-19 survey data. *JMIR Mental Health*. 2024; 11: e50150.

[4] Weger R, Lossio-Ventura JA, Rose-McCandlish M, Shaw JS, Sinclair S, Pereira F, et al. Trends in language use during the COVID-19 pandemic and relationship between language use and mental health: text analysis based on free responses from a longitudinal study. *JMIR Mental Health*. 2023; 10(1): e40899.

[5] Birjali M, Kasri M, Beni-Hssane A. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*. 2021; 226: 107134.

[6] Mehta P, Pandya S. A review on sentiment analysis methodologies, practices and applications. *International Journal of Scientific and Technology Research*. 2020; 9(2): 601-609.

[7] Sudhir P, Suresh VD. Comparative study of various approaches, applications and classifiers for sentiment analysis. *Global Transitions Proceedings*. 2021; 2(2): 205-211.

[8] Hatzivassiloglou V, McKeown K. Predicting the semantic orientation of adjectives. In: *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics; 1997. p.174-181.

[9] Vermeij M. The orientation of user opinions through adverbs, verbs and nouns. In: *3rd Twente Student Conference on IT, Enschede June*. Citeseer; 2005.

[10] Benamara F, Cesarano C, Picariello A, Recupero DR, Subrahmanian VS. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. *ICWSM*. 2007; 7: 203-206.

[11] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. *Ain Shams*

Engineering Journal. 2014; 5(4): 1093-1113.
[12] Ahmad M, Aftab S, Muhammad SS, Ahmad S. Machine learning techniques for sentiment analysis: A review. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*. 2017; 8(3): 27.
[13] Pushp PK, Srivastava MM. Train once, test anywhere: Zero-shot learning for text classification. *arXiv: 171205972*. 2017. Available from: https://arxiv.org/abs/1712.05972.
[14] Corchado JM, López S, Garcia R, Chamoso P. Generative artificial intelligence: Fundamentals. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*. 2023; 12(1): e31704.
[15] Koppel M, Schler J. The importance of neutral examples for learning sentiment. *Computational Intelligence*. 2006; 22(2): 100-109.
[16] Hamed A, Qiu R, Li D. The importance of neutral class in sentiment analysis of Arabic tweets. *International Journal of Computer Science and Information Technology*. 2016; 8: 17-31.
[17] Saura JR, Palos-Sanchez P, Grilo A. Detecting indicators for startup business success: Sentiment analysis using text data mining. *Sustainability*. 2019; 11(3): 917.
[18] Liu B. *Sentiment Analysis and Opinion Mining*. Springer Nature; 2022.
[19] Mæhlum P, Samuel D, Norman RM, Jelin E, Bjertnæs ØA, Øvrelid L, et al. It's difficult to be neutral-human and LLM-based sentiment annotation of patient comments. *arXiv: 240418832*. 2024. Available from: https://arxiv.org/abs/2404.18832.
[20] Hutto C, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International Aaai Conference on Web and Social Media*. 2014; 8: 216-225.
[21] Valdez D, Ten Thij M, Bathina K, Rutter LA, Bollen J. Social media insights into US mental health during the COVID-19 pandemic: Longitudinal analysis of Twitter data. *Journal of Medical Internet Research*. 2020; 22(12): e21418.
[22] Al Mansoori S, Almansoori A, Alshamsi M, Salloum SA, Shaalan K. Suspicious activity detection of twitter and facebook using sentimental analysis. *TEM Journal*. 2020; 9(4): 1313.
[23] Scholz J, Jeznik J. Evaluating geo-tagged twitter data to analyze tourist flows in Styria, Austria. *ISPRS International Journal of Geo-Information*. 2020; 9(11): 681.
[24] Isnan M, Elwirehardja GN, Pardamean B. Sentiment analysis for TikTok review using VADER sentiment and SVM model. *Procedia Computer Science*. 2023; 227: 168-175.
[25] Bonta V, Kumaresh N, Janardhan N. A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*. 2019; 8(S2): 1-6.
[26] Illia F, Eugenia MP, Rutba SA. Sentiment analysis on pedulilindungi application using textblob and vader library. *Proceedings of The International Conference on Data Science and Official Statistics*. 2021; 2021(1): 278-288.
[27] Diyasa IGSM, Mandenni NMIM, Fachrurrozi MI, Pradika SI, Manab KRN, Sasmita NR. Twitter sentiment analysis as an evaluation and service base on python textblob. In: *IOP Conference Series: Materials Science and Engineering*. IOP Publishing; 2021. p.012034.
[28] Chandrasekaran G, Hemanth J. Deep learning and TextBlob based sentiment analysis for coronavirus (COVID-19) using twitter data. *International Journal on Artificial Intelligence Tools*. 2022; 31(01): 2250011.
[29] Braig N, Benz A, Voth S, Breitenbach J, Buettner R. Machine learning techniques for sentiment analysis of COVID-19-related twitter data. *IEEE Access*. 2023; 11: 14778-14803.
[30] Cambria E, Schuller B, Liu B, Wang H, Havasi C. Statistical approaches to concept-level sentiment analysis. *IEEE Intelligent Systems*. 2013; 28(3): 6-9.
[31] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. *arXiv:1310.4546.* 2013. Available from: https://doi.org/10.48550/arXiv.1310.4546.
[32] Bouabdallaoui I, Guerouate F, Sbihi M. Hybrid text embedding and evolutionary algorithm approach for topic clustering in online discussion forums. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*. 2024; 13: e31448.
[33] Islam MS, Kabir MN, Ghani NA, Zamli KZ, Zulkifli NSA, Rahman MM, et al. Challenges and future in deep learning for sentiment analysis: A comprehensive review and a proposed novel hybrid approach. *Artificial Intelligence Review*. 2024; 57(3): 62.
[34] Kumar S, Roy PP, Dogra DP, Kim BG. A comprehensive review on sentiment analysis: Tasks, approaches and applications. *arXiv: 231111250*. 2023. Available from: https://arxiv.org/abs/2311.11250.
[35] Singh N. Sarcasm text detection on news headlines using novel hybrid machine learning techniques. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*. 2024; 13: e31601.

[36] Afifah K, Yulita IN, Sarathan I. Sentiment analysis on telemedicine app reviews using xgboost classifier. In: *2021 International Conference on Artificial Intelligence and Big Data Analytics*. IEEE; 2021. p.22-27.
[37] Shakya SR, Ceh-Varela E, Shakya S, Parten C, Zhou Z. Boosting stock predictions with sentiment analysis and deep learning models. In: *2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC)*. IEEE; 2025. p. 1-6.
[38] Ahuja R, Chug A, Kohli S, Gupta S, Ahuja P. The impact of features extraction on the sentiment analysis. *Procedia Computer Science*. 2019; 152: 341-348.
[39] Rana S, Kanji R, Jain S. Comparison of SVM and naïve bayes for sentiment classification using BERT data. In: *2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)*. IEEE; 2022. p.1-5.
[40] Wang Z, Xie Q, Feng Y, Ding Z, Yang Z, Xia R. Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv: 230404339*. 2023. Available from: https://arxiv.org/abs/2304.04339.
[41] Ceh-Varela E, Imhmed E, Chavez J. Comparative analysis of methods for sentiment labeling with limited labeled data. In: *2024 17th International Conference on Advanced Computer Theory and Engineering (ICACTE)*. IEEE; 2024. p.140-144.
[42] Chaisen T, Charoenkwan P, Kim CG, Thiengburanathum P. *A Zero-Shot Interpretable Framework for Sentiment Polarity Extraction*. IEEE Access; 2023.
[43] Zhang J, Lertvittayakumjorn P, Guo Y. Integrating semantic knowledge to tackle zero-shot text classification. *arXiv: 190312626*. 2019. Available from: https://arxiv.org/abs/1903.12626.
[44] Pelicon A, Pranjić M, Miljković D, Škrlj B, Pollak S. Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*. 2020; 10(17): 5993.
[45] Shu L, Xu H, Liu B, Chen J. Zero-shot aspect-based sentiment analysis. *arXiv: 220201924*. 2022. Available from: https://arxiv.org/abs/2202.01924.
[46] Kumar P, Pathania K, Raman B. Zero-shot learning based cross-lingual sentiment analysis for sanskrit text with insufficient labeled data. *Applied Intelligence*. 2023; 53(9): 10096-10113.
[47] Tesfagergish SG, Kapočiūtė-Dzikienė J, Damaševičius R. Zero-shot emotion detection for semi-supervised sentiment analysis using sentence transformers and ensemble learning. *Applied Sciences*. 2022; 12(17): 8662.
[48] Giachanou A, Crestani F. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*. 2016; 49(2): 1-41.
[49] Ceh-Varela E, Maes L, Shakya SR. Machine learning analysis of factors contributing to Diabetes Development. *Cloud Computing and Data Science*. 2024; 5(1): 157-182.
[50] Pang B, Lee L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*. 2005. Available from: https://arxiv.org/abs/cs/0506075.
[51] Kotzias D, Denil M, De Freitas N, Smyth P. From group to individual labels using deep features. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery; 2015. p.597-606.
[52] McAuley J, Leskovec J. Hidden factors and hidden topics: Understanding rating dimensions with review text. In: *Proceedings of the 7th ACM Conference on Recommender Systems*. Association for Computing Machinery; 2013. p.165-172.
[53] Malo P, Sinha A, Korhonen P, Wallenius J, Takala P. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*. 2014; 65(4): 782-796.
[54] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery; 2016. p.785-794.
[55] Liu H, Chen X, Liu X. A study of the application of weight distributing method combining sentiment dictionary and TF-IDF for text sentiment analysis. *IEEE Access*. 2022; 10: 32280-32289.
[56] Chiny M, Chihab M, Bencharef O, Chihab Y. LSTM, VADER and TF-IDF based hybrid sentiment analysis model. *International Journal of Advanced Computer Science and Applications*. 2021; 12(7): 256-275.
[57] Hoang M, Bihorac OA, Rouces J. Aspect-based sentiment analysis using bert. In: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Linköping University Electronic Press; 2019. p.187-196.
[58] Lin HY, Moh TS. Sentiment analysis on COVID tweets using COVID-Twitter-BERT with auxiliary sentence approach. In: *Proceedings of the 2021 ACM Southeast Conference*. Association for Computing Machinery; 2021: p.234-238.
[59] Roelleke T, Wang J. TF-IDF uncovered: A study of theories and probabilities. In: *Proceedings of the 31st Annual*

International Acm Sigir Conference on Research and Development in Information Retrieval. Association for Computing Machinery; 2008. p.435-442.
[60] Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv: 190810084*. 2019. Available from: https://arxiv.org/abs/1908.10084.
[61] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv: 181004805*. 2018. Available from: https://arxiv.org/abs/1810.04805.
[62] Yin W, Hay J, Roth D. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv:190900161*. 2019. Available from: https://arxiv.org/abs/1909.00161.
[63] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv: 191013461*. 2019. Available from: https://arxiv.org/abs/1910.13461.
[64] Halder K, Akbik A, Krapac J, Vollgraf R. Task-aware representation of sentences for generic text classification. In: *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics; 2020. p.3202-3213.
[65] Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, Vollgraf R. FLAIR: An easy-to-use framework for state-of-the-art NLP. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Association for Computational Linguistics; 2019. p.54-59.
[66] Effrosynidis D, Symeonidis S, Arampatzis A. A comparison of pre-processing techniques for twitter sentiment analysis. In: *Research and Advanced Technology for Digital Libraries: 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings 21*. Springer International Publishing; 2017. p.394-406.
[67] Saripilli P, Aruna Kumari G, Vannemreddy CS, Shaik K, Madishetty S. A sentiment analysis of tweets by using TF-IDF vectorizer and lemmatization with POS tagging. In: *XVIII International Conference on Data Science and Intelligent Analysis of Information*. Springer; 2023. p.377-386.
[68] Farr JN, Jenkins JJ, Paterson DG. Simplification of Flesch reading ease formula. *Journal Of Applied Psychology*. 1951; 35(5): 333.
[69] Kincaid JP, Delionbach LJ. Validation of the automated readability index: A follow-up. *Human Factors*. 1973; 15(1): 17-20.