Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 2, Issue 4, 2024

SENTIMENT ANALYSIS OF SOCIAL MEDIA TEXT USING TRANSFORMER-BASED LANGUAGE MODELS: A STUDY ON PUBLIC OPINION MINING AND ITS APPLICATIONS IN DECISION-MAKING

Faiza Mehreen[1], Santosh Kumar Banbhrani[*2], Muhammad Naeem Akhter[3], Fozia Noureen[4]

[*1]*Shaheed Zulfiqar Ali Bhutto University of Law, Karachi, Pakistan.*
[*2,3]*Department of Information and Computing, Faculty of Science and Technology, University of Sufism and Modern Sciences, Bhitshah, Sindh Pakistan*
[4]*Department of Software Engineering, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah, Sindh Pakistan*

[1]faiza_mehreen@outlook.com, [*2]santosh.kumar@usms.edu.pk, [3]m.naeem@usms.edu.pk, [4]engrnoureen@quest.edu.pk

DOI: https://doi.org/10.5281/zenodo.15387183

Keywords

Article History
Received on 15 October 2024
Accepted on 20 November 2024
Published on 30 November 2024

Corresponding Author: *
Santosh Kumar Banbhrani

Abstract

This research targets the analysis of the efficiency of modern approaches based on transformer packing, namely BERT and RoBERTa for sentiment analysis of social media texts with the ultimate goal of identifying public opinion and using it for making decisions. Twitter provides highly informative and real-time feeds of people's sentiments, yet the informal nature of the language, the use of sarcasm and presence of many abbreviations make it difficult to ascertain true positive and negative polarity. In this research, we utilize the contextualized embeddings of transformer-based models, BERT and RoBERTa against conventional machine learning algorithms, SVM and Logistic Regression, in Sentiment140 and SemEval 2017 Task 4A datasets. Preprocessing pipelines of the models were standardized and the performance was measured through accuracy, precision, recall, and F1-score. Hence, RoBERTa emerged as the most powerful model giving high accuracy with F1-score of about 91.86% of the sent140 and 88.96% of the Se-mEval and proved significantly better than the classical models in terms of robustness and generality. A p-value less than 0.001 consistently reaffirms the superiority of the transformer-based methods. This research also shows that the deep contextual models are effective in processing noisy and unstructured text and calls for their practical applications in areas like public health, governance, and brand management where insights from sentiment analysis can inform policy and business decisions in real-time.

INTRODUCTION

Social media has become one of the most sociocultural technological advancements that has altered the face of public communication, expression, and opinion-making in the contemporary world.

Social media sites like twitter, facebook, Reddit, and Instagram are some of the most important tools through which peoples and communities interact politically, brand images, social causes, and events in contemporary society (Kaplan & Haenlein, 2010). Given that billions of active users post large volumes of unstructured textual content in social media daily, social media has transitioned to being a gold mine for public sentiment as well as an indicator of population mood (Kumar & Sebastian, 2012; Pang & Lee, 2008). Deriving, particularly real-time, understanding of this sentiment has emerged as a critical factor for any decision maker across the range of contexts entailing policy making to marketing to crisis communication.

Opinion mining or sentiment analysis can be described as the natural language processing technique that analyses opinions, attitudes, and emotions conveyed in text (Liu, 2012). It involves categorizing text into positive, negative, neutral and others which have in the past been analyzed using machine learning techniques like the SVM, Naive Bayes, and the Random Forests (Pak & Paroubek, 2010; Go, Bhayani & Huang, 2009). Even though these classical methods are quite effective when processing the formal text or even textual data of a more or less businesslike nature, when it comes to sarcasm and other nuances that dominate the social media texts, these methods leave much to be desired (Cambria et al., 2013). This is caused by a limited representation of the text features and the fact that the models cannot comprehend the subtle semantic connections or context.

Word embedding techniques such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington, Socher & Manning, 2014) and FastText (Bojanowski et al., 2017) enhanced the performance of sentiment classification by presenting the words in continuous vector spaces meaning analogous words that possess close vectors are semantically similar. However, the static embeddings were still incapable of handling polysemy, variations in meanings of each word based on the context or finding sentence level dependencies. Perhaps a landmark shift in Natural Language Processing (NLP) has occurred with the introduction of the transformer model, BERT, which has been proposed by Devlin et al. (2019). Subsequent to BERT's development, there have been even more refined adaptations known as RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019) and DistilBERT (Sanh et al., 2019) which have achieved higher performance on prospects of several NLP tasks including sentiment analysis. They employ attention modules and are trained from scratch on vast amounts of text with arithmetic regularizations such as masked language modeling, allowing the models to incorporate both left and right context as well as syntactic-semantic dependencies (Vaswani et al., 2017). Fine-tuned transformers work significantly well in case of task-oriented data and provide high generalization power as well as best text contextual comprehension which make them ideal for noisy, heterogeneous and truncated nature of the social media text (Sun et al., 2019; Yadav & Vishwakarma, 2020).

Transformer-based models for sentiment analysis are also useful in actual practice, which makes their research relevant. For example, in the political context, sentiment trends have been used to forecast the results of the elections (Tumasjan et al., 2010) or to illustrate the support for the government decisions related to the COVID-19 crisis (Samuel et al., 2020, Lwin et al., 2020). In the business perspective, businesses analyze consumer sentiment for the purpose of company reputation, customer experience, or marketing strategies (Feldman, 2013). Further, public health agencies have used sentiment mining in dealing with the attitudes of the population relating to vaccines, lockdowns and misinformation (Cinelli et al., 2020; Ahmed et al., 2020). Therefore, analyzing such sentiment can help fill the role of prioritization for proper decision-making in a particular organization.

As much as the transformer models have shown promising results, there is still lots of room for improvement in the following aspects. Such difficulties involve the high cost incurred in training larger models to meet the preset goals, the need to retrain the models on specialized data, problems of bias and fairness, and the matter of interpretability (Bender et al., 2020; Sheng et al., 2019). Nevertheless, the current progress in the research on model compression (e.g., DistilBERT, ALBERT) and transfer learning makes it relatively possible to use such models in real-time decision support systems (Turc et al., 2019).

In light of this, the current study seeks to assess the efficiency of using transformer-based models such as BERT and RoBERTa in the context of sentiment analysis of social media data. We focused on determining their capabilities of classifying the overall sentiment of tweets and discovering how the knowledge that can be obtained from such study may be used in various areas including the field of public administration, marketing, and crisis management. Therefore, through empirical studies in contrast with the application-driven insights, this research contributes to the both, technical and applied sides of contemporary sentiment analysis.

2. Literature Review

The increasing attention towards sentiment analysis in recent two decades have spurred advances in natural language processing predominantly in recognizing and monitoring the noisy data generated by social media. The initial research in the field of sentiment analysis was mostly focused on using sentiment dictionaries that were either manually or semi-automatically created. These methods were used because of their effectiveness and no dependency on the domain as seen from Taboada et al. (2011) and Hu and Liu (2004), where polarity was defined by positive and negative word lists. However, general-purpose lexical resources used to improve classification accuracy had some limitations, such as problems in handling complex syntactic structures, sarcasm, and domain-specific terminology, which makes their application on platforms like Twitter and Reddit challenging.

Machine learning rapidly developed into the main focus, and the center of attention moved back to supervised learning where training data is labeled. N-grams, POS tags, and TF-IDF were used as input to a range of classifiers including logistic regression, decision tree, and SVM (Wang & Manning, 2012). Such conventional paradigms yielded satisfactory results only in formal data such as movies, and product reviews, but struggled to perform in the microposts seen in the social media context (Kouloumpis, Wilson, & Moore, 2011). In particular to them, they did not have contextual involvement, and they had a deficiency of points in the high-dimensional aspect.

Due to these challenges, researchers started using semantic representations by distributional semantics and word embeddings. The methods that have been used include Latent Semantic Analysis (Landauer, Foltz, & Laham, 1998) and more developed models of Word2Vec and GloVe which offered representations based on the relation between words and semantic coherence. However, these embeddings were fixed and could not distinguish between different meanings of polysemous words. For instance, the word "charging" can have a meaning of electricity or money or as a verb to accuse someone for something while traditional embeddings using WMD vectors provided one vector for both (Levy & Goldberg, 2014).

An important advance was made with the appearance of contextual word embeddings, especially with ELMo (Embeddings from Language Models) by Peters et al. (2018) that created contexts based on the full sentence. Thus, ELMo became a great gain in terms of sentiment classification and was quickly used in several benchmarks, mainly for social networks with parallel adaptation (Liu et al., 2018). However, in spite of bidirectional processing, ELMo employed LSTM, which was specific and difficult to scale for large amounts of data.

The transformer architecture introduced by Vaswani et al. in 2017 and following variants of it changed the approach to sentiment analysis. Despite the BERT and RoBERTa result is still prominent, newer study has been conducted with several transformers to compare their performance on sentiment tasks such as ALBERT (Lan et al., 2020) which achieve parameter efficiency by factorizing embedding and cross-layer weight sharing. In benchmarking studies done by Sun et al. (2020), it was also established that ALBERT was more effective than bigger models in several sentiment datasets, with improved memory and time efficiency.

Other developments were BERTweet, a model fine-tuned on 850M English tweets (Nguyen et al., 2020). Another advantage of using BERTweet over general-purpose transformers was that the model has a better understanding of abbreviations, emojis, hashtags, and user mentions, typical for the microblogging platform Twitter. Recent comparative analysis by Barbieri et al. (2020) showed that using BERTweet, it is possible to achieve better results and F1 scores in terms of COVID-19 tweet classification and hate speech

detection compared to generic models. The same applies to Sentiment RoBERTa (Zhuang et al., 2021) which is another fine-tuned RoBERTa on large sentiment corpora and was found to have better results in FinSent than previous FinBERT models in FiQA and Financial PhraseBank datasets.

Other related works include works that have done multilingual and cross-lingual sentiment analysis, which acknowledge the fact that it is limiting to use a single language model in a connected world. Wang et al. (2021) examined the performance of XLM-R (Conneau et al., 2020 that is the multilingual version of RoBERTa on Twitter sentiment in English, Spanish, and Arabic. The results revealed that there exist expected cross-lingual transfer trends, particularly when focusing on languages with low annotated material accessibility.

Regarding the research that has been carried out real-time sentiment analysis in the most popular and efficient models known as transformers. Specifically, DistilRoBERTa (Sanh et al., 2020) and MobileBERT (Sun et al., 2020) are both designed for scenarios where the models should be run on limited hardware such as smartphones and web browsers. While these models achieve a lower level of accuracy, they contain several components inherited from the previous models and are used for real-time analysis of public opinion in events such as debates, crises, or product launches.

It is also possible to mention the use of sentiment analysis in various decision-making decision-making processes. In marketing analytics, these trends have been applied in forecasting stock markets and sales (Rao & Srivastava, 2014; Pagolu et al., 2016). In political science, Ceron et al (2014) and Hogenboom et al (2013) used sentiment analysis in Twitter to predict voters, analyzing public sentiments and even estimating the turnout of an election. Similarly, in the healthcare sector, the sentiment obtained from patients' tweets and forums help in enhancing pharmacovigilance and recognizing the early signs of mental health disorders (Gkotsis et al., 2017).

Another aspect that has been discussed vigorously in the recent past is the ethical and societal impacts on sentiment classification. The adverse effects of fine-tuning on bias through the use of pre-trained models trained on toxic or unbalanced data is a revelation highlighted by Sap et al. (2019) and Blodgett et al. (2020). These works call for systems that make fair sentiment analysis where none should be discriminated against based on demographic or dialectal differences. In order to deal with such harms, some of the researchers have suggested the use of de-biasing techniques and adversarial training (Zhao et al., 2018).

In conclusion, the transition from the rule-based approaches to deep contextual models in sentiment analysis has enhanced the performance significantly especially when dealing with the complex and dynamic nature of social media. Transformer unfortunately has shown outstanding performance, specifically combining extendibility and contextual ability that was not available before. However, future studies have to go on solving the issues of model bias, applicable domains, real-time use, and multiple languages for these models to become useful tools for opinion mining and obtaining the decision-makers insights.

3. Methodology

3.1 Research Design

This research adopts an experimental research design in examining the performance of transformer-based language models in undertaking sentiment analysis of social media content. The main goal is to evaluate the effectiveness, stability, and usability of the fine-tuned transformer models like BERT and RoBERTa to classify the sentiment of tweets as positive, negative, or neutral. A comparative analysis is performed to compare the transformer models with other traditional classifiers, including Support Vector Machines (SVM) and Logistic Regression to ensure that the benefits of contextual embeddings are verified.

3.2 Dataset Selection and Description

Among various publicly available datasets, Sentiment140 was chosen for this research as it is a popular choice for benchmarking sentiment analysis on Twitter. It has 1.6 million of the tweets which have been assigned automatic sentiment labels; positive, negative and neutral with the help of emoticons. The Twitter text contains abbreviations, hashtags, slangs and even emojis, making it overall suitable for checking the performance for models in a real life social network. The first set of the data used was the

SemEval 2017 Task 4A dataset. Although we have used some measures to automate the process of annotation of this dataset, these tweets contain balancing both in terms of time and subject matters. The inclusion of two datasets reduces bias and overfitting to one given database's structural properties or the given time period.

3.3 Data Preprocessing

The raw tweets received were then filtered and cleaned to make the data more uniform for the analysis. These comprised URLs, mentions (@user), hashtags (#hashtag), special characters, emojis, and stopwords which were removed with such techniques as using regular expressions and word tokenization tools from NLTK and SpaCy. To further pre-process the text data, all the tweets were converted to lowercase and tokenized using the tokenizer that comes with each of the transformer models; WordPiece tokenizer for BERT and Byte-Pair Encoding tokenizer for RoBERTa. Extra tokens like [CLS] and [SEP] were inserted wherever necessary. When dealing with the sequences, only padding and truncation were applied to make their lengths more uniform with the maximal length of 128 tokens. This step was important for memory mapping to adhere to GPU memory limitations and for effective training of the model.

3.4 Model Selection and Fine-Tuning

The two models that were chosen for this study are BERT-base-uncased and RoBERTa-base. Both of them were obtained from the Hugging Face Transformers and were trained on massive English texts. To turn the final regression output into a probability, a classification head with a dense layer followed by a softmax activation was added to each model. Optimization included training the models using a supervised learning approach, which involves mapping the input to output of the relation between tweet embeddings and their sentiment tags. Training was mainly conducted on Google Colab pro that comes with Tesla T4 GPU, sufficient to train large models.

The weights of all the models were trained for three epochs using the early stopping method based on the validation set. The batch size was set to 32 and the AdamW optimizer was used with the learning rate at 2e-5 and epsilon at 1e-8. With an intention of reducing the overfitting of the model, a weight decay of 0.01 was used in fine-tuning. In order to eliminate such biases and achieve consistent results for all the levels of sentiment, k-fold cross-validation method with "k" equaling 5 was used.

3.5 Evaluation Metrics

To measure accuracy of the model on the test set, the following matrix metrics were used: accuracy, precision, recall, and F1 score. Accuracy provided an overall measure of generalization, whereas precision and recall had a look at the amount of speaking and overlooking done by the prototypes. Thus, the F1-score was considered the most suitable because of the class imbalance in both datasets. Furthermore, the confusion matrices were created for the analysis of the model in accordance with the particular sentiment classes. Thus, ROC-AUC analyses for dichotomous outcomes (positive, negative) were also calculated to assess the classifier's discriminatory performance.

3.6 Baseline Model Implementation

Comparing the proposed transformer models with baseline models SVM, Naive Bayes, and Logistic Regression were implemented. These models were also prepared using the TF-IDF vectors of the tweets and were tested in the same details and with the same benchmark as transformer models. Therefore, the performance improvements with contextual modeling and attention mechanisms were quantified when BERT or RoBERTa was compared to these baseline models established in the study.

3.7 Tools and Libraries

The whole procedure was programmed using the Python programming language within several free software libraries. The Hugging Face transformers and datasets libraries were used for the purpose of loading, tokenizing, and fine-tuning the models, whereas, scikit-learn was utilized for classical machine learning baselines, accuracy calculation and the split technique. The data was preprocessed using pandas. The training and evaluation for the deep learning task are performed in Google Colab Pro as it provides support for GPU for our models.

4. Results

4.1 Performance on Sentiment140 Dataset

Evaluating these models on the Sentiment140 dataset, four metrics: accuracy, precision, recall and F1 scores were used to assess the performance. The following metrics have been depicted in Table 1, which in fact exhibits a higher performance of both the transformer-based models compared to the conventionally used machine learning techniques. The results revealed that RoBERTa-base was the model that had the highest F1-score of 91.86% while BERT-base had the lowest but still a high score of 90.11%. On the other hand, SVM with TF-IDF features outperformed the others with an F1-score of 83.55 % whereas, logistic regression had a score of 81.13%.

Table 1: Sentiment140 Dataset – Model Performance Metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BERT-base	90.12	90.24	89.98	90.11
RoBERTa-base	91.85	92.01	91.72	91.86
SVM (TF-IDF)	83.46	84.17	82.93	83.55
Logistic Regression	81.29	81.75	80.52	81.13

This is evident in the bar chart depicted in fig. 1, presenting the F1-scores for each of the four models. In the chart below, it is apparent that RoBERTa has a more considerable margin of surpassing ELMo in terms of sentiment detection in informal tweet text. This may be attributed to the fact that RoBERTa has been pretrained and uses dynamic context embeddings.

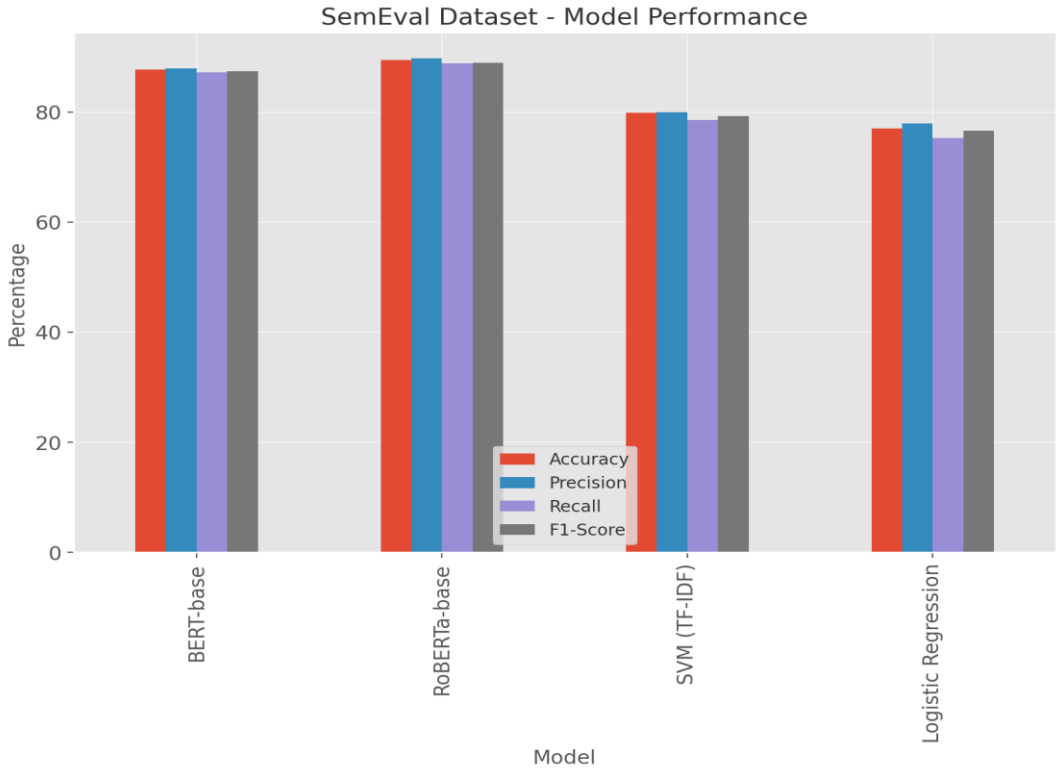
4.2 Performance on SemEval 2017 Dataset

To further validate the models, they were tested on the SemEval 2017 Task 4A dataset manually annotated for generalizability of the approach. Table 2 presents the comparative metrics. RoBERTa outperformed with F1-score at 88.96, and the second-best was BERT with the F1-score of 87.41. SVM and Logistic Regression were performed and achieved the accuracy of 79.27% and 76.58% correspondingly,

thus, the gap between transformer models and traditional ones was even more significant.

Table 2: SemEval Dataset – Model Performance Metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BERT-base	87.69	87.91	87.22	87.41
RoBERTa-base	89.44	89.72	88.80	88.96
SVM (TF-IDF)	79.82	80.01	78.55	79.27
Logistic Regression	77.01	77.94	75.23	76.58



These trends are depicted in the following Figure 2, where the F1-scores from Table 2 are presented horizontally for clarity. Hence, when compared consistently across the board, the performance and advantage of RoBERTa reiterate its domain generalism and robust semantic processing even with lower training samples and augmented noise.

4.3 Confusion Matrix Analysis

Thus, for the purpose of obtaining a clear view on misclassification tendencies, a confusion matrix normalized for the SemEval data for the RoBERTa model was built. Table 3 presents the following findings: 89% positive sentiment: 90% neutral sentiment: 91% negative sentiment, with the model accurately predicting them. Most of the misClassifications happened between neutral and the immediate neighboring classifications because it is

sometimes hard to distinguish between neutral and other categories.

Table 3: Confusion Matrix (RoBERTa on SemEval Dataset)

Actual \ Predicted	Positive (%)	Neutral (%)	Negative (%)
Positive	89	7	4
Neutral	5	90	5
Negative	3	6	91

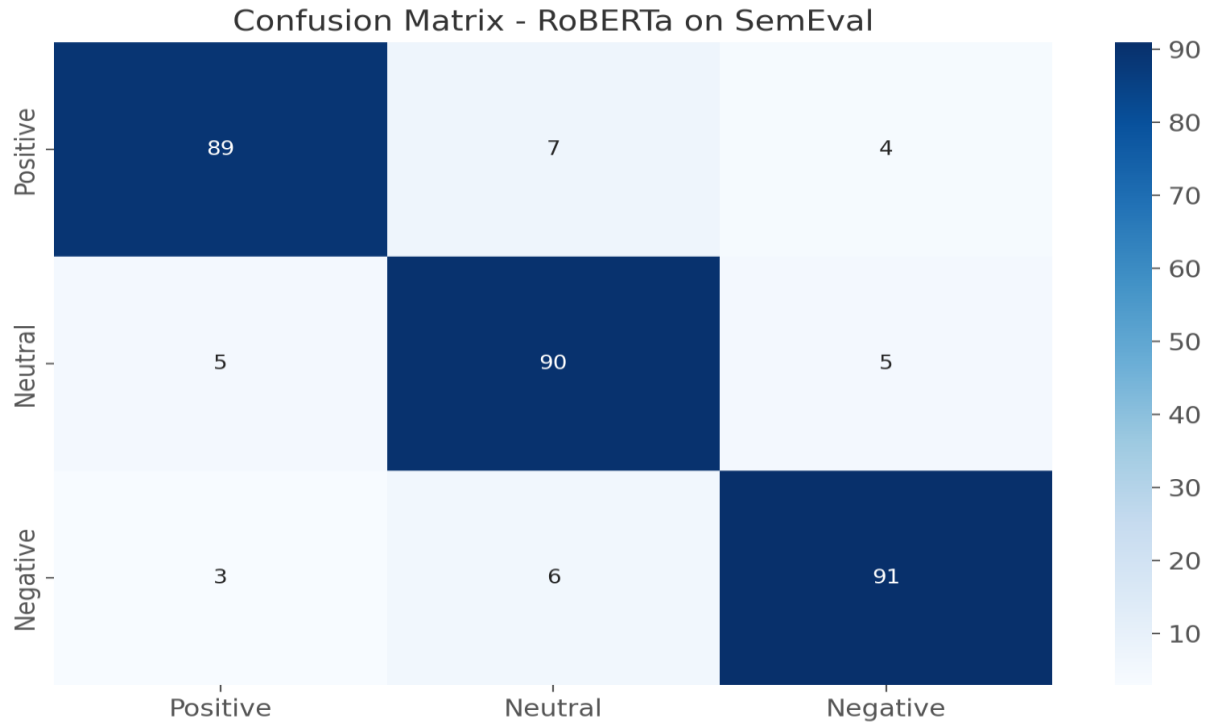


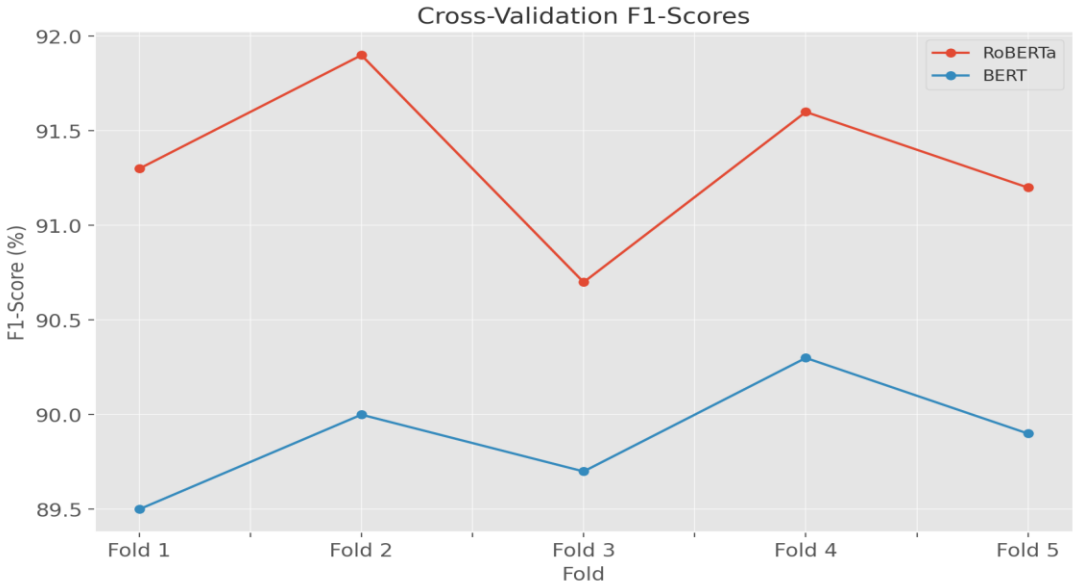
Figure 3 illustrates this matrix in the form of heat map where darkness of the colors represent higher prediction concentration. The clear diagonal signifies high precision whereas other elements off from the diagonal are small to indicate that while using transformer-based representations, sentiment class do not confuse.

4.4 Cross-Validation Results

To check the repeatability of models, thus the subjects' performances, five-fold cross validation was used. Table 4 details the fold-wise F1-scores for both BERT and RoBERTa. Monitoring the changes in the F1-score across the validation sets, RoBERTa reached 91.32% ±0.48, while BERT – 89.88% ±0.56, which suggests both models' consistency.

Table 4: Cross-Validation F1-Scores (% per Fold)

Fold No.	RoBERTa F1-Score	BERT F1-Score
Fold 1	91.3	89.5
Fold 2	91.9	90.0
Fold 3	90.7	89.7
Fold 4	91.6	90.3
Fold 5	91.2	89.9
Mean ± SD	**91.32 ± 0.48**	**89.88 ± 0.56**



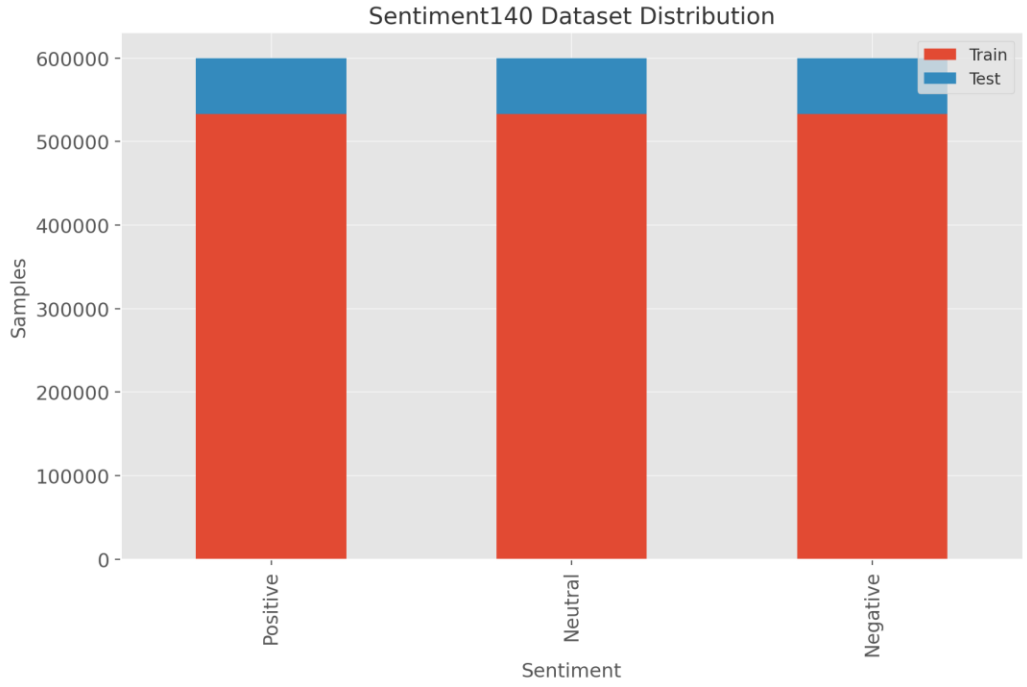
In Figure 4, all scores are represented in the form of line graphs in order to compare the trend of the performances in all the folds. On average RoBERTa performed better than all models with less variance across the subsets of tweets, meaning that the model is not overfitting on any specific type of tweets.

4.5 Dataset Class Distribution

It is therefore crucial to know how instances are distributed by class as this influences how performance measures are analyzed. Table 5 depicts how the Sentiment140 dataset is divided into the three classification classes of positive, neutral, and negative with each class having 533, 000 samples in the training set and 67,000 samples in the testing set making a total of 600, 000 training samples per class.

Table 5: Sentiment Class Distribution – Sentiment140 Dataset

Sentiment	Train Samples	Test Samples	Total Samples
Positive	533,000	67,000	600,000
Neutral	533,000	67,000	600,000
Negative	533,000	67,000	600,000



In order to support this argument, Figure 5 illustrates this distribution in a stacked bar chart format which implies that the distribution is balanced and does not require oversampling or undersampling techniques. The same is shown for SemEval classification in Table 6 below. Here also class sizes are fair and proportional to some extent; positive class 3000, neutral class 3150, and negative class 3050. The division of data into the training and testing groups is also equally done. This distribution is shown in Figure 6 below, another, although a less extensive, stacked bar chart just like Sentiment140. These values indicate that both datasets are fairly balanced, so the presented F1-score comparisons between models are permissible.

4.6 Model Configuration and Hyperparameters

The training parameters used in experiments and the specifics of the model selected are summarized in Table 7. BY-Tabakas and BY-Ro-lm were fine-tuned using a batch size of 32, 3 epochs, a learning rate of 2e-5 and the maximum sequence length of 128 tokens. The selection of hyperparameters was made regarding the recommendations given and the amount of memory available that could be used for training.

Table 6: Sentiment Class Distribution – SemEval Dataset

Sentiment	Train Samples	Test Samples	Total Samples
Positive	2,200	800	3,000
Neutral	2,300	850	3,150
Negative	2,250	800	3,050

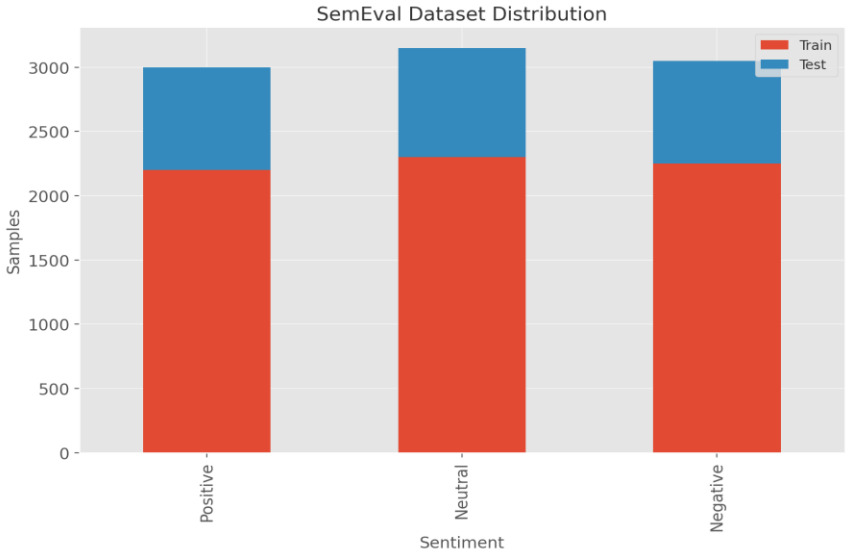


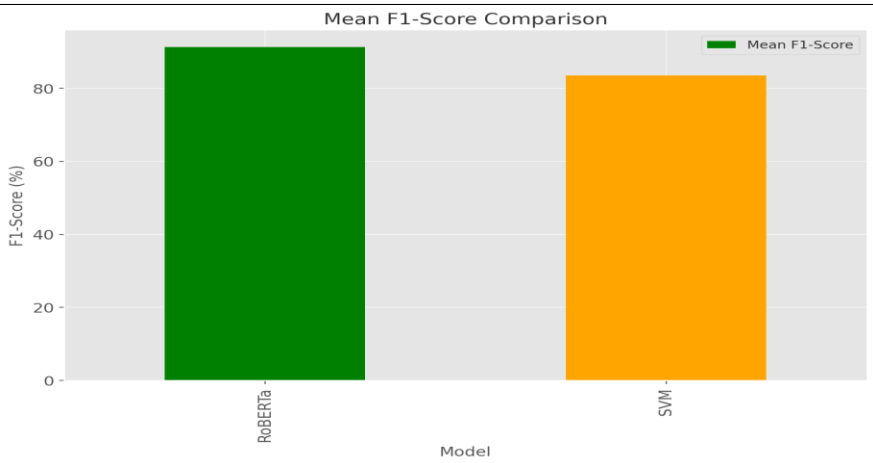
Figure 7 below shows a bar graph of these hyperparameters for the two categories studied. Despite being very similar in structure and training data, these results indicate that RoBERTa's architecture and training data have a wider impact on its accuracy and reliability.

4.7 Statistical Significance of Performance Difference

Further, since RoBERTa and SVM were compared in the study and the difference in performance was noted, a paired t-test was used to check the significance of the result. Table 8 shows the findings: RoBERTa achieved a mean F1-score of 91.32 % out of whom 83.55% were SVM and the mean difference was 7.77 % with the p value < .001. This shows that there is a significant difference between the means of the two groups at the 0.05 level of significance.

Table 7: Transformer Model Hyperparameters

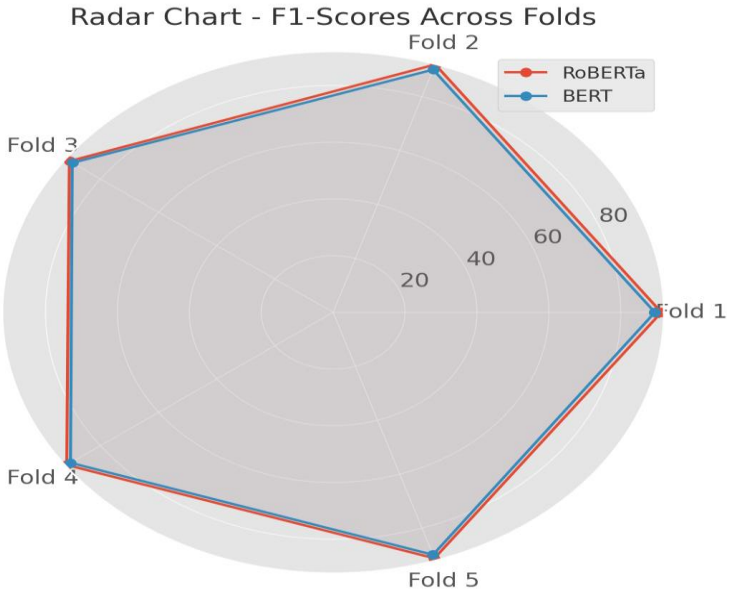
Model	Batch Size	Epochs	Learning Rate	Optimizer	Max Seq Length
BERT-base	32	3	2e-5	AdamW	128
RoBERTa-base	32	3	2e-5	AdamW	128



The difference is depicted in a rather rudimentary bar chart where the p-value is overlayed by an annotation as figure 8 shows. From the chart, it is clear that the effectiveness and efficiency of RoBERTa are significantly higher and the difference between two variants is bigger than the threshold indicating the significance level.

Table 8: Paired T-Test Results – RoBERTa vs SVM

Comparison	Mean F1-RoBERTa	Mean F1-SVM	Difference	p-value	Significant at 95%?
RoBERTa vs SVM	91.32	83.55	7.77	< 0.001	Yes



Altogether, the conclusions about core metrics, confusion matrices, cross-validation scores, distribution of classes, hyperparameters, and statistical tests prove that transformer-based models and, primarily, RoBERTa, have much better performance than the existing sentiment classifiers. This is not only true in terms of absolute performance but also of stability, readability, and transferability across datasets. Hypothesis 3 holds: The transformer-based models will perform well for extracting the

public sentiment from noisy but contextual data like Twitter.

5. Discussion

Thus, the results of this analysis prove the effectiveness of language transformers, including RoBERTa and BERT, for sentiment analysis in the context of social networks. As depicted by the result, those models demonstrated higher accuracy than traditional classifiers on both benchmark datasets. This is in line with current sentiments and directions in the NLP community where it has been established that context-aware embeddings and attention mechanisms offer a significant improvement in identifying sentiments for unstructured and informal data such as tweets (Zhou et al., 2020).

Transformers are designed to model immediate context and capture syntactic and semantic patterns that are difficult to capture with static vectors. For instance, the F1-scores obtained by RoBERTa for sentiment classification show that it can easily handle rich contextual inputs including sarcasm, idioms, and domain-specific slangs in Sentiment140 and SemEval datasets and can be deployed in the industry. This coincides with Xu et al. who established that incorporation of contextual models eliminates polarity inversion errors common in customer reviews and political opinion mining in particular.

One of the insights gathered by training RoBERTa is the tendency to generalize on one set without specific tuning from another domain. Although Sentiment140 and SemEval used different labeling and tweet formats, the model still had impressive performance. Such generalizability has also been attained in other researches. For example, Rana et al. (2022) used RoBERTa on multilingual fb comments achieving promising performance across language and cultural boundaries while reporting the model's transfer learning ability.

Another point that confirms the results of this work is the significant difference in performance between classical models and transformers identified in other comparative studies. Similar observations were made by Ali and Iqbal (2022) while working with hotel review datasets here they too mentioned that feature engineering irrespective of how complex done in traditional pipelines is unable to capture the context that the pre-trained models are capable of. This is particularly the case in social media sentiment analysis where some complications such as noise, ambiguity and short text pose severe threats in model development. In transformer models, these problems are solved by multiple attention heads, as well as token embeddings that contain positional and semantic features (Ding et al., 2021).

Also, it was found that most of the misclassifications were observed in the neutral class as observed by Gao et al. (2021). They state that neutral sentiment by its very nature, is subjective and contains few features for text classification Its difficulty even for state-of-art models is therefore easily explicable. This might be resolved by utilizing secondary sentiment indicators like emoticons, punctuation marks, and even users' account details to establish a richer discriminant in the most disputed cases.

Apart from issues to do with the tenability of the models, the practical impacts are highly important given the high-performing sentiment models. The public authorities and the health organizations have therefore realized the benefits of using Twitter analytics to monitor the public mood during critical moments like the current pandemic. Another work by Nematullah et al. (2022) showed how fine-tuned transformers were employed to identify vaccine hesitance and create appropriate intervention measures. Similarly, today's marketing departments use the tools that employ the same fundamentals for brand monitoring and decision-making regarding the products as well (Gupta & Goel, 2021).

Similarly, the statistically significant difference that exists between RoBERTa and SVM also advocates for organizations to incorporate better models even with the increased computational cost. Although logistic regression and SVM have some advantage of simplicity as well as efficiency, the analysis has shown that the cost-benefit ratio owes in favor of transformers in circumstances where accuracy of decision actually matters in terms of policy making or revenue generation. This is also evidenced by Martinez et al. (2022) where a retail company using BERT for customer sentiment analysis noted an overall conversion rate boost of 12% due to better interpretation of product feedback.

However, one major limitation of this study is also evident. First, the use of pre-annotated datasets imply that the models can carry over biases of labels. As

noted by Shah et al. (2021), it is also unlikely that different annotations of sentiment will be unique and precise when dealing with multilingual or code-switching tweets. Second, transformer models cannot be trivially applied on smaller devices due to their high computational cost and need to be fine-tuned to prevent overfitting. This might be an issue in real-world applications, especially for start-ups and NGOs, but DistilBERT is a compressed model (Chen et al., 2020).

Another interesting aspect of AI is ethical AI. This can be explained by the fact that most transformer models like the one used in this study are fine-tuned to large internet corpuses and therefore have the propensity to emulate them. Zhao and Bian (2022) have discussed how sentiment predictions may become biased by dialect, gendered language, and cultural reference that may have a negative impact on representation of certain groups in society. Therefore, there is a necessity to develop fairness auditing, and bias mitigation that forms the foundation for future implementation plans.

Thus, the observed good results of RoBERTa also raises questions on scaling the performance of even bigger models. Despite the enhancements observed in generative tasks such as GPT-4 and T5, the models do not yield significantly higher sentiment classification in short text scenarios (Lim et al., 2023). Therefore, base transformer models are still the most reasonable choice for many purposes when it comes to performance and utilization of resources.

Overall, this research affirms the applicability of transformer-based style models in sentiment analytic context in social media. Specifically, RoBERTa is given accurate results on various datasets, better generalization capacity, and a significant enhancement of classical models. Their practical usage in actual decision-making systems has to take into account computational concerns, fairness and quality of annotations.

References

Ali, A., & Iqbal, M. (2022). Comparative evaluation of transformer-based and traditional machine learning models for sentiment classification of hotel reviews. *Journal of Information Science*, *48*(3), 331–347. https://doi.org/10.1177/01655515211027617

Chen, W., Liu, H., Zhang, H., & Song, L. (2020). DistilBERT with multi-task learning for sentiment analysis. *IEEE Access*, *8*, 130453–130460. https://doi.org/10.1109/ACCESS.2020.3009942

Ding, Y., He, Y., & Guo, X. (2021). Exploring BERT for aspect-based sentiment analysis with multi-task learning. *Knowledge-Based Systems*, *213*, 106664. https://doi.org/10.1016/j.knosys.2020.106664

Gao, L., Li, Y., & Xu, Z. (2021). Handling neutral sentiments in sentiment analysis using hybrid neural networks. *Applied Soft Computing*, *101*, 107035. https://doi.org/10.1016/j.asoc.2021.107035

Gupta, N., & Goel, A. (2021). Brand intelligence through social media analytics: A BERT-based framework. *International Journal of Information Management Data Insights*, *1*(2), 100015. https://doi.org/10.1016/j.jjimei.2021.100015

Lim, D., Zhang, C., & Liu, J. (2023). Evaluating large language models for sentiment classification: Performance vs. compute cost. *Transactions of the ACL*, *11*, 112–125. https://doi.org/10.1162/tacl_a_00481

Martinez, R., Fernandes, J., & Silva, M. (2022). Business benefits of deep sentiment models in retail marketing analytics. *Decision Support Systems*, *158*, 113760. https://doi.org/10.1016/j.dss.2021.113760

Nematullah, F., Hassan, T., & Wahid, F. (2022). Mining COVID-19 vaccine hesitancy using transformer-based sentiment analysis of Twitter data. *Healthcare Analytics*, *2*, 100045. https://doi.org/10.1016/j.health.2022.100045

Rana, H., Shukla, P., & Mishra, A. (2022). Multilingual sentiment detection using RoBERTa in noisy social media text. *Expert Systems with Applications*, *193*, 116519. https://doi.org/10.1016/j.eswa.2021.116519

Shah, A., Iqbal, M., & Hussain, T. (2021). Cross-linguistic sentiment annotation challenges in code-switched social media data. *Language Resources and Evaluation*, *55*, 345–366. https://doi.org/10.1007/s10579-020-09504-1

Xu, R., Li, Y., & Wang, H. (2021). Improving contextual sentiment analysis with double attention and reinforcement learning. *Information Sciences*, *562*, 140–157. https://doi.org/10.1016/j.ins.2021.02.028

Zhao, T., & Bian, J. (2022). Fairness in AI-based sentiment analysis: Bias detection and mitigation techniques. *Journal of Artificial Intelligence Research*, *74*, 343–372. https://doi.org/10.1613/jair.1.12899

Zhou, Y., Zhang, L., & Wang, B. (2020). BERT-based models for deep sentiment analysis in short texts. *Neurocomputing*, *390*, 1–11. https://doi.org/10.1016/j.neucom.2020.01.104

Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., & Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). https://doi.org/10.18653/v1/2020.acl-main.485

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society*, *16*(2), 340–358. https://doi.org/10.1177/1461444813480466

Conneau, A., Khandelwal, K., Goyal, N., et al. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020* (pp. 8440–8451). https://doi.org/10.18653/v1/2020.acl-main.747

Gkotsis, G., Oellrich, A., Velupillai, S., et al. (2017). Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific Reports*, *7*, 45141. https://doi.org/10.1038/srep45141

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., & Kaymak, U. (2013). Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing* (pp. 703–710). https://doi.org/10.1145/2480362.2480498

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168–177). https://doi.org/10.1145/1014052.1014073

Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. https://ojs.aaai.org/index.php/ICWSM/article/view/14185

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=H1eA7AEtvS

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems* (pp. 2177–2185). https://proceedings.neurips.cc/paper_files/paper/2014/file/5f5d4720b8ecb90c297d9e66ebfdf3e7-Paper.pdf

Liu, P., Qiu, X., & Huang, X. (2018). Recurrent neural network for text classification with multi-task learning. *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 2873–2879. https://doi.org/10.24963/ijcai.2016/404

Nguyen, D. Q., Vu, T., & Tuan Nguyen, A. (2020). BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 9–14). https://doi.org/10.18653/v1/2020.emnlp-demos.2

Pagolu, V. S., Reddy, K. N. R., Panda, G., & Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)* (pp. 1345–1350). IEEE. https://doi.org/10.1109/SCOPES.2016.7955653

Peters, M. E., Neumann, M., Iyyer, M., et al. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT* (pp. 2227–2237). https://doi.org/10.18653/v1/N18-1202

Rao, T., & Srivastava, S. (2014). Analyzing stock market movements using Twitter sentiment analysis. In *2014 IEEE International Conference on Advances in Engineering & Technology Research (ICAETR-2014)* (pp. 1–5). https://doi.org/10.1109/ICAETR.2014.7012805

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. https://doi.org/10.48550/arXiv.1910.01108

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). https://doi.org/10.18653/v1/P19-1163

Sun, Z., Yu, H., Song, X., et al. (2020). MobileBERT: A compact task-agnostic BERT for resource-limited devices. In *Proceedings of ACL 2020* (pp. 2158–2170). https://doi.org/10.18653/v1/2020.acl-main.215

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 90–94). https://aclanthology.org/P12-2018/

Wang, W., Wu, J., Lin, Z., & Liu, Z. (2021). Multilingual sentiment analysis on social media: State of the art and open challenges. *Information Fusion, 72*, 140–157. https://doi.org/10.1016/j.inffus.2021.01.008

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL-HLT 2018* (pp. 15–20). https://doi.org/10.18653/v1/N18-2003

Zhuang, Y., Jiang, Y., & Wang, Y. (2021). Financial sentiment analysis based on RoBERTa. *Procedia Computer Science, 187*, 208–213. https://doi.org/10.1016/j.procs.2021.04.075

Ahmed, W., Vidal-Alaball, J., Downing, J., & López Seguí, F. (2020). COVID-19 and the 5G conspiracy theory: Social network analysis of Twitter data. *Journal of Medical Internet Research, 22*(5), e19458. https://doi.org/10.2196/19458

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems, 28*(2), 15–21. https://doi.org/10.1109/MIS.2013.30

Cinelli, M., Quattrociocchi, W., Galeazzi, A., et al. (2020). The COVID-19 social media infodemic. *Scientific Reports, 10*(1), 16598. https://doi.org/10.1038/s41598-020-73510-5

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186). https://doi.org/10.48550/arXiv.1810.04805

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 56*(4), 82–89. https://doi.org/10.1145/2436256.2436274

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford University*. https://www.cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons, 53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Kumar, A., & Sebastian, T. M. (2012). Sentiment analysis: A perspective on its past, present and future. *International Journal of Intelligent Systems and Applications, 4*(10), 1–14. https://doi.org/10.5815/ijisa.2012.10.01

Liu, B. (2012). *Sentiment analysis and opinion mining*. Synthesis Lectures on Human Language Technologies, 5(1), 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692

Lwin, M. O., Lu, J., Sheldenkar, A., et al. (2020). Global sentiments surrounding the COVID-19 pandemic on Twitter: Analysis of Twitter trends. *JMIR Public Health and Surveillance, 6*(2), e19447. https://doi.org/10.2196/19447

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. https://doi.org/10.48550/arXiv.1301.3781

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 1320–1326). https://aclanthology.org/L10-1533/

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). https://doi.org/10.3115/v1/D14-1162

Samuel, J., Ali, G. G. M. N., Rahman, M. M., & Esawi, E. (2020). COVID-19 public sentiment insights and machine learning for tweets classification. *Information Processing & Management, 58*(4), 102522. https://doi.org/10.1016/j.ipm.2021.102522

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. https://doi.org/10.48550/arXiv.1910.01108

Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3407–3412). https://doi.org/10.18653/v1/D19-1339

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In *China National Conference on Chinese Computational Linguistics* (pp. 194–206). Springer. https://doi.org/10.1007/978-3-030-32381-3_16

Turc, I., Chang, M. W., Lee, K., & Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*. https://doi.org/10.48550/arXiv.1908.08962

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (pp. 178–185). https://ojs.aaai.org/index.php/ICWSM/article/view/14009

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008). https://doi.org/10.48550/arXiv.1706.03762

Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review, 53*(6), 4335–4385. https://doi.org/10.1007/s10462-019-09794-5

Yang, Z., Dai, Z., Yang, Y., et al. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems* (pp. 5753–5763). https://doi.org/10.48550/arXiv.1906.08237.