PoS tag-based Attention for Feature Selection in Sentiment Analysis

K. H.S.L Kanakkahewa (sahanbcsrh@gmail.com)
University of Ruhuna
W. A. Mohotti
University of Ruhuna
L. D.C.S. Subhashini
University of Sri Jayewardenepura

Research Article

Keywords: Sentiment analysis, BERT, PoS Tagging, Attention mechanism, Polarity words

Posted Date: July 13th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3151544/v1

Additional Declarations: No competing interests reported.

PoS tag-based Attention for Feature Selection in Sentiment Analysis

K.H.S.L Kanakkahewa [a], W.A. Mohotti [b], L.D.C.S. Subhashini [c]

[a] Department of Computer Science, University of Ruhuna, Sri Lanka, sahanbcsrh@gmail.com

[b] Department of Computer Science, University of Ruhuna, Sri Lanka, wathsala@dcs.ruh.ac.lk

[c] Department of Information Technology, University of Sri Jayewardenepura, Sri Lanka, subhashini@sjp.ac.lk

Abstract

The growth in the usage of the Internet and Web-based applications has led to an exponential increase in text data, including customer reviews, social media comments, and blogs. Sentiment analysis is an essential technique to analyze this data and determine the polarity of opinions. However, traditional sentiment analysis methods face challenges in feature selection and representation due to the unstructured nature of the text and associated noise. We propose a novel framework called PoPoSBert that uses BERT to capture syntax, semantic, and contextual features effectively. We leverage part-of-speech (PoS) tags which assign a unique identifier to every single token in a text corpus to indicate the part of speech, to capture polarity words that reveal the sentiment behind the text. We use the transformer architecture's attention mechanism to improve the weights for polarity words, resulting in high-quality feature selection for sentiment classification. We conducted an extensive experimental analysis to evaluate the suggested framework against the existing methods. The suggested framework is assessed using several standard evaluation measures such as accuracy, precision, recall, and F1-score against the traditional baseline methods such as TF-IDF and existing models like BERT, Word2Vec, and GloVe techniques. The findings indicate that the PoS tag-based attention concept improves the accuracy of sentiment classification, achieving superior performance over existing methods on several benchmark datasets. The suggested framework can be utilized in various applications, including market research, monitoring social media and analyzing customer feedback to make informed decisions.

Keywords

Sentiment analysis, BERT, PoS Tagging, Attention mechanism, Polarity words.

1. Introduction

The increase in text data is exponential over the internet and web-based platforms, and it is expected that there will be substantial expansion and increased focus in the upcoming years. (Tam et al., 2021). Digital platforms including forums, Wikis, blogs, news feeds, e-marketplaces, and emails play a vital role in allowing the sharing of thoughts and social viewpoints via text communication between users. Digital communication platforms such as Twitter, and Instagram also disseminate information that is currently trending using short-text communication (Dashtipour et al., 2015). People heavily use these platforms to publicly share their opinions and reviews.

Researchers and practitioners use different text-mining methods to obtain interesting patterns inherent in these opinions and reviews to improve business decisions. Sentiment analysis is a supervised text-mining method of opinion classification that is used to analyze a vast amount of review data (Rehman et al., 2019). Sentiment analysis, alternatively referred to as opinion mining, refers to a person's perspective, emotion, judgment, or evaluation of a given good or service. Usually, opinion is considered either positive or negative in binary classification. (Srividya & Mary Sowjanya, 2019). Some studies treat sentiment analysis as a multi-class classification(Zhao et al., 2021) problem with the inclusion of the choice of neutral and variations of positive/negative options (Pontiki et al., 2014).

Sentiment analysis is useful for people and organizations who want to explore customers' or public opinions about a specific topic, issue, product, or event (Srividya & Mary Sowjanya, 2019). Sentiment analysis has replaced conventional and web-based surveys used by businesses to ascertain feedback about their items for brand monitoring and improving customer satisfaction. Political decision-makers can use sentiment analysis to understand public opinion and the sentiment of political discussions. Through the analysis of sentiment in various sources such as social media posts and news articles, political decision-makers can gain insights into how different policies or issues are perceived by the public. This can help them make better educated judgments that will be supported by the public.

Figure 1 depicts the three primary steps of text-based sentiment analysis. It initiates with the review data that is medium in text vector length or social media data that are extremely short. Preprocessing deals with unstructured, messy, and noisy raw text data by cleaning, organizing, and standardizing it to make it more suitable for analysis. The feature selection and representation techniques capture and model the relationship between the features and assist classification tasks to identify groups. The classifier takes in a piece of text as input and uses algorithms to analyze the words and phrases used, as well as their context, to determine the text's emotional tone. The classifier then assigns the text to one of several predefined categories, to assess the sentiment of the text, such as positive, negative, or neutral. The feature selection and representation have a higher influence in selecting and extracting high-quality features for classification tasks and thereby control the overall accuracy of the sentiment classification (Tam et al., 2021). Therefore, it is essential to employ effective feature representation techniques to capture syntax, existing semantic and contextual ties in the text and to identify the polarity of words that highlight opinions to end up with a better decision-making process.

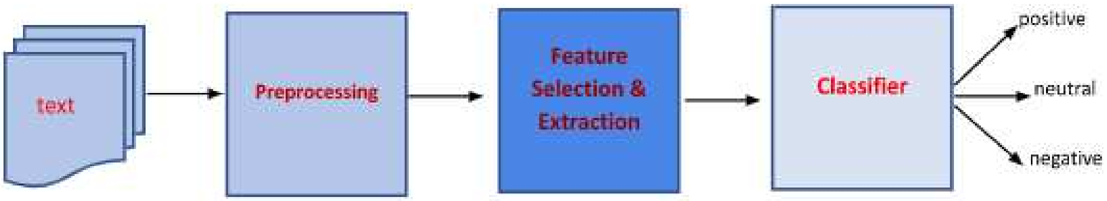


Figure 1:Phases of Sentiment analysis

Different machine learning concepts have been explored in recent years to effectively extract and represent features with syntax and semantics for classification in sentiment analysis (Pang et al., 2002). The contemporary paradigm shift toward contextual language embedding with BERT has given fantastic success by capturing syntax, semantics together with contextual features in feature representation (Devlin et al., 2019). However, the approach employed in the proposed method is entirely different from it. It addresses the gaps in effectively identifying the polarity words while capturing contextual information in representation.

In summary, this research proposes a novel framework for dealing with sentiment analysis using novel high-quality techniques for feature selection and extraction with deep learning concepts and cutting-edge feature representation methods named “Polarity PoS Bert” (PoPoSBert). Part-of-speech tagging involves assigning a grammatical category to each word depending on its context which is used in sentiment analysis work and obtaining scores for each token calculated via sentiment libraries to unveil the sentiment of a text. The assistance of Point of Speech (PoS) tagging information is used to improve the features vector representation in the proposed approach. It employed a more elaborate representation of positive and negative tags to ascertain the sentiment conveyed in the text. Then PoS vector representation that highlights polarity words and Bert representation that capture local contextual information of the text is combined with the attention concept to combine both information to obtain a high-quality feature.

To the extent of our current understanding, PoPoSBert is the first method that extends the contextual feature embedding in BERT to combine with PoS tag information to capture the polarity words via the attention concept. By conducting empirical analysis on various reviews and social media data for sentiment analysis, it has been observed that PoPoSBert demonstrates a high level of accuracy in effectively classifying opinions when compared to other state-of-the-art methods.

2. **Related work**

Over the past few years, substantial research efforts have been dedicated to exploring the field of sentiment analysis. According to Subhashini et al (Subhashini et al., 2021a), a significant advancement in this domain involves the categorization of sentiment analysis into three primary levels: document level, sentence level, and aspect/feature level. These levels reflect the scope and granularity of the analysis, with document level encompassing the general tone of a document, sentence level focusing on the sentiment of individual sentences within a document, and aspect/feature level examining the sentiment towards specific aspects or features within a document or sentence. Specifically, document-level classification involves evaluating an entire document as a customer review and determining if it reflects a favorable or negative viewpoint about a product. Sentence-level categorization involves analyzing a single sentence and determining if it reflects a positive or negative opinion. Aspect/feature-level classification involves identifying entities, characteristics, and relationships among opinions. Understanding the nuances of these classification levels has allowed researchers to make significant advancements in the field and propose improvements for the accuracy and effectiveness of sentiment analysis techniques. We concentrate on document-level sentiment analysis in this study.

Sentiment analysis models deal with extensive quantities of textual data exhibiting a broad spectrum of characteristics. The text data are unstructured in nature and high in dimensions (Tam et al., 2021). Feature selection and extraction techniques (Rehman et al., 2019) play a vital role to enhance the precision of sentiment analysis.. They effectively identify key characteristics to determine the opinion reducing the workload of sentiment analysis classifiers. In the early stages of sentiment analysis one hot encoding, a bag of words, and N-Gram (Liu & Forss, 2014) techniques were used for feature selection. These techniques perform well with machine learning and rule-based sentiment analysis models. However, the sparseness of text data due to its high-dimensional nature led researchers to investigate improved feature representation mechanisms such as word embedding.

2.1. **Word embedding for feature selection**

Multiple research studies utilized word embedding algorithms as input to extract and select text features for sentiment classification (Tam et al., 2021). The word embedding concept is used to address the issue of data sparsity in the Bag of Word approach. The computational cost of sparse matrices is high due to the significant number of unnecessary zeros that exist in their matrix structure. The issue of having a large size significantly amplifies the complexity of space usage, making it difficult to discern meaningful features. Word embedding is a method for learning features that associates each word or phrase in a given vocabulary with a vector of real numbers in N dimensions. Numerous word embedding techniques have been devised to transform individual words into meaningful input for machine learning systems. Popular techniques include context2vec (Melamud et al., 2016), Word2Vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), FastText (Bojanowski et al., 2017), and BERT (Devlin et al., 2019).

Tomas et al (Bojanowski et al., 2017) introduced an enhanced word embedding architecture called "word to vector" representation. To generate high-dimensional vectors for individual words, the Word2Vec technique utilizes shallow neural networks featuring two hidden layers, specifically the Continuous Bag-of-Words (CBOW) and Skip-gram models. Word2Vec is a predictive model that is trained to predict context words based on a target word (using the skip-gram method) or predict a target word given its context (using the CBOW method). This technique is incredibly effective in capturing relationships within a text corpus and determining word similarity by maximizing the probability. For instance, in the vector space, the embeddings of "big" and "bigger" would be considered close to each other in terms of their values. However, Word2vec focuses on local information in prediction.

In contrast, GloVe which stands for Global Vectors captures vector statistics in the global setting (Pennington et al., 2014). It is another potent method of word embedding that has been applied to sentiment analysis. Similar to the Word2Vec approach, this method represents each word as a high-dimensional vector and trains it using the surrounding words within a large corpus. However, it learns its vectors by doing dimensionality reduction on the co-occurrence counts matrix considering the whole text corpus. The GloVe method has limitations in capturing sequential information and context, which may negatively impact its effectiveness in sentiment analysis tasks as sentiment analysis heavily depends on the word sequence and the context of words to infer meaning.

2.2. **Attention models**

The problem of processing long sequences of data in neural networks has been a significant challenge for existing deep learning models. Traditional neural networks, which are commonly used for these types of tasks, process all inputs independently without considering their order or position in the sequence. Therefore, as a solution to the issue of fixed-length context vectors, attention models were developed (Vaswani et al., 2017). In many tasks, attention processes have evolved into a crucial component of persuasive sequence modeling and transduction models, enabling the modeling of dependencies without taking into account their proximity in the input or output sequences (Nathani et al., 2019).

Current state-of-the-art models face limitations in efficiently processing long sequences of data (Hameed & Garcia-Zapirain, 2020a). This reduction in resolution can be mitigated by using Multi-Head Attention. Attention functions take a query and a collection of key-value pairs as input and produce an output, all represented as vectors. The output is calculated by taking a weighted sum of the values, with each weight determined by a compatibility function that evaluates the relationship between the query and the corresponding key (Vaswani et al., 2017). Transformers utilize self-attention mechanisms to analyze the connections between words within a text and extract contextual information from the entire sequence. This capability enables them to capture long-range dependencies in the text, which is crucial for comprehending the text's meaning and sentiment.

2.3. **BERT**

BERT is an acronym for Bidirectional Encoder Representations from Transformers. BERT is both conceptually and empirically simple (Devlin et al., 2019). BERT aims to pre-train deep bidirectional representations from unlabeled text by considering both the left and right contexts simultaneously across all layers. Thus, Bert embeddings of a word are not static. The embeddings of a word in BERT vary depending on the context of the surrounding words. For instance, in a sentence like "one bird was flying below another bird," the two embeddings of the word "bird" will differ. The pre-trained BERT model can be fine-tuned by adding just one additional output layer, enabling the creation of cutting-edge models for various downstream tasks, including sentiment analysis, question answering, and language inference. Remarkably, these fine-tuned models achieve impressive performance without substantial modifications to the task-specific architecture.

2.4. PoS Tagging

Part-of-speech (PoS) tagging involves assigning a grammatical label to each word based on its context (Wang et al., 2018). It categorizes words in a sentence into their respective parts of speech, such as nouns, verbs, adjectives, or adverbs, by applying PoS tagging rules. PoS tagging is widely applied in various tasks, including text classification, speech recognition, and automatic machine translation. The process of PoS tagging comprises two stages: the training stage and the tagging stage. In the training stage, a corpus is utilized to observe words in different contextual settings, allowing the extraction of rules to determine the word's lexical class based on contextual information. In the tagging stage, the most likely tag for a word is determined by calculating the probability of its context and immediate neighbors appearing together (Wang et al., 2018).

PoS tagging is often used in aspect-based sentiment analysis, which involves three tasks: extracting aspects, identifying and orienting associated opinions, and producing a final summary. (Samha et al., 2014) propose opinion extraction from customer reviews based on PoS tagging. This work uses sentence importance to determine the sentiment by calculating the weights for all "adjectives, adverbs, and verbs for each sentence. The calculation is done by adding up all weights for each "adjective, adverb, and verb "within the sentence. Subramanian (Subrahmanian & Reforgiato, 2008) used the PoS tags to determine the sentiment of sentences. As a result, PoS tagging can be leveraged to enhance the feature representation process in sentiment analysis, leading to improved overall performance in sentiment analysis tasks. However, PoS tagging is often used with rule-based classification methods and hardly combined with classification techniques in existing work.

3. Research Objectives

In this study, we have introduced an innovative framework with PoS tags, attention concept and BERT model that capture syntax, semantic and contextual features to effectively capture high quality features for sentiment analysis. The primary goals of the research can be summarized as follows.

- The initial aim was to effectively model polarity words for sentiment identification.

  In sentiment analysis identifying polarity words is important in identifying sentiment. Researchers have use dictionary-based approaches, rule-based approaches and PoS tagging(Tam et al., 2021) in identifying polarity words. Among them PoS tags play an important role in identifying positive and negative opinion identification based on more detailed grammatical categories(Tubishat et al., 2018) to improve the accuracy of identifying sentiments. Therefore, we further analyze grammatical categories to support polarity word identification.

- The second objective was to embed polarity information together with the syntax and semantic information, as well as local context features(Yan et al., 2015) to improve the accuracy of sentiment analysis.

  Researchers have investigated different consensus strategies to include polarity information for sentiment analysis. Attention functions play an extremely important role in highlighting important words in the text representation. BERT pre-trained deep bidirectional representations also learn contexts simultaneously through the self-attention concept. Therefore, this research further investigates the mechanism to capture all this information with an attention technique to the BERT model.

- The third objective was to prove that the proposed approach outperforms state-of-the-art feature selection and extraction methods in sentiment analysis.

  There exist multiple research works that propose to select and extract features for sentiment analysis(Subhashini et al., 2021b). Therefore, the proposed PoPoSBert is evaluated against those baselines using multiple datasets to confirm its accuracy.

4. **Proposed approach**

Our research aim was to develop a novel framework to extract high-quality features relevant to sentiment analysis and to improve the accuracy of sentiment classification on medium and short text (i.e., movie reviews and tweets). The proposed framework uses the PoS tags and attention-based approach to represent polarity words with the BERT feature embedding after data pre-processing. Specifically, the proposed PoPoSBert selects Bert embedding as the base feature representation method to capture syntax, semantic and contextual information that exists within the document collection and couples it with the PoS tag base embedding that captures polarity words using the attention concept. Figure 2 provides an overview of the proposed framework called PoPoSBert.

PoPoSBert uses transformer architecture consisting of standard 12 Bert layers for better feature extraction. To enable contextual predictions in BERT encoding, certain tokens within the input sequence need to be masked and replaced with a special [MASK] token. The output is generated in a non-autoregressive manner, meaning that all tokens are computed simultaneously without any self-attention mask. This computation is conditioned on the non-masked tokens that exist in the same input sequence as the masked tokens. Therefore, our case for sentiment analysis only requires this encoding process for contextual feature representation not using decoders.

The segment embedding and positional embedding were considered as inputs in the general approach of BERT to generate embedding. In the proposed PoPoSBert, additionally, we input the PoS tag embedding to capture the polarity words to capture the opinion.

The PoPosBert used PoS tags useful for determining the sentiment such as nouns, verbs, adjectives, adverbs, comparative adjectives, comparative adverbs, superlative adverbs, and superlative adjectives aligning with Subrahmanian (Subrahmanian & Reforgiato, 2008) work. Samaha et al, Click or tap here to enter text. prove the use of more detailed grammatical categories to improve the accuracy of identifying sentiments and utilize PoS tags to serve this purpose (Samha et al., 2014). Thereby, they accurately identify polarity words as positive or negative. In our research, we established distinct dictionaries aligned with the grammatical categories suggested in (Samha et al., 2014). These dictionaries include positive nouns, positive verbs, positive adverbs, positive adjectives, negative nouns, negative verbs, negative adverbs, and negative adjectives. Based on the PoS tags and the polarities we recognized 16 categories for the PoS tag-based feature representation. As an example, for the category's positive nouns, negative nouns, neutral nouns, positive adjectives, and negative adjectives like categorization was used. Then these PoS tags are assigned categorical weights for the feature representation and then attention mechanisms are used to combine the feature representation.

Finally, we combined PoS-based embedding with Bert embedding using the attention mechanism in PoPoSBert for feature selection and extraction. Additive attention is used for combining both embeddings together to obtain quality feature representation.

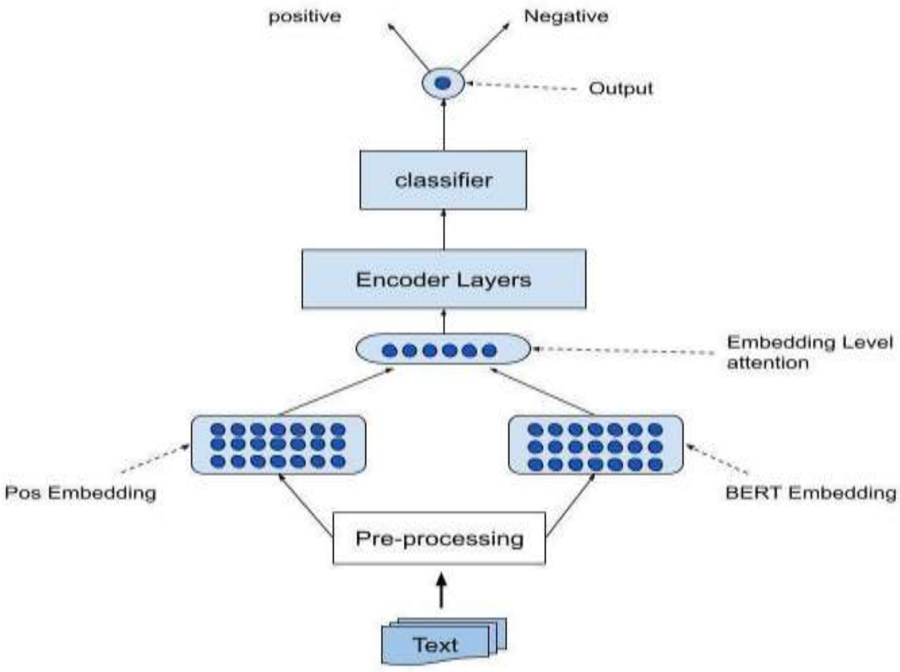


Figure 2:Proposed PoPoSBert Framework

The attention weights are formulated as given in Eq (1) using element-wise tanh operation through W_a and U_a is weights taken from Bert embedding x and PoS embedding p for a text. Thereby we form the final enriched embedding X in Eq. (3) for text using the dot product with a scalar parameter V which we experimentally set as given in Eq. (2) on resultant weight.

$$a = (tanh (W_a x + U_a p)) \qquad (1)$$

$$a = a @ V \qquad (2)$$

$$X = a.x + (1 - a)p \qquad (3)$$

Algorithm 1 explains this enriched embedding process utilizing equation 1-3.

We used Layer Normalization (LayerNorm) and Dropout techniques that are commonly used in deep learning models to improve their performance and generalization ability. Layer Normalization is a technique that is applied to the activations of a layer in a neural network to stabilize the training process and prevent activations from becoming too large or too small. Dropout is a regularization technique that sets a percentage of activations to zero during training to prevent overfitting by reducing model complexity. Both Layer Normalization and Dropout are applied to hidden layers.

Algorithm 1: PoPoSBert Embedding

```
Inputs:   x- input Bert embedding vector
          p - PoS categorical input embedding vector
          $W_a$ - weights Vector of Bert Embedding
          $U_a$ - weights Vector of PoS Embedding
          V - Scaler parameter.
Output: X – Enriched Embedding
$a = (\tanh (W_a x + U_a p))$      Eq. (1)
$a = a @ V$Eq. (2)
$X = ax + (1-a) p$     Eq. (3)
$X = \mathrm{LayerNorm}(X)$
$X = \mathrm{Dropout}(X)$
return X
```

After this enrichment in embedding is completed, we send the output through 12 encoding layer stacks in BERT Transformer architecture to form the contextual embedding considering both left and right contexts simultaneously.

5. **Experiment**

The experiments were designed to evaluate the accuracy of the PoPoSBert framework in text sentiment classification on short/medium text data. Also, we have conducted experiments to analyze the sensitivity of PoS tag-based embedding and attention formula. The details of the experimental setup, evaluation measure, datasets, and different experiments together with results are given in the following sections.

5.1. Benchmarks and experimental setup

Experiments were conducted using google colab pro facilities using Python 3. Keras Pytorch and NLTK libraries are used in experiments within the program.

As the bassline for the evaluation of PoPoSBert, we utilized

1. Tf-idf feature selection with the naïve Bayes(Pang et al., 2002)

2. Word2vec (Wang et al., 2018)

3. Glove. (Hameed & Garcia-Zapirain, 2020b)

4. Bert Embedding (Munikar et al., 2019)

As the classifier for the sentiment classification, we used a linear classifier to be consistent in all the experiments except in the baseline that uses Tf-idf feature selection with the naïve Bayes.

5.2. Dataset

The datasets used for sentiment analysis such as tweets and movie review datasets are short or medium in vector length. We have used well-known datasets in sentiment analysis research fields as given in Table 1. IMDB and SST2 are datasets of movie reviews that are considered to be medium context datasets. Twitter and the SamEval dataset are related to Twitter and are therefore shorter in length, making them short context datasets. IMDB, SST2, and Twitter are datasets for binary classification tasks, meaning that they involve categorizing text into two predefined classes, such as positive and negative sentiment. On the other hand, the SamEval dataset is a 3-class classification dataset, meaning that it involves categorizing text into three predefined classes, such as positive, negative, and neutral sentiment.

Our study employed a subset of the Tam Sakiran Twitter dataset (Tam et al., 2021), which consisted of 20,162 positive tweets and 29,838 negative tweets, totaling 50,000 tweets. The original dataset was collected from Twitter users located in Chicago over a two-month period from September 1 to October 31, 2019, and was sentiment labeled using a binary classification scheme, using a binary coding scheme, positive tweets are assigned the value 1, while negative tweets are assigned the value 0. The dataset is extensive and encompasses a broad range of topics and sentiment expressions, rendering it a valuable resource for sentiment analysis model training and evaluation. In our research, our focus was specifically directed towards a specific aspect of sentiment analysis and hence selected a subset of the Tam

Sakiran Twitter dataset, ensuring that the tweets chosen were pertinent to our investigation. It is crucial to acknowledge that our subset may not fully represent the entire dataset, and it is important to recognize potential limitations and biases in our approach.

The Internet Movie Database (IMDB) is a comprehensive database containing a vast collection of movies, TV shows, and other forms of entertainment content (Maas et al., 2011). The IMDB dataset specifically designed for sentiment analysis comprises 50,000 movie reviews sourced from the IMDB website. These reviews are annotated with binary sentiment polarity, indicating whether they are positive or negative. The dataset is evenly split into a training set of 25,000 reviews and a test set of 25,000 reviews. Each review is labeled based on the star rating given by the reviewer, where reviews with ratings of 7 or higher are labeled as positive, and those with ratings of 4 or lower are labeled as negative. Reviews with ratings between 5 and 6 are excluded from the dataset due to their ambiguous sentiment. Notably, the reviews within the IMDB dataset tend to be longer than those in the SST-2 dataset, averaging around 230 words per review.

The Stanford Sentiment Treebank (SST) is a widely recognized dataset extensively used in sentiment analysis research (Socher et al., 2013). It comprises sentences extracted from movie reviews, each labeled with binary sentiment polarity (positive or negative). Specifically, the SST-2 subset of the dataset contains a total of 67,346 sentence examples, which are further divided into a training set consisting of 56,195 sentences and a test set comprising 11,151 sentences. Each sentence is assigned a label indicating whether it expresses a positive or negative sentiment. These labels are derived from a 5-star rating system, where ratings of 1-2 stars are considered negative and ratings of 4-5 stars are considered positive. Sentences with 3-star ratings are excluded from the dataset due to their ambiguous nature in assigning a clear sentiment polarity. The sentences in the SST-2 dataset are typically shorter than those in other sentiment analysis datasets, with an average length of around 19 words per sentence.

The SemEval-2013 Task 2 (Pontiki et al., 2014) dataset is based on Twitter data and consists of 9,846 tweets. The tweet dataset used in this study was gathered during the period of January to February 2013. The selection of tweets aimed to encompass diverse topics and incorporate informal language, including slang and misspellings commonly found on Twitter. The dataset was deliberately designed to present challenges due to the informal nature of tweet language and the restricted length of tweets, which was limited to 140 characters at the time of the challenge.

Dataset		Positive	Negative	Neutral	Dataset size

Twitter(Tam et al., 2021)	2	20,162	29,838	_	50,000
IMDB (Maas et al., 2011)	2	25,000	25,000	_	50,000
SST-2(Socher et al., 2013)	2	37569	29780		67349
SamEval(Pontiki et al., 2014)	3	3,246	3,423	3,177	9,846

Table 1:Datasets

5.3. **Evaluation metrics**

To assess the performance of the PoPoSBert framework, various evaluation metrics such as Accuracy, Precision, Recall, and F1-Score were employed. These metrics were used to compare the performance of PoPoSBert against state-of-the-art methods, enabling a comprehensive evaluation of the framework's effectiveness.

Accuracy

The evaluation metric in Eq. (4) quantifies the model's ability to accurately predict the correct class for a given input.

$$Accuracy(Acc) = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (4)$$

Precision

Precision is a measure of the proportion of positive predictions that are actually correct as in Eq. (5).

$$Precision(P) = \frac{(TP)}{(TP+ FP)} \qquad (5)$$

Recall

Eq. (6) represents Recall as a metric that quantifies the proportion of actual positive cases that were accurately predicted by the model.

$$Recal(R) = \frac{(TP))}{(TP+FN)} \qquad (6)$$

F1-Score

Eq. (7) defines the F1-score as a metric that captures the balance between precision and recall. It is calculated as the harmonic mean of precision and recall, and it serves as a valuable metric for evaluating the overall performance of a classifier.

$$F1\_Score(F1) = \frac{2*(Precision* Recal)}{(Precision+Recall)} \qquad (7)$$

5.4. Data preprocessing

Preprocessing steps have been used to convert unstructured noisy text to structured and useful outcome. We remove stop words, punctuations, and tokenize words while removing the most frequent words and least frequent words in the text document collections. However, we did not remove negation-related stop words because it helps to deal with the negative sentiment. The tweet dataset was handled using special preprocessing techniques such as replacing emojis with the meaning of the emoji. Specifically, chat words(slang) are replaced with long-term meanings, and emoticons are also replaced with meaning. Additionally, the negation of short terms is replaced with long terms.

5.5. Hyper-parameters setting

The hyperparameter data used for the experiment are shown in Table 2. PoS embedding vocabulary includes 22 categories, it means that there are 22 different PoS tags and categories being used in the model. Due to limitations in computation power and training time, we set a maximum of 10 epochs for the training process. It is possible that further improvements in results could have been achieved with a larger number of epochs. In summary, the proper selection of hyperparameters plays a crucial role in achieving high-quality results in our experiments. As per the Hyperparameter configuration presented in Table 2, the pertinent outcomes are exhibited in Table 3 and Table 4.

hyperparameter	value
Word embedding size	768
PoS Embedding size	768
Word embedding vocab size	30522
PoS embedding vocab size	22
Hidden size	768
padding	VALID
dropout	0.1
Batch size	32
epoch	10
Learning rate	0.00001

Table 2: Hyperparameters

5.6. **Experiments**

Accuracy results against the baselines are given in Table 3. The findings indicate that the PoPoSBert model, which incorporates advanced feature representation techniques to capture word polarity and contextual information, outperforms other state-of-the-art models in sentiment analysis across all datasets. This is the primary factor contributing to the superior performance of the PoPoSBert model. Among the baselines, BERT which is used as the foundation for PoPoSBert outperforms the other word embedding models as it captures the syntax, semantic as well as context information in a document collection. Although word2vec-based models are commonly used for representing fractures in sentiment analysis, they have certain limitations. While word2vec models are able to represent features based on semantic and syntactic characteristics, they neglect contextual relationships. Word2vec models use a simple averaging technique to represent documents or sentences with word vectors, but this approach can lead to the loss of important syntactic and semantic details. On the other hand, GloVe-based models capture global statistical characteristics while still maintaining the semantic and syntactic properties found in word2vec. GloVe represents words as vectors based on co-occurrence statistics, but it doesn't take into account the specific context in which the words appear. As a result, GloVe may be limited in its ability to fully capture the meaning of a word or phrase in a given context. This limitation is demonstrated in the results presented in Table 3. In contrast, the PoPoSBert model utilizes more advanced techniques, such as transformers, which are able to capture the context and relationships between words more effectively while capturing syntax and semantic information.

The SamEval dataset, commonly used for evaluating sentiment analysis models, has been observed to yield lower levels of statistical significance when compared to other datasets. One possible explanation for this is that the SamEval dataset consists of three classes, namely positive, neutral, and negative, which could affect the evaluation results. On the other hand, the other datasets used in this study are with only two sentiment categories (positive and negative). It is generally more challenging to accurately classify text into multiple categories, as there are more possible combinations and distinctions that need to be made.

In summary, Tf-Idf weights are unable to capture semantic relationships between words accurately as it considers syntactic information. Although word embedding models consider word co-occurrences, they also have their own limitations as per the technique used to capture relationships. In contrast, BERT embeddings take into account the relationships between words and their surrounding context, enabling a more comprehensive understanding of the meaning and sentiment of the text. However, it also does not capture certain important linguistic features or nuances that could impact the sentiment of the text such as polarity words. The proposed PoPoSBert model incorporates additional information such as part-of-speech tags to detect polarity words, which allows it to achieve a more enriched and high-quality feature representation. The PoPoSBert model also utilizes the attention mechanism, which helps to focus on specific words or phrases within the text and their relationships to other words, enhancing the model's ability to accurately classify sentiments.

Table 3: Experimental results of models

Datasets	PoPoSBert		Bert		Glove		Word2vec		Tf-Idf	
	accuracy	f1-score	accuracy	f1-score	accuracy	f1-score	accuracy	f1-score	accuracy	f1-score
IMDB	0.93	0.93	0.92	0.92	0.72	0.72	0.79	0.79	0.87	0.87
SST-2	0.95	0.95	0.94	0.94	0.84	0.84	0.60	0.58	0.87	0.87
SamEval	0.71	0.71	0.67	0.67	0.58	0.58	0.54	0.52	0.60	0.60
Twitter	0.98	0.98	0.95	0.95	0.90	0.90	0.69	0.68	0.84	0.84
Average	0.89	0.89	0.87	0.87	0.76	0.76	0.65	0.64	0.79	0.79

Datasets	PoPoSBert				Pos Bert				Bert			
	accuracy	precision	recall	f1-score	accuracy	precision	recall	f1-score	accuracy	precision	recall	f1-score
IMDB	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.92	0.92	0.92	0.92
SST-2	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
SamEval	0.71	0.71	0.71	0.71	0.68	0.68	0.68	0.67	0.67	0.67	0.67	0.67
Twitter	0.98	0.98	0.98	0.98	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Average	0.89	0.89	0.89	0.89	0.88	0.88	0.88	0.88	0.87	0.87	0.87	0.87

Table 4: Experimental results of PoS tag based embeddings

6. RESULTS

6.1. Sensitivity Analysis of PoPoSBert

We have analyzed the PoPoSBert with different settings in (1) PoS tag-based embedding and (2) Attention formula

6.1.1. **PoS tag-based embedding**

To examine the effect of incorporating polarity words in the PoPoSBert model, we compared the results of using PoS embedding without these words to those obtained with PoSBert, PoPosBert, and Bert results shown in Table 4. In the PoSBert model, no polarity-based dictionaries are used for categorical weighting; only PoS tags are identified. In this case, the tags of the words are first identified and then weights are assigned without considering whether the identified PoS tag word is positive or negative. A total of 16 categories are identified in this manner. In contrast, the PoPoSBert model also incorporates the use of defined polarity PoS tag-based dictionaries for comparison. As Shown in Table 4 results showed that PoSBert had a significant improvement compared to Bert due to the inclusion of PoS tags for enhanced feature extraction. However, the best results were obtained with the PoPosBert model, which utilized both PoS tagging and handling of polarity words to achieve high-quality feature extraction with attention mechanisms.

6.1.2. **Attention Formula**

In order to obtain high-quality feature extraction using attention mechanisms, we experimented with various attention equations utilizing SoftMax functions and different activation functions (e.g., sigmoid, tanh, ReLU). Specifically, we considered Eq(4): X=W_a x + U_a p , Eq(5): a=Sigmoid(tanh (W_a x + U_a p)), X=a.x + (1-a)p, Eq(6 a=(tanh (W_a x+ U_a p)), X=a.x+(1-a)p, X = softMax(X), Eq(7): a=(tanh (W_a x+ PreLU(U_a p))), X=a.x+(1-a)p and Eq(8): a=Sigmoid(tanh (W_a x+ PReLU(U_a p)) and used as the baselines against the Eq(1) in Section 4 - 4. Proposed approach. After obtaining results for these attention equations as shown in the Figure 3, the best performing equation was selected for combining features in the PoPoSBert model.

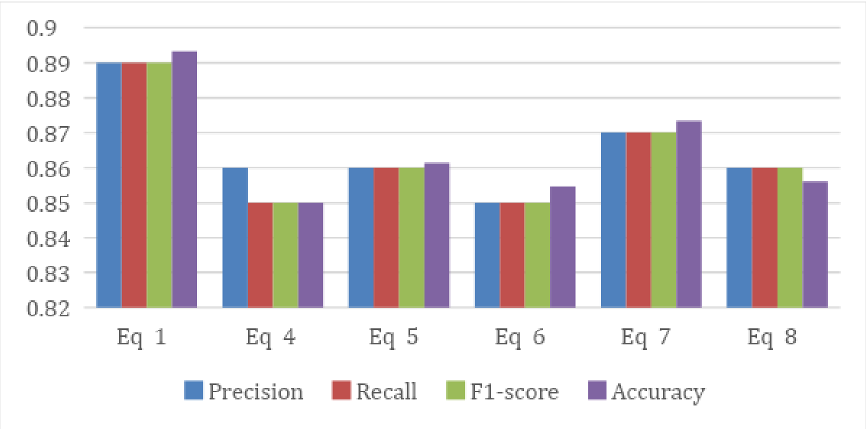
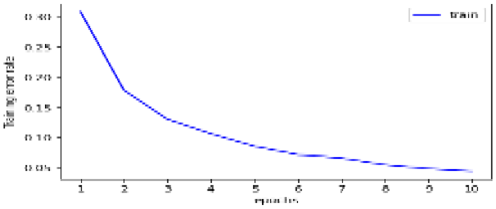


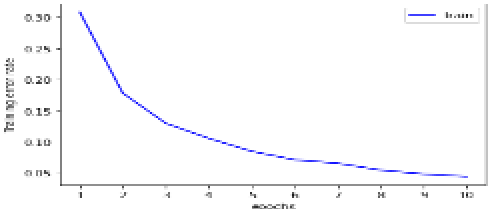
Figure 3:Attention Equations

6.1.3. **Convergence of PoPoSBert**

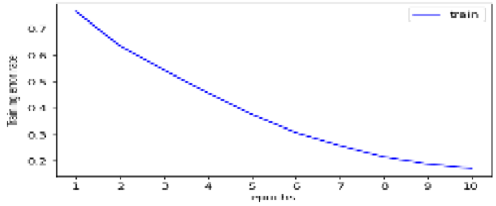
Our study involved conducting experiments to test the convergence of the proposed algorithms. Based on Figure 4(a)-4(d), we can conclude that the "PoPoSBert" algorithm achieves at least a local minimum within the 10 epochs. In general, the experimental results affirm the convergence of the PoPoSBert algorithm and its capacity to learn from the available data, leading to improved performance over time.



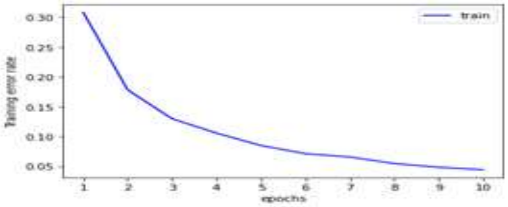
IMDB Error rate (2-class)



SST-2 Error rate (2-class)



SamEval Error rate (3-class)



Twitter Error rate (2-class)

DISCUSSION

This research proposed a novel feature selection and extraction method based on PoS tags and attention concept that is used together with the BERT model and analyses its performance against various sentiment analysis models on different datasets. The PoPoSBert model that we proposed has been found to outperform other state-of-the-art models on every dataset owing to its advanced feature representation techniques, which include the use of transformers, attention mechanisms, and part-of-speech tags to detect polarity words. Specifically, the use of PoS tags to identify polarity words extended in PoPoSBert to have detailed polarity-based dictionaries for categorical weighting. On the other hand, when compared to BERT-based approaches, state-of-the-art models like tf-idf, word2vec, and GloVe are limited in their ability to fully comprehend the meaning of a word or phrase within a specific context. It is evident in PoPoSBert as well.BERT embedding considers the relationships between words and their surrounding context, it still fails to capture certain significant linguistic features or nuances that may affect the sentiment of the text. Additionally, combining polarity information identified together with contextual feature representation through the attention mechanism allows PoPoSBert to accurately classify sentiments. Furthermore, the results indicate that the SamEval dataset has lower levels of significance compared to other datasets, potentially because it has three classes (i.e., positive, neutral, and negative), making it more challenging to accurately classify text into multiple categories.

7. Conclusion

In sentiment analysis, feature selection and feature extraction techniques face challenges due to the higher feature dimensionality and unstructured nature of the text. In this study, we aimed to identify the high-quality representative features including polarity words that improve the accuracy of sentiment analysis. Specifically, we propose utilizing PoS tags to form polarity-based dictionaries and aid in improving the feature representation. Thereafter, an attention mechanism is utilized to combine this information with basic text feature representation to enhance feature extraction. We empirically select the BERT embedding technique as the basic feature representation to capture syntax, semantic and contextual information. Thereby we introduce a new and innovative algorithm for sentiment classification "PoPoSBert" which utilized part of speech tagging based polarity words identification that combined with the Bert embedding using the attention mechanism.

The proposed model PoPoSBert successfully extracts high-quality features by combining polarity words through PoS tagging and contextual feature representation through Bert embedding. Our proposed algorithm significantly enhances the accuracy of sentiment analysis. We conducted a comparison between our model and existing state-of-the-art methods using movie review and Twitter datasets. The evaluation results clearly demonstrate that our framework surpasses other relevant baselines and achieves higher scores in terms of evaluation metrics. As a result, the PoPoSBert algorithm we propose is well-suited for real-world applications involving sentiment analysis.

We are to investigate other feature selection and extraction in future. Specifically, we hope to consider other contextualized word embedding techniques, such as Elmo to further enhance feature selection and extraction. Furthermore, there is an opportunity for future research to focus on enhancing the quality of PoS tagging and exploring its influence on feature representation. This would require developing more sophisticated techniques for PoS tagging.

8. REFERENCES

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*. https://doi.org/10.1162/tacl_a_00051/1567442/tacl_a_00051.pdf

Dashtipour, K., Soujanya Poria, •, Hussain, • Amir, Cambria, E., Ahmad, •, Hawalah, Y. A., Gelbukh, • Alexander, & Zhou, • Qiang. (2015). Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques. *Cognitive Computation*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186.

Hameed, Z., & Garcia-Zapirain, B. (2020a). Sentiment Classification Using a Single-Layered BiLSTM Model. *IEEE Access*, *8*, 73992–74001. https://doi.org/10.1109/ACCESS.2020.2988550

Hameed, Z., & Garcia-Zapirain, B. (2020b). Sentiment Classification Using a Single-Layered BiLSTM Model. *IEEE Access*, *8*, 73992–74001. https://doi.org/10.1109/ACCESS.2020.2988550

Liu, S., & Forss, T. (2014). Combining N-gram based Similarity Analysis with Sentiment Analysis in Web Content Classification. *Special Session on Text Mining. Vol. 2. SCITEPRESS*, 530–537.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics,* , 142–150.

Melamud, O., Goldberger, J., & Dagan, I. (2016). context2vec: Learning Generic Context Embedding with Bidirectional LSTM. *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL),* 51–61. http://www.cs.biu.ac.il/nlp/resources/

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv Preprint ArXiv:1301.3781* . http://arxiv.org/abs/1301.3781

Munikar, M., Shakya, S., & Shrestha, A. (2019, October 4). Fine-grained Sentiment Classification using BERT. *Artificial Intelligence for Transforming Business and Society (AITB). Vol. 1. IEEE*. http://arxiv.org/abs/1910.03474

Nathani, D., Chauhan, J., Sharma, C., & Kaul, M. (2019). Learning Attention-based Embeddings for Relation Prediction in Knowledge Graphs. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4710–4723.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment Classification using Machine Learning Techniques*. EMNLP. http://www.cs.cornell.edu/people/pabo/movie-review-data/.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543. http://nlp.

Pontiki, M., Papageorgiou, H., Galanis, D., Androutsopoulos, I., Pavlopoulos, J., & Manandhar, S. (2014). SemEval-2014 Task 4: Aspect Based Sentiment Analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation*, 27–35. http://alt.qcri.

Rehman, A. U., Malik, A. K., Raza, B., & Ali, W. (2019). A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis. *Multimedia Tools and Applications*, *78*(18), 26597–26613. https://doi.org/10.1007/s11042-019-07788-7

Samha, A. K., Li, Y., & Zhang, J. (2014). ASPECT-BASED OPINION EXTRACTION FROM CUSTOMER REVIEWS. *ArXiv Preprint ArXiv:1404.1982*.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. http://nlp.stanford.edu/

Srividya, K., & Mary Sowjanya, A. (2019). Aspect based sentiment analysis using POS tagging and TFIDF. *International Journal of Engineering and Advanced Technology*, *8*(6), 1960–1963. https://doi.org/10.35940/ijeat.F7935.088619

Subhashini, L. D. C. S., Li, Y., Zhang, J., Atukorale, A. S., & Wu, Y. (2021a). Mining and classifying customer reviews: a survey. *Artificial Intelligence Review*, *54*(8), 6343–6389. https://doi.org/10.1007/s10462-021-09955-5

Subhashini, L. D. C. S., Li, Y., Zhang, J., Atukorale, A. S., & Wu, Y. (2021b). Mining and classifying customer reviews: a survey. *Artificial Intelligence Review*, *54*(8), 6343–6389. https://doi.org/10.1007/s10462-021-09955-5

Subrahmanian, V. S., & Reforgiato, D. (2008). *AVA: Adjective-Verb-Adverb Combinations for Sentiment Analysis*. IEEE Computer Society. www.computer.org/intelligent

Tam, S., Said, R. Ben, & Tanriöver, Ö. (2021). A ConvBiLSTM Deep Learning Model-Based Approach for Twitter Sentiment Classification. *IEEE Access*, *9*, 41283–41293. https://doi.org/10.1109/ACCESS.2021.3064830

Tubishat, M., Idris, N., & Abushariah, M. A. M. (2018). Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges. *Information Processing and Management*, *54*(4), 545–563. https://doi.org/10.1016/j.ipm.2018.03.008

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *31st Conference on Neural Information Processing Systems*. http://arxiv.org/abs/1706.03762

Wang, J.-H., Liu, T.-W., Luo, X., & Wang, L. (2018). An LSTM Approach to Short Text Sentiment Classification with Word Embeddings. *The 2018 Conference on Computational Linguistics and Speech Processing*, 214–223.

Yan, Z., Xing, M., Zhang, D., & Ma, B. (2015). EXPRS: An extended pagerank method for product feature extraction from online consumer reviews. *Information and Management*, *52*(7), 850–858. https://doi.org/10.1016/j.im.2015.02.002

Zhao, L., Liu, Y., Zhang, M., Guo, T., & Chen, L. (2021). Modeling label-wise syntax for fine-grained sentiment analysis of reviews via memory-based neural model. *Information Processing and Management*, *58*(5). https://doi.org/10.1016/j.ipm.2021.102641