Towards Optimal Sentiment Analysis Model Selection: Assessing Performance Across Varied Text Types and Embedding Methods

Madhurima Dasi, Swathi Mounika Bheemana and Venkat Kanakamedala

Finance Decision Sciences and Real Estate Department, Wichita State University

Abstract

Text classification plays a crucial role in natural language processing, encompassing applications such as sentiment analysis, spam filtering, and topic identification. This study aims to evaluate and compare the performance of diverse machine learning models on text classification tasks across multiple domains, including tweets, movie reviews, and Yelp reviews. The models investigated comprise logistic regression with TF-IDF vectorization, LSTM with trainable embeddings, LSTM with pre-trained GloVe embeddings, and LSTM with trainable embeddings and average pooling. The research assesses these models' performance using standard metrics like accuracy and F1-score, highlighting the hyperparameters and architectural factors influencing their effectiveness. Furthermore, this study emphasizes the benefits of using advanced techniques, including pre-trained embeddings, pooling strategies, transfer learning, and large language models (BERT, RoBERTa, GPT-3, and DeBERTa v3) for sentiment analysis. By adopting these approaches, researchers, businesses, and organizations can improve sentiment analysis accuracy and efficiency across various domains and text types, informing optimal text classification strategies and enabling more effective public sentiment analysis.

Keywords: Sentiment analysis, roBERTa, DeBERTa v3, TF-IDF, GPT, BERT Trainable embeddings, GloVe embeddings, Average pooling

---

Introduction:

Sentiment analysis, alternatively referred to as opinion mining, has emerged as a crucial component in the examination and interpretation of human emotions and opinions present in a variety of textual formats. These formats encompass informal social media posts, formal news articles, and reviews. With the exponential growth of digital text, the significance of accurate sentiment analysis techniques in areas such as marketing, customer support, and product development is becoming increasingly evident. The progress made in the field of Natural Language Processing (NLP) has facilitated the development of advanced models capable of capturing the subtleties of sentiment in diverse text types and various levels of formality.

On the other hand, BERTweet [4] and DeBERTa v3 [5] are other advanced transformer-based models. While BERTweet is designed for sentiment analysis on Twitter data, DeBERTa v3 uses a disentangled attention mechanism and relative positional encoding to achieve top-tier results across a wide range of NLP tasks [5]. While both models excel in various contexts, they may not be as specialized for Twitter data as RoBERTa [6].

In addition to these transformer-based models, other techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) combined with Logistic Regression [2], trainable embeddings, and pre-trainable GloVe embeddings [10] have also been employed in sentiment analysis tasks. These methods, particularly TF-IDF with Logistic Regression, have proven effective for traditional machine learning approaches to sentiment analysis [2]. The incorporation of trainable embeddings, either independently or in conjunction with pre-trainable GloVe embeddings, has further enriched the feature representation of text data, enabling models to better discern sentiment [10]. Moreover, the use of average pooling techniques with trainable embeddings has been demonstrated to enhance the overall performance of sentiment analysis models [11].

These recent advancements underscore the ongoing refinement of sentiment analysis models and their adaptability to the ever-evolving domain of digital text. Consequently, these models serve as indispensable tools for researchers, businesses, and organizations seeking to glean valuable insights into public sentiment and emotional reactions across a wide range of text types and levels of formality.

Literature review:

Recent advancements in sentiment analysis have led to the development of various techniques focusing on feature extraction and sentiment classification [1]. A common feature extraction process encompasses data pre-processing, TF-IDF, and selection methods such as Odds Ratio and Chi-Square [2]. Sentiment classification methods generally fall into two categories: lexicon-based and machine-learning-based approaches, which involve support vector machines, neural networks, and trainable Bayesian networks [1].

In 2016, Li proposed a model utilizing Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) layers [3]. RNNs account for the time-order structure of the entire text, while LSTM layers address the issue of long intervals between related previous texts and the current position [3]. Li's model demonstrated improved performance compared to traditional RNNs [3]. However, its applicability to texts with varying lengths and formality remains uncertain, as it was only tested on English movie reviews. The present project aims to compare three models (logistic regression with TF-IDF, RNN with LSTM layer and word embedding, and RNN with LSTM layer, word embedding, and average pooling) on three datasets containing texts of different lengths and formality to determine the best-performing model.

RoBERTa (Robustly optimized BERT pretraining approach) is a modified version of the BERT model, which has demonstrated significant improvements in various NLP tasks [6]. RoBERTa was first introduced by Liu et al. in their paper "RoBERTa: A Robustly Optimized BERT Pretraining Approach" published in 2019 [6].

BERTweet, a pre-trained language model specifically designed for sentiment analysis on Twitter data, leverages the RoBERTa architecture and a vast amount of tweets for pre-training [4]. This model demonstrates remarkable performance in capturing the unique language patterns and nuances present in short, informal texts like tweets.

DeBERTa v3 (Decoding-enhanced BERT with Disentangled Attention), an enhanced version of the BERT model, features a disentangled attention mechanism and relative positional encoding [5]. DeBERTa v3 has achieved state-of-the-art results on various NLP tasks, including sentiment analysis, by improving the representation of words and the understanding of their relationships in the text.

In this literature review, we examine the existing body of research on sentiment analysis models and techniques, with a particular focus on the recent advancements brought forth by transformer-based models such as RoBERTa, BERTweet, and DeBERTa v3. In our analysis, we focused on the performance of RoBERTa, as it was the only pre-trained model among the three specified that we managed to use. We examined its effectiveness in diverse textual contexts, including texts with varying lengths and formality levels, to evaluate how well it captures and interprets human emotions and opinions. Although BERTweet and DeBERTa v3 were not utilized in this study, our findings using RoBERTa offer valuable insights into the capabilities of state-of-the-art transformer-based models in sentiment analysis tasks.

Furthermore, we discuss traditional machine learning approaches to sentiment analysis, such as TF-IDF combined with Logistic Regression, and the incorporation of trainable and pre-trained embeddings, including GloVe embeddings. The review also covers the utilization of average pooling techniques with trainable embeddings to enhance the performance of sentiment analysis models.

Methods:

Data Collection and Pre-processing:

Tweets dataset – In the Twitter dataset we have two files which are test and train with 3535 and 27481 instances, respectively with an average text length of 7.5 characters. In this dataset we have three classes namely negative, neutral, and positive which are defined as 0, 1 and 2 respectively.

IMDB movie reviews dataset – In the IMDB movie reviews dataset we have three files which are train, test, and validation with a total of 50003 instances, with an average text length of 136.2 characters. In this dataset we have two classes namely negative and positive which are defined as 0 and 1 respectively.

Yelp reviews dataset – In the Yelp reviews dataset we have two Train and Test files with 598000 instances, with an average text length of 72.2 characters. In this dataset we have two classes namely negative and positive which are defined as 1 and 2 respectively.

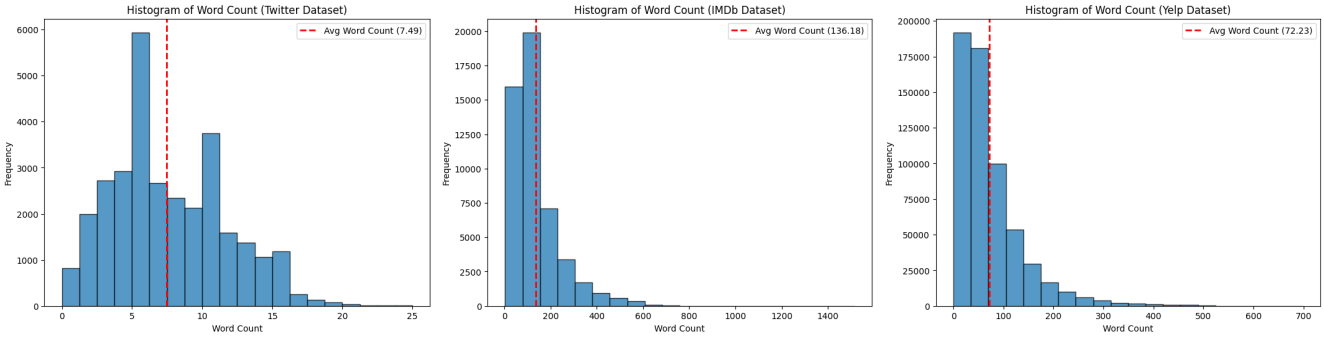

Figure 1: Processed Text Length Histograms

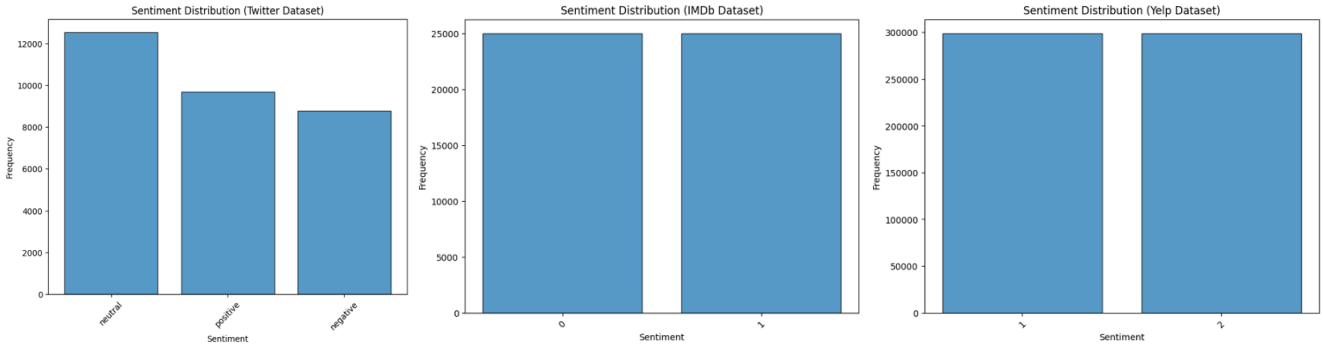

Figure 2: Sentiment Class Distribution

Pre-processing:

As a part of preprocessing, we used a pre-trained RoBERTa model which is fine-tuned on sentiment analysis and a tokenizer for Twitter Tweets data along with the below define process.

For IMBD and Yelp reviews, the preprocessing steps are implemented using Python libraries, including NLTK and contractions. The text preprocessing function consists of several steps. Expand contractions is used for converting contractions to their full form to standardize the text.

Remove punctuation and stress marks for removing any special characters, punctuation, and stress marks from the text. Tokenize for breaking down the text into individual words. Lemmatize for converting words to their base form. Remove stop words for removing common words such as "the", "and", and "a" that do not carry significant meaning but excluding negation words.

The text preprocessing function is then applied to each instance in the training and test sets, and the preprocessed text is saved as a new column in the corresponding Data Frame.

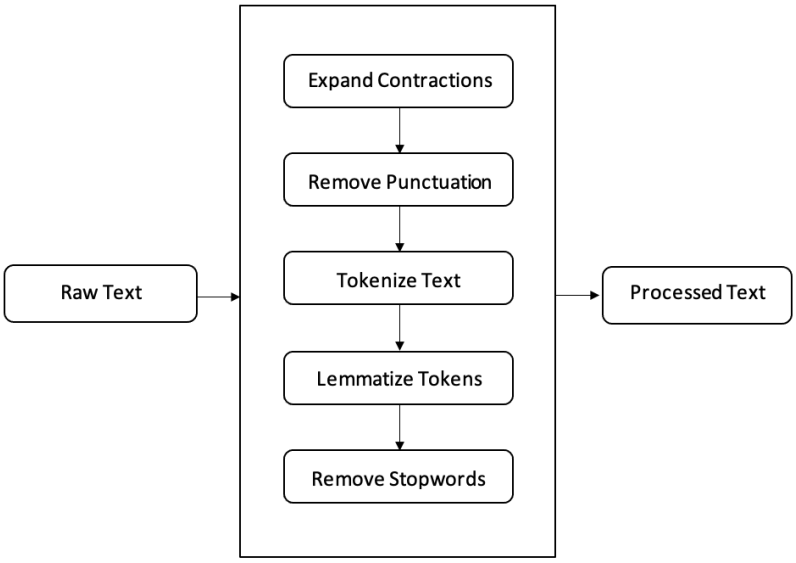

Figure 3: Data Pre-processing

This study employs a diverse set of methodologies to compare the performance of different sentiment analysis techniques on various text types and formality levels. First, a logistic regression model with Term

Frequency-Inverse Document Frequency (TF-IDF) is implemented, a widely-used baseline technique for transforming text data into numerical features for classification tasks. Second, a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) layers and trainable word embeddings is used, allowing the model to learn contextual representations from the text while considering the sequential nature of language.

Third, another RNN with LSTM layers is applied, but this time utilizing pre-trained word embeddings, such as GloVe, which leverage large-scale corpora to generate word representations that capture semantic and syntactic relationships. This approach can provide a stronger starting point for the model by incorporating prior knowledge of language structure. Finally, a fourth model is developed, which comprises an RNN with LSTM layers, adjustable word embeddings, and a GlobalAveragePooling1D layer. Incorporating the GlobalAveragePooling1D layer aids in summarizing the information extracted by the LSTM layers across all time steps, which may improve the model's capability to identify the sentiment present in the provided text.

Each of these models will be trained and evaluated on three distinct datasets, comprising texts of different lengths and varying degrees of formality. By comparing their performance, this study aims to identify the most effective sentiment analysis technique for diverse textual contexts, contributing valuable insights to the ongoing development of sentiment analysis methodologies in the field of Natural Language Processing.

Results/discussion:

Across the four models evaluated on each dataset, we observed that LSTM with pre-trained GloVe embeddings, Nested CV Logistic Regression with TF-IDF, and LSTM with trainable embeddings yielded the highest accuracy rates for Twitter, IMDB, and Yelp datasets, respectively.

Accuracy			
Model	**Twitter**	**IMDB**	**Yelp**
Nested Cross Validation using Logistic with TF-IDF	0.71	0.90	0.93
LSTM with trainable embeddings	0.69	0.86	0.94
LSTM with pre-trained GloVe embeddings	0.72	0.87	0.93
LSTM with trainable embeddings and average pooling	0.69	0.86	0.93

Table 1: Accuracy of all the models for three datasets

For the Twitter dataset, the highest accuracy of 72% was attained using the LSTM model with pre-trained GloVe embeddings. In this instance, we employed predicted sentiment labels from the RoBERTa model as the target variable for training a sentiment classification model with pre-trained GloVe, leading to enhanced performance. Regarding the IMDB dataset, the highest accuracy of 90% was achieved utilizing Nested CV [7]. In this case, we employed logistic regression combined with TF-IDF vectorization to transform the text data, contributing to improved performance. For the Yelp dataset, the highest accuracy of 94% was reached using the LSTM model with trainable embeddings. Here, we implemented the Keras Sequential API, incorporating an Embedding layer followed by an LSTM layer, a Dense layer, a Batch Normalization layer, and an Activation layer, which resulted in better model performance.

The other metrics (precision, recall, and F1-score) for the most accurate models for each dataset are presented below:

Model & Dataset	Class	Precision	Recall	F1- Score
LSTM with pre-trained GloVe embeddings for Twitter	Positive	0.8	0.75	0.74
	Neutral	0.69	0.54	0.6
	Negative	0.66	0.85	0.78
Nested Cross Validation using Logistic with TF-IDF for IMDB	Positive	0.88	0.92	0.9
	Negative	0.91	0.88	0.9
LSTM with trainable embeddings for Yelp	Positive	0.95	0.93	0.94
	Negative	0.94	0.95	0.94

Table 2: Precision, Recall and F1-Scores for best performing model for each dataset.

The performance of the models was influenced by tuning the following hyperparameters for each model:

LSTM with trainable embeddings: embedding_dim, lstm_units, vocab_size, initial_learning_rate, decay_steps, decay_rate, batch size, and number of epochs used during training.

Nested Cross Validation using Logistic with TF-IDF: Logistic regression model was tuned by varying the 'C' hyperparameter with a single value of 10 using 5 outer folds and 5 inner folds. The TfidfVectorizer was employed to convert text data into a numerical format. Additionally, n_splits, shuffle, random_state, penalty, solver, max_iter, and smooth_idf were used.

LSTM with pre-trained GloVe embeddings: embedding_dim, lstm_units, initial_learning_rate, decay_steps, decay_rate, and the patience parameter for early stopping to prevent overfitting were used.

LSTM with trainable embeddings and average pooling: embedding_dim, lstm_units, initial_learning_rate, decay_steps, decay_rate, batch size, and number of epochs used during training were considered as hyperparameters affecting performance. Other architecture-related hyperparameters, such as the use of bidirectional LSTMs or dropout layers, may also impact performance.

The results indicate that using pre-trained embeddings and keeping them fixed led to better performance compared to trainable embeddings.

Results comparison:

Our results showed a slight improvement in accuracy for the IMDB dataset, achieving a 90% accuracy compared to the previous works 89% [9]. However, we did not achieve the same accuracy for the Twitter dataset. For the Yelp dataset, our findings were consistent with the previous study across all models.

Furthermore, we observed improvements in precision and recall for both positive and negative classes in the IMDB dataset. Specifically, precision for the negative class increased from 89% to 91%, while recall for the positive class increased from 89% to 92%. Additionally, the F-1 score for both classes showed a slight increase.

For the Yelp dataset, our LSTM model with trainable embeddings achieved a precision of 95% for the positive class and a recall of 95% for the negative class, which is a slight improvement over the previous study results of 94%.5.

Conclusion:

In conclusion, this study offers an in-depth examination of diverse machine learning models applied to text classification tasks across multiple domains, such as tweets, movie reviews, and Yelp reviews. The results emphasize the significance of pre-trained embeddings, which demonstrate superior performance when compared to trainable embeddings. Furthermore, the inclusion of pooling strategies, such as average pooling, max pooling, or dynamic pooling, potentially enhances model effectiveness.

Future Work and Recommendations:

Transfer learning has emerged as a prominent technique in sentiment analysis, offering the potential to significantly improve model performance by leveraging pre-trained language models such as BERT, RoBERTa, GPT-3, and DeBERTa v3. Researchers should consider incorporating transfer learning in their approaches to sentiment analysis, as it allows for the fine-tuning of these pre-trained models on specific tasks without the need to train a model from scratch. By utilizing the vast knowledge base encoded in these pre-trained models, transfer learning can lead to more accurate and efficient sentiment analysis across various domains and text types [8].

Large language models, such as GPT-3 and DeBERTa v3, have demonstrated remarkable capabilities in generating human-like text and understanding complex language structures. As part of future research in sentiment analysis, it is important to explore the potential of these large language models in capturing subtle nuances and context-dependent emotions that may be overlooked by traditional machine learning methods. Incorporating these large language models into the domain of sentiment analysis could result in more sophisticated models capable of handling a wider range of text types, languages, and sentiment expressions. By focusing on these advanced techniques, researchers, businesses, and organizations can better harness the power of sentiment analysis and derive more accurate insights into human emotions and opinions across diverse textual data.

References:

1. Liu, B., 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), pp.1-167.
2. Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3(Mar), pp.1289-1305.
3. Li, X. and Roth, D., 2016. Neural sentiment classification with user and product attention. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp.1650-1659.
4. Nguyen, D.Q., Vu, T., and 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.10-15.
5. He, P., Liu, X., Chen, W. and Gao, J., 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In Proceedings of the 2021 Conference of the Association for Computational Linguistics (ACL), 2021.
6. 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.
7. Aletras, N., Chamberlain, B.P., and Hepple, M., 2015. Modeling the Severity of Complaints in Social Media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1323-1329). Association for Computational Linguistics.
8. Ruder, S., 2019. Neural Transfer Learning for Natural Language Processing. Ph.D. thesis, National University of Ireland, Galway.
9. Chen, J., 2021. Sentiment Analysis behind Text with Different Length and Formality.
10. Pennington, J., Socher, R., and 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.
11. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A., 2017. Supervised learning of universal sentence representations from natural language inference data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp.670-680.