Unveiling the Verdict - They Said We Did Good (Or Bad?) Sentiment Analysis Using BERT and Word2Vec with RNN-based Classification

Nishika Abeytunge
W266: Natural Language Processing
UC Berkeley School Of Information
nishika.abey@ischool.berkeley.edu

Abstract

This research presents a comprehensive analysis of multiclass sentiment analysis models by comparing Word2Vec-based LSTM, BiLSTM, and BERT-based models. The study aims to determine the model that best captures sentiment nuances from text data. Through thorough evaluation, the BERT-BiLSTM hybrid model emerges as the top performer, achieving a precision and F1-score of 0.82, coupled with a recall of 0.81. This outcome underscores the synergy between BERT's contextual embeddings and BiLSTM's sequential understanding, making the BERT-BiLSTM model the optimal choice for accurate and nuanced sentiment classification tasks.

1 Introduction

In today's technology-driven world, expressing feelings and ideas has become increasingly common through various digital platforms, surpassing traditional face-to-face communication. Just as individuals use the internet to express sentiments, businesses seek to gain insights into consumer perceptions of their products or services. However, the vast amount of available data and its time sensitivity make it challenging and time-consuming to interpret each comment individually.

The utilization of Natural Language Processing (NLP) tools and techniques in sentiment analysis is a well-established practice. These algorithms efficiently determine the contextual polarity of the text, which is vital for businesses to evaluate their performance and identify areas for improvement or further enhancement in the market. Accurate sentiment classifications play a crucial role in making informed decisions and staying competitive in the industry.

The objective of my research is to conduct a comparative analysis of various models for text classification tasks using word embeddings and contextual embeddings. More specifically, the study involves comparing a Word2Vec-based LSTM and BiLSTM model with a BERT-based model equipped with a classification token (CLS) pooling layer and BiLSTM. The primary focus is to evaluate which model exhibits superior performance concerning accuracy, generalization, and efficiency in the context of text classification tasks.

2 Project Overview

2.1 Hypothesis

The hypothesis for this research is that the BERT-based model with the CLS token pooling and BiLSTM will outperform the Word2Vec-based LSTM and BiLSTM models in terms of accuracy. This is based on the rationale that BERT's contextual embeddings capture more fine-grained semantic information, and the CLS token pooling helps in aggregating the context for classification tasks. Furthermore, the

inclusion of BiLSTM layers provides better understanding of sequential context in the BERT-based model, enhancing its performance in text classification tasks.

2.2 Related work

Different approaches and algorithms in sentiment analysis are evolving and getting better everyday. They are improved by better quality data and more diverse training data. Researchers also invent new algorithms that can use this improved data more effectively.

Machine Learning Approaches like Support Vector Machines (SVM), Naive Bayes, and Logistic Regression existed since the mid 2000's and deep learning approaches in natural language processing like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks evolved a few years later. Now one of the most prominent is hybrid Approaches in which dictionary based methods are combined with machine learning based approaches which are capable of gaining the advantages of improved performance, flexibility, handling domain specific challenges and many more.

To the best of our knowledge, a deep learning-based method in which a unified feature set which is representative of word embedding, sentiment knowledge, sentiment shifter rules, statistical and linguistic knowledge, has not been thoroughly studied for a sentiment analysis[1].

In the research paper published in 2019 proposed a new deep-learning-based method to classify a user's opinion expressed in reviews (called RNSA) which employs the Recurrent Neural Network (RNN) which is composed of Long Short-Term Memory (LSTM). Here they used bag-of-word (BOW) as the embedding method and fed it into the RNN-LSTM layer to generate a sentence-wide feature set. For the highest F_measuee of 0.7246 they were able to get Precision of 0.8404 and Recall 0.6369. Model also performed better than other CNN and Recursive methods they used in the classification [1].

In 2020 a team researched on how BERT transfer learning strategy can be used for Bengala sentiment analysis using a model developed with CNN- BiLSTM and compared against different embeddings such as Word2Vec, GloVe, and fastText. Bangla-BERT model was able to achieve the highest accuracy among all [2].

In a research done in 2021, Sentiment Analysis Algorithm Based on BERT and Convolutional Neural Network, they used Word2Vec and BERT with support vector machine (SVM) and CNN models with different fine tunings. They observed that the CNN model with BERT embeddings resulted in the highest F1 score[3].

2.2 Dataset Used

The Twitter US Airline Sentiment dataset sourced from Kaggle serves as the foundational dataset for sentiment analysis and model comparison purposes (https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiments). This dataset comprises tweets directed at six prominent US airlines, collected during February 2015. With a total of 9,178 negative, 3,099 neutral, and 2,363 positive comments, the dataset demonstrates an inherent imbalance. It's worth noting that a study conducted by Zendesk, a software company, in 2013, revealed a tendency for individuals to share negative experiences more prominently than positive ones. Consequently, I consider this dataset as reflective of real-world scenarios, encapsulating the diversity of sentiments encountered in online discussions about airline experiences.

2.3 Data Preparation and Embeddings

Text data in the dataset is scraped data from Twitter and was as is. Eg. “@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse”. Data cleaning included removal of html tags, replacement of contractions, removal of numbers and then using NLTK library to remove Stopwords, tokenization and lemmatization.

Length of the texts were mostly between 10-30 words in length. Therefore I used the maximum length of the texts as 35 and used padding as for the shorter texts.(Fig.1)

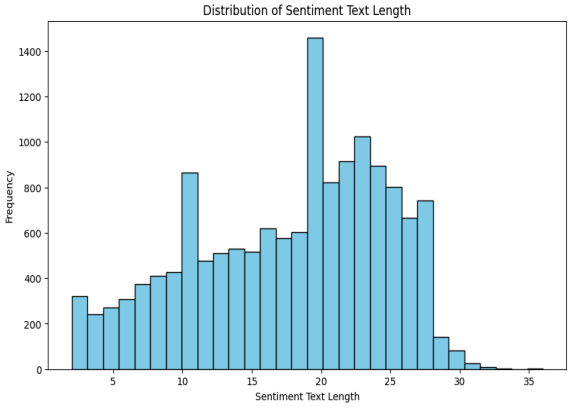

Fig. 1 Distribution of Sentiment Text Length

In this research I used both Word2Vec and BERT encoders. Word2Vec is a feature expansion which aims to process non-numeric data into numeric data. Word embedding Word2Vec is often used as a feature expansion for some text classifications[7]. This model operates by constructing a vocabulary from the provided training data and subsequently analyzing it to create vector representations for each word. It is employed in text classification due to its capability to capture semantic and contextual relationships between words through word embedding. There are two models available for Word2Vec: Skip-Gram and Continuous Bag of Words (CBOW).

BERT, short for "Bidirectional Encoder Representations from Transformers," provides a pre-trained bidirectional representation of text input, incorporating both left and right context conditioning. Thanks to its pre-training, the model can undergo faster fine-tuning for specific tasks. During the training phase, it may learn from both the left and right sides of the tokens, known as bidirectional learning, which is helpful when learning about the many meanings of the same term[8].

3 Model

3.1 Baseline Model

The baseline models for this study consist of two variants: a Word2Vec-based Long Short-Term Memory (LSTM) model and a Bidirectional LSTM (BiLSTM) model. The Word2Vec-based LSTM employs pre-trained word embeddings to capture semantic relationships within the input text. The sequence of word embeddings is processed through an LSTM layer followed by a global max pooling operation, providing a fixed-length representation of the sequence. Subsequent dense layers with dropout regularization are applied, culminating in a softmax activation layer for classification. Similarly, the Bidirectional LSTM model enhances the sequence understanding by employing a BiLSTM layer, which processes the input sequence in both forward and backward directions. The rest of the architecture follows a similar pattern, encompassing pooling, dense, and dropout layers. These baseline models serve as a comparative benchmark to evaluate the performance improvements brought by subsequent enhancements in accuracy, generalization, and efficiency for text classification tasks.

3.2 Potential of BERT and BiLSTM Integration

I believe that the model with BiLSTM integrated into BERT will exhibit enhanced performance in multiclass text classification. In this comparative evaluation, the research encompasses two advanced models alongside the previously discussed Word2Vec (W2V) models. The first of these models centers on a simple yet effective multiclass classification architecture. This architecture leverages the Pooler Output from a pre-trained BERT model, with all BERT layers set as trainable. Input sequences are processed through BERT, utilizing the pooler output to capture contextual sentence representations. A subsequent dense hidden layer, featuring ReLU activation and dropout for regularization, is employed, culminating in a softmax-based classification layer. While this model capitalizes on BERT's contextual embeddings, the second model further augments this approach.

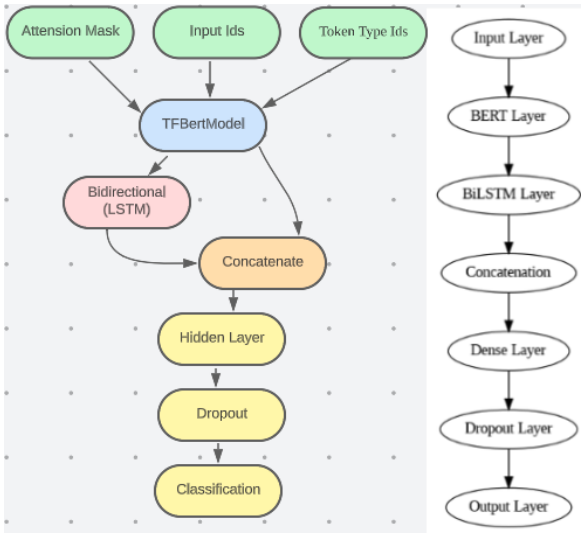

Fig.2 Model Architecture

The second model (Fig.2) integrates BERT's contextual embeddings with the sequence understanding capabilities of Bidirectional Long Short-Term Memory (BiLSTM). Through the fusion of these strengths, this hybrid architecture aims to provide an effective solution for multiclass classification tasks, offering improvements in accuracy, generalization, and efficiency.

4 Results

The results from these four models, as shown in Table 1, encompass precision, recall, and F1 scores, calculated based on the Weighted Average.

Meodel	Precision	Recall	F1-Score
Word2Vec- LSTM	0.79	0.79	0.79
Word2Vec-BiLSTM	0.79	0.79	0.79
BERT-CLS	0.81	0.81	0.81
BERT-BiLSTM	0.82	0.81	0.81

Table.1 Results

6 Conclusion and Discussion

In conclusion, the evaluated models have demonstrated consistent and competitive performance across precision, recall, and F1-score metrics. Both the Word2Vec-based LSTM and BiLSTM models exhibit commendable precision, recall, and F1-score values, each achieving a score of 0.79. The BERT-CLS model showcases a slight improvement with a precision, recall, and F1-score of 0.81, suggesting its effective capture of contextual information and classification accuracy. Notably, the BERT-BiLSTM model emerges as the top performer, displaying the highest precision and F1-score of 0.82, and an equally strong recall of 0.81. This underscores the potential synergy between BERT's contextual embeddings and BiLSTM's sequence understanding. In light of these findings, the BERT-BiLSTM model stands out as the optimal choice for multiclass sentiment analysis, offering a well-balanced blend of precision, recall, and F1-score for accurate classification and nuanced sentiment representation.

A notable finding of this study is the significant increase in training time for BERT models compared to Word2Vec models. While the investigation focused on a dataset of 14,640 sentiments within a specific business domain, the generalizability of this observation could be enhanced by expanding the dataset's size and incorporating diverse data sources. Additionally, tailoring BERT encoders to specific industries could yield more refined results. Leveraging processors with higher GPU capabilities could also contribute to finer-tuned outcomes during model training. These avenues offer potential for optimizing training efficiency and improving the models' overall performance.

References

[1] Asad Abdi, Siti Mariyam Shamsuddin, Shafaatunnur Hasan, Jalil Piran, Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion, Information Processing & Management, Volume 56, Issue 4, 2019

[2] Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning: Nusrat,J., Abdullah, As., Md, K., Saydul, A., Anupam, K., Mehedi, M., & Mohammed, B.(2022). Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning

[3] R. Man and K. Lin, "Sentiment Analysis Algorithm Based on BERT and Convolutional Neural Network," 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 2021, pp. 769-772, doi: 10.1109/IPEC51340.2021.9421110

[4] Dimensional Research. (April 2013). Customer service and business results: A survey of customer service from mid-size companies. Retrieved from: http://cdn.zendesk.com/resources/whitepapers/Zendesk_WP_Customer_Service_and_Business_Results.pdf

[5] W. Yue and L. Li, "Sentiment Analysis using Word2vec-CNN-BiLSTM Classification," 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), Paris, France, 2020, pp. 1-5, doi: 10.1109/SNAMS52053.2020.9336549

[6] K. L. Tan, C. P. Lee, K. S. M. Anbananthen and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," in IEEE Access, vol. 10, pp. 21517-21525, 2022, doi: 10.1109/ACCESS.2022.3152828

[7] Rayhan Rahmanda, & Erwin Budi Setiawan. (2022). Word2Vec on Sentiment Analysis with Synthetic Minority Oversampling Technique and Boosting Algorithm. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 6(4), 599 - 605. https://doi.org/10.29207/resti.v6i4.4186

[8] Aksh Patel, Parita Oza, Smita Agrawal, Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model, Procedia Computer Science, Volume 218, 2023 Pages 1245-1259, ISSN 0306-4573