DistilBERT vs. VADER: Unveiling the Depths of Movie Review Sentiment Analysis

Sajad Ahmadi Jozdani

Prof. dr. Antal van den Bosch

Introduction

Today millions of online users put their opinions on the internet about various topics. As a result, other users can know the strengths and weaknesses of the product from the experiences shared by people on the Web, which can be useful for them in making purchasing decisions (Agarwal & Mittal, 2014, 2016). One of the methods to analyze the sentiment behind these opinions is sentiment analysis. Opinion mining or sentiment analysis is known as the mining of behavior, opinions, and sentiments of the text, chat, etc. using natural language processing and information retrieval methods(Neshan & Akbari, 2020).

Sentiment analysis proves valuable for various issues relevant to human-computer interaction professionals, researchers, and individuals in fields like sociology, marketing, psychology, economics, and political science.(Hutto & Gilbert, 2014) .As a result, the advent and measuring the performance of models that can be used to assess the sentiment of these opinions as accurately as possible is an area of interest for both researchers and businesses.

This paper aims to perform comparative sentiment analysis on textual IMDB movie reviews using a machine-learning-based approach and a rule-based approach. As the rule-based approach, VADER lexicon-based algorithm is chosen and as the machine learning-based approach, DistilBERT is used. The motivation behind choosing VADER is that based on Hutto's research (Hutto & Gilbert, 2014), VADER outperformed the other conventional lexicon-based algorithms. In addition, BERT has shown great performance in sentiment analysis (Alaparthi & Mishra, 2021) and DistilBERT, is the cheaper and faster version of BERT while preserving over 95% of BERT's performance (DistilBERT, 2023). DistilBERT also showed better performance (an accuracy score of 91.46%) in sentiment analysis of tweets regarding COVID-19-related hashtags on Twitter(Ranganathan & Tsahai, 2022). As a result, the research question of this study is: "In comparison between DistilBERT and VADER for analyzing the sentiment of movie reviews, which one shows a better performance?".

The study is of value to analytics professionals and academicians working on text analysis as it offers insight into the sentiment classification performance of two different algorithms.

Literature Review

Liu and Zhang (Liu & Zhang, 2012) define sentiment analysis as “the computational study of people’s opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics, and their attributes.” Sentiment analysis is also mentioned as opinion mining in other researches (Sun, Luo, & Chen, 2017; Varathan, Giachanou, & Crestani, 2017). These reviews are referred by potential customers in the market when making purchase decisions(Agarwal & Mittal, 2014). There are studies that show movie reviews can affect the box office performance of movies. Chintagunta et al.(Chintagunta, Gopinath, & Venkataraman, 2010) analyzed the ticket sales of 148 movies released in the United States during a 16-month period and found that the main driver of box office performance for movies is the valence of reviews and not the volume.

While advancements have been achieved in consumer review analysis, criticism persists due to concerns about accuracy and predictability. Goncal et al. (Gonçalves, Araújo, Benevenuto, & Cha, 2013) says “it is unclear which method is better for identifying the polarity (i.e., positive or negative) of a message as the current literature does not provide a method of comparison among existing methods.”. Since that time, various research has been done in different contexts to compare the performance of sentiment analysis methods. Musto et al. (Musto, Semeraro, & Polignano, 2014) found that SentiWordNet has the best performance based on accuracy score (0.59) for measuring the sentiment of microblog posts in comparison to other lexicon-based methods. Musto et al. (Musto et al., 2014) stated that VADER outperformed other lexicon-based methods (F1 and accuracy score = 0.96 and 0.84, respectively) when it comes to measuring the sentiment of social media texts. On IMDB 50K reviews and multiple other datasets, baseline algorithms such as CNNs, KNN, LSTM-CNN and LTSM have achieved an overall accuracy of 96% after optimization via grid search(Joshy & Sundar, 2022). As we see, different studies have found different methods as the best-performing methods in various contexts.

In general, there are four main methods in analyzing sentiment, (1) unsupervised lexicon-based models, e.g., VADER; A sentiment lexicon is a list of lexical features (e.g., words) that are generally labeled according to their semantic orientation as either positive or negative (Liu, 2010). VADER is a rule-based and lexicon-based framework for sentiment analysis, with support for intensity estimation(Borg & Boldt, 2020). (2) traditional supervised machine learning models, e.g., logistic regression; (3) supervised deep learning models, e.g., Long Short-Term Memory (LSTM); and (4) advanced supervised deep learning models, e.g., Bidirectional Encoder Representations from Transformers (BERT)(Alaparthi & Mishra, 2021). BERT was Developed by Devlin et al.(Devlin, Chang, Lee, & Toutanova, 2018) of Google AI Language. It is pre-trained on two unsupervised tasks—masked language modeling and next sentence prediction, thus making it an effective technique for sentiment classification (Trivedi, 2019). In this study, the researcher used DistilBERT, which is a lighter, cheaper, and faster version of the BERT model while preserving over 95% of BERT’s performances(DistilBERT, 2023).

In this study, the researcher only compared the lexicon-based approach (using VADER) and the advanced supervised deep learning approach (using DistilBERT).

Data and Methodology

Data:

This study uses a dataset of 50 thousand movie reviews from the IMDB website. IMDB is “the world's most popular and authoritative source for movie and movie reviews”(IMDb.) Every review in the dataset has been pre-labeled as either positive or negative sentiment, with an equal distribution of positive and negative reviews. Importantly, there are no neutral reviews included in the dataset.

Tools:

The study uses different packages to implement various techniques at each. Data analysis was carried out on a Jupyter Notebook on Google Colab with the help of Python 3.10.12 using T4 GPU runtime. Preprocessing of the reviews was done using the data preprocessing package natural language toolkit (NLTK) 3.8.1 and largely based on the steps recommended therein (Neshan & Akbari, 2020). For data exploration, the Pandas package (Team, 2020) and NumPy(Harris et al., 2020) are used. VADER was accessed using NLTK 3.8.1 (Jiwani, Gupta, & Whig, 2022). The implementation of DistilBERT utilizes Keras's API in TensorFlow version 2.14.0 and the Transformers library version 4.35.2 (Abadi et al., 2016).

Data preprocessing

Text is classified as unstructured data, requiring substantial preprocessing before applying text mining techniques, such as sentiment analysis(Vijayarani, Ilamathi, & Nithya, 2015). The researcher preprocessed the text using the following steps:
(1) the data contains HTML tags like <br /> , so I removed them using the BeautifulSoup module of sb4 4.11.2 library. (2) removal of special characters and brackets like “@”, “&”, and “[“ using the re (Regular expression operations) 3.12.1 library and r'[^a-zA-z0-9\s]' as the regex pattern to be substituted (3) getting the lemma of words, for example, the lemma of “corpora” is “corpus”. (4) lower-casing the words, for example, replacing “Movie” with “movie” (5) removal of stop words like “the”. Stop words are words that happen highly frequently in documents and do not contribute to the meaning (Wilbur & Sirotkin, 1992).
The preprocessed text was then partitioned and analyzed by VADER and DistilBERT models.

Data Partitioning

The data was partitioned into 35% training set(17,500 reviews), 15% validation set(7,500 reviews) and 50% test set(25,000 reviews). This partitioning was because DistilBERT is already trained on large corpora of text and it might overfit the current dataset(Das & Chen, 2007). In addition, It requires high computational power for large training datasets which was beyond the available resources for this study. Furthermore, since a lexicon-based model like VADER doesn’t require training, having a larger portion of data for comparison purposes, which is the main purpose of this study, increases the stability of results.

Model Implementation

The DistilBERT model was compiled by the Adam algorithm. Two different learning rates with values of 1e-3 and 5e-5 were tested using the validation set. The 5e-5 value for learning rate was chosen due to its better performance regarding accuracy score. As the tokenizer, a DistilBertTokenizerFast class with the 'distilbert-base-cased' parameter, padding, truncation and maximum length of 128 was used. Data partitions were then transformed to TensorFlow batched data objects. Finally, the model was trained using 10 epochs.

The VADER model was tested across different thresholds from -0.5 to +0.5 for the compound value to determine the sentiment of reviews using the training set. The compound score is a metric that measures the sum of all the lexicon ratings normalized between -1 and 1(Bonta, Kumaresh, & Janardhan, 2019). Finally, the 0.5 value was chosen as the threshold due to its higher accuracy score. Reviews with compound values higher than 0.5 were considered as positive sentiment and vice versa.

Evaluation Metrics

The metrics used to analyze the performance of models are based on Hutto and Gilbet's research (Hutto & Gilbert, 2014). In addition, a confusion matrix for each model based on test data is plotted. The 2 × 2 confusion matrix is composed of cells that represent various scenarios: true positives (TP), where cases are predicted to belong to the positive sentiment class correctly (TN), where cases are predicted to belong to the negative sentiment class correctly; false positives (FP), where cases are predicted to belong to the positive sentiment class but actually belong to the negative sentiment class; and false negatives (FN), where cases are predicted to belong to the negative sentiment class but in fact belong to the negative sentiment class (James, Witten, Hastie, & Tibshirani, 2013). Only the true positive and true negative cases are predicted correctly.

The accuracy score shows the overall performance of the model across all classes:

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \tag{1}$$

The precision shows the performance of the model in classifying positive sentiment cases.

$$\text{Precision} = TP/(TP + FP) \tag{2}$$

The recall or sensitivity (also called true positive ratio) demonstrates the robustness of the model since it reflects the ability of the model to correctly predict the reviews having a positive sentiment(Alaparthi & Mishra, 2021).

$$\text{Recall} = TP/(TP + FN) \tag{3}$$

The F1 score, which is the harmonic mean of precision and recall, represents both precision and recall in one metric(" F-score," 2023).

$$\text{F1 score} = 2 \ast ((\text{Precision} \ast \text{Recall})/(\text{Precision} + \text{Recall})) \tag{4}$$

Results

The runtime for the DistilBERT model was 26 minutes, while it took 5 minutes for the VADER model. The performance metrics for the two algorithms are given in Table 1. VADER showed a lower overall accuracy score (69.6%) vs. the DistilBERT model (86.7%). Additionally, since successfully predicting positive sentiments is more important than negative sentiments due to its challenges(Alaparthi & Mishra, 2021), the precision and recall metrics for the positive class should be considered. In both metrics, DistilBERT showed a higher performance, specifically in the precision metric, in which it showed a 0.857 score (0.205 higher than VADER's score). This score means that DistilBERT's capability to get a better success rate among those reviews predicted to have a positive sentiment was greater than VADER.

Furthermore, the F1 score for the DistilBERT algorithm was 0.868, which was greater than VADER's value of 0.731. This metric means considering both sides, i.e., precision and recall, the DistllBERT model showed a better performance.

Approach	Algorithm	Accuracy	Precision	Recall	F1 Score
Rule-based	VADER	0.696	0.652	0.833	0.731
Machine learning	DistilBERT	0.867	0.857	0.879	0.868

Table 1: Classification performance outputs

DistilBERT's loss graph, figure 1, shows a decrease after each epoch for the training set, while in the third epoch, it increases for the validation set, until the fifth epoch when it shows a steady decline again.

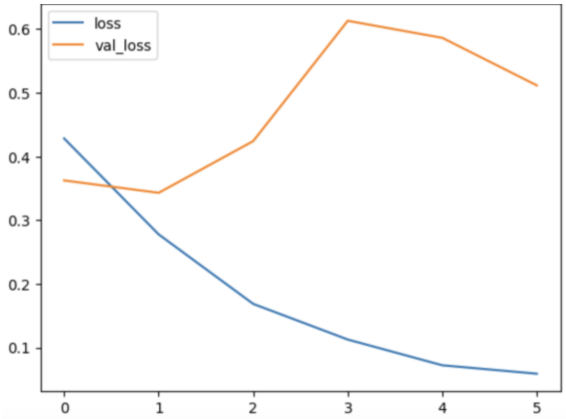

Fig. 1: loss graph of the DistilBERT model

Discussion and Conclusion

This study was performed to compare DistilBERT and VADER in the context of sentiment analysis. The answer to the research question of the study is that the machine-learning approach, using DistilBERT, outperformed the rule-based approach(using VADER) across all the measured scores. This outperformance might be attributed to DistilBERT's pre-trained nature as a Transformer model, which has been trained on large corpora of data. The model benefits from the application of attention mechanisms in its structure, allowing it to comprehend the entire concept of the corpus(Joshy & Sundar, 2022). In contrast, VADER relies on a pre-defined list of words to measure sentiment in the corpus(Elbagir & Yang, 2019)

In comparison to Ranganathan and Tsahai's study(Ranganathan & Tsahai, 2022), the application of DistilBERT in this study got a lower accuracy score (86.7% vs. 91.46%). This might be due to the low number of epochs (10) in this study and not optimizing for different hyperparameters.

Furthermore, the accuracy score of VADER in this study (69.6%) was lower than what was observed in Musto el al.'s research (Musto et al., 2014), in which it was 84%. This difference in results might be due to the difference in the sources of data(ie, social media text vs. movie review text).

Based on my knowledge, it was the first comparison between DistilBERT and VADER for analyzing movie reviews. The results of this study might help analytics professionals and academicians working in the field of sentiment analysis. In addition, since in Chintagunta et al.'s study (Chintagunta et al., 2010), the sentiment of movie reviews showed to be effective in box office performance of the movies, the results of this study might be in the interest of businesses in the field of entertainment and movie production.

Although the researcher tried to make a fair comparison between the models, this study was conducted under some limitations. Firstly, the number of epochs used for training the DistilBERT model was limited due to limitations in computational power. This limitation also affected the search for optimized hyperparameters, in which only one parameter (learning rate) was tested. A greater computational power could help to have a more tuned model with higher performance scores. Second, the loss curve of the DistillBERT model didn't show a declining behavior across all epochs. This might be due to overfitting to the training data at the beginning epochs. Therefore, a study with a higher number of epochs might be helpful to test different implementations of the model. Finally, In this study, a cross-validation approach was not employed during data partitioning and performance evaluation of the DistilBERT model. It is acknowledged that utilizing a cross-validation approach in data partitioning may contribute to the stability of the results(James et al., 2013). For future studies, using this approach might increase the reliability of the results.

References:

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., . . . Isard, M. (2016). *{TensorFlow}: a system for {Large-Scale} machine learning.* Paper presented at the 12th USENIX symposium on operating systems design and implementation (OSDI 16).

Agarwal, B., & Mittal, N. (2014). Semantic feature clustering for sentiment analysis of English reviews. *IETE Journal of Research, 60*(6), 414-422.

Agarwal, B., & Mittal, N. (2016). Prominent feature extraction for review analysis: an empirical study. *Journal of Experimental & Theoretical Artificial Intelligence, 28*(3), 485-498.

Alaparthi, S., & Mishra, M. (2021). BERT: A sentiment analysis odyssey. *Journal of Marketing Analytics, 9*(2), 118-126.

Bonta, V., Kumaresh, N., & Janardhan, N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology, 8*(S2), 1-6.

Borg, A., & Boldt, M. (2020). Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications, 162*, 113746.

Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing science, 29*(5), 944-957.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management science, 53*(9), 1375-1388.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DistilBERT. (2023). DistilBERT. Retrieved from https://huggingface.co/docs/transformers/model_doc/distilbert

Elbagir, S., & Yang, J. (2019). *Twitter sentiment analysis using natural language toolkit and VADER sentiment.* Paper presented at the Proceedings of the international multiconference of engineers and computer scientists.

F-score. (2023). Retrieved from https://en.wikipedia.org/w/index.php?title=F-score&oldid=1174585267

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). *Comparing and combining sentiment analysis methods.* Paper presented at the Proceedings of the first ACM conference on Online social networks.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., . . . Smith, N. J. (2020). Array programming with NumPy. *Nature, 585*(7825), 357-362.

Hutto, C., & Gilbert, E. (2014). *Vader: A parsimonious rule-based model for sentiment analysis of social media text.* Paper presented at the Proceedings of the international AAAI conference on web and social media.

IMDb. homepage. Retrieved from https://www.imdb.com/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.

Jiwani, N., Gupta, K., & Whig, P. (2022). *Analysis of the potential impact of omicron crises using NLTK (natural language toolkit).* Paper presented at the Proceedings of Third Doctoral Symposium on Computational Intelligence: DoSCI 2022.

Joshy, A., & Sundar, S. (2022). *Analyzing the Performance of Sentiment Analysis using BERT, DistilBERT, and RoBERTa.* Paper presented at the 2022 IEEE International Power and Renewable Energy Conference (IPRECON).

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing, 2*(2010), 627-666.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415-463): Springer.

Musto, C., Semeraro, G., & Polignano, M. (2014). *A Comparison of Lexicon-based Approaches for Sentiment Analysis of Microblog Posts.* Paper presented at the DART@ AI* IA.

Neshan, S. A. S., & Akbari, R. (2020). *A combination of machine learning and lexicon based techniques for sentiment analysis.* Paper presented at the 2020 6th international conference on web research (ICWR).

Ranganathan, J., & Tsahai, T. (2022). *Sentiment Analysis of Tweets Using Deep Learning.* Paper presented at the International Conference on Advanced Data Mining and Applications.

Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information fusion, 36*, 10-25.

Team, T. (2020). Pandas development Pandas-dev/pandas: Pandas. *Zenodo, 21*, 1-9.

Trivedi, K. (2019). Multi-label text classification using bert–the mighty transformer. In: Medium. Retrieved from https://medium. com/huggingface/multi-label ….

Varathan, K. D., Giachanou, A., & Crestani, F. (2017). Comparative opinion mining: a review. *Journal of the Association for Information Science and Technology, 68*(4), 811-829.

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks, 5*(1), 7-16.

Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of information science, 18*(1), 45-55.