Jurnal Info Sains : Informatika dan Sains
Volume 14, Number 03, 2024, DOI 10.54209/infosains.v14i03
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Sentiment Analysis Using Transformers

Ahmad Fadhil N
Faculty of Computing, President University, Jababeka Education Park, Cikarang, Bekasi

Article Info
Keywords:
Sentiment Analysis,
IMDb,
BERT,
DistilBERT,
Transformers

ABSTRACT
This study examines how transformer-based models, such as BERT and DistilBERT, can be used for sentiment analysis of IMDb movie reviews. The goal of the experiment was to find a balance between accuracy and computational efficiency, evaluating how well both models performed with different training parameters. BERT was able to reach a peak accuracy of 91.39% in three epochs, taking a total of 54 minutes to train. On the other hand, DistilBERT achieved a similar accuracy of 91.80% in only 38 minutes and 25 seconds. Although there was a slight variance in accuracy, DistilBERT proved to be a much more efficient option for training, thus becoming a feasible substitute for environments with limited resources. The findings were contrasted against R. Talibzade's (2023) research, which obtained a 98% accuracy rate using BERT but needed 12 hours of training, illustrating the balance between accuracy and training duration. Potential upcoming tasks involve refining further, testing with bigger datasets, investigating alternative transformer models, and utilizing more resource-efficient training methods to improve performance without sacrificing efficiency.

Corresponding Author:
Ahmad Fadhil N
Faculty of Computing, President University, Jababeka Education Park, Cikarang, Bekasi
fadhil.naswir@president.ac.id

INTRODUCTION

In the current era of digitalization, technological advancements are progressing at an unprecedented rate, with artificial intelligence (AI) undergoing substantial development. Natural Language Processing (NLP), a critical subfield of AI, is designed to analyze sentiment, mood, or emotion within a given context, typically represented as text. NLP offers a novel perspective on traditional text classification methods. The task of classifying text based on emotional content remains a significant and complex area of research within NLP, with ongoing investigations in the domain of text mining (R. Talibzade, 2024).

One of the most common applications of sentiment analysis is identifying the sentiment expressed in reviews. The Internet Movie Database (IMDb) serves as a rich platform containing numerous movie reviews and ratings contributed by viewers, providing valuable insights into audience preferences and opinions on films (Singh, 2023). Understanding what viewers appreciate, criticize, and perceive in specific aspects of a movie is essential for enhancing film quality, shaping marketing strategies, and informing the development of new projects (Ramadhan, 2022).

Jurnal Info Sains : Informatika dan Sains
Volume 14, Number 03, 2024, DOI 10.54209/infosains.v14i03
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Transformers are an innovative neural network structure created for handling sequential data, with a strong focus on performing well in natural language assignments. In 2017, the Transformer model's self-attention mechanism changed how models deal with long-distance connections and contextual relationships among words in a sequence. This progress has resulted in top-notch results in many NLP tasks, such as language translation, text summarization, and sentiment analysis (Devlin, 2018). An example is the effective use of Transformers in conducting sentiment analysis on IMDB, with the aim of categorizing movie reviews as either positive or negative. Transformers surpass traditional models by understanding the full meaning of words in reviews, resulting in more precise sentiment forecasts.

Hugging Face is both a company and an open-source platform offering tools and libraries to assist in creating and deploying machine learning models, specifically in the field of natural language processing (NLP). Most notably recognized for its Transformers library, known for providing pre-trained models for a range of NLP tasks, such as text classification, translation, question answering, and more. The transformers models that are commonly used to do sentiment analysis are BERT and DistilBERT.

BERT is a bidirectional transformer pre-trained using a combination of masked language modeling objective and next sentence prediction on a large corpus comprising the Toronto Book Corpus and Wikipedia. Meanwhile, DistilBERT is a small, fast, cheap and light Transformer model trained by distilling BERT base. It has 40% less parameters than google-bert/bert-base-uncased, runs 60% faster while preserving over 95% of BERT's performances as measured on the GLUE language understanding benchmark. In this paper, by using several models from Hugging Face, namely BERT and DistilBERT as the model, it could identify which is the optimal model for sentiment analysis with high accuracy and minimal training time duration.

Related Work

(S. K. Singh, 2023) applied machine learning and deep learning algorithms in order to conduct sentiment analysis. The methods utilized include gathering data, preparing data, training the model, and comparing results. The dataset used in this research is IMDb dataset compiled by Andrew Maas consisting of 50.000 film reviews. The dataset is divided into training and testing sets with balanced labels. Various preprocessing steps are conducted including removal of HTML tags, special characters, and punctuation, converting text to lowercase, eliminating stopwords, and tokenization. The study incorporated machine learning algorithms such as Naïve Bayes, Logistic Regression, LSVM, Decision Tree, and deep learning algorithms like LSTM and BiLSTM. The top performance is achieved by LSTM and BiLSTM with an accuracy of 91%. Nevertheless, the research could be enhanced by incorporating and comparing various transformer models and algorithms.

(R. Talibzade ,2023) utilized transformers in the study conducted in 2023. The BERT model is employed for analyzing sentiments. The dataset used is 50.000 movie reviews provided by Andrew Maas. This study evaluates conventional machine learning methods, such as Multinomial Naive Bayes (MNB), Logistic Regression, Support Vector Machine (SVM), against a transformer model called BERT. The findings show that the BERT model performs better than all other machine learning algorithms, achieving a 98% accuracy rate. Despite

Jurnal Info Sains : Informatika dan Sains
Volume 14, Number 03, 2024, DOI 10.54209/infosains.v14i03
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

achieving very good results, the training process took a considerable amount of time, specifically 12 hours.

(S. Ouyang, 2023) also used the IMDB dataset for sentiment analysis. In this research, there is only one preprocessing done namely Tokenization, N-gram is used as the feature extraction for enhancing the model's understanding of the capture contextual information. Model architecture is designed with a sequence of layers namely embedding layer and global average pooling 1D layer, two fully connected layers, and the last layer is dense layer for the binary classification for the prediction of positive or negative sentiment. The result of this research is 95% accuracy. However there are other several preprocessing methods that can be implemented and there was no comparison between this research with other research.

(G. Cahyani et al., 2022) and (N. G. Ramadhan, 2022) employed a conventional method utilizing Support Vector Machines (SVM). In their study, (G. Cahyani et al., 2022) utilized a dataset of 2,000 IMDb movie reviews, split into 80% for training and 20% for testing, while (N. G. Ramadhan, 2022) specifically concentrated on reviews of the Squid Game series. Both studies utilized a range of preprocessing techniques such as tokenization, filtering, excluding stopwords, and stemming. (G. Cahyani et al., 2022) reached an accuracy of 86.5%, while N. (N. G. Ramadhan, 2022) achieved a lower accuracy of 79%. In contrast to other studies, these two research projects showed somewhat lower outcomes. Expanding the size of the dataset and comparing the results with cutting-edge algorithms could boost the performance even more than the already impressive results obtained.

(S. M. Qaisar 2020) utilized deep learning algorithms, specifically LSTM, as the model for sentiment analysis. The dataset employed consists of 25000 entries from IMDb gathered by Andrew Maas. Eliminating symbols and punctuation, changing to lowercase, removing links, excluding stop words, and stemming are the preprocessing stages in this research. Vectorization is also carried out for feature extraction before training the model. The research showed an accuracy of 89%, which is comparatively lower than similar studies utilizing deep learning as the primary algorithm. Furthermore, there is a lack of comparison of the findings with other research using deep learning for sentiment analysis.

METHODS

In this research, a transformers model is proposed to do the sentiment analysis. The pre-trained models used are BERT and DistilBert. The accuracy and training time will be compared as the result of this experiment. The methodology for this experiment is as follows:

1. Data Collection
   The dataset consists of movie reviews from IMDb collected by Andrew Maas and is commonly utilized for sentiment analysis studies. In this study, 50,000 reviews were used, with half being positive and half negative. Every review is tagged appropriately, creating a well-rounded dataset that acts as a standard for assessing sentiment analysis model effectiveness. The dataset contains a wide variety of viewpoints, making it a strong tool for evaluating the efficiency of different methods in sentiment classification.
2. Data Preparation
   During data preparation, string labels were converted to numeric labels to help train the models. This transformation was needed to convert the sentiment labels, like "positive"

Jurnal Info Sains : Informatika dan Sains
Volume 14, Number 03, 2024, DOI 10.54209/infosains.v14i03
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

and "negative," into numerical values for the algorithm. Furthermore, the dataset columns were updated by changing the name of the text column to "text" to make it compatible with the HuggingFace model framework. An initial examination was done with a sample of 100 data points to determine if the model could be effectively trained. This stage functioned as a first test to confirm the viability of the comprehensive training process and to recognize any potential problems in the data pipeline.

3. Preprocessing
   During the data preprocessing stage, various important actions were taken to get the dataset ready for analysis. Initially, common non-informative words were eliminated by removing stopwords to enhance the sentiment analysis by focusing on more meaningful content. After that, all text was transformed to lowercase in order to maintain consistency and avoid inconsistencies due to variations in capitalization. Special characters and digits were eliminated to clean the text and concentrate only on significant word content. The converted data was then transformed into the Hugging Face dataset format, necessary to work with the Hugging Face framework and its models. In the end, the text was tokenized to separate it into individual tokens or words, necessary for the models. Every single one of these preprocessing stages played a vital role in standardizing and enhancing the data, improving the efficiency and precision of the sentiment analysis model.
4. Training and Testing Data
   The dataset was partitioned into training and test sets to assess the sentiment analysis model's performance during the data split. In accordance with the Pareto principle, which states that around 80% of results stem from 20% of causes, the data was separated to mirror this pattern. More precisely, 80% of the data was assigned to the training set, offering a significant amount of information for model training and enabling the algorithm to effectively learn from a sample that represents the dataset. The test set, comprising 20% of the data, was set aside to evaluate how well the model performed and generalized to new data. This division guarantees that the model is assessed on a distinct subset, providing a trustworthy evaluation of its precision and stability while following recommended methods for training and validation.
5. Define Training Argument
   Two models were employed in the training process: BERT and DistilBERT. The training arguments for both models were configured as follows:
   a. Batch size during the training phase is set to 16, chosen to strike a balance between memory usage and computational efficiency. Reducing the batch size can improve the stability of training, particularly with complex models and extensive datasets.
   b. Setting the evaluation batch size to 64 leads to more efficient processing of validation data, enabling faster assessment of model performance.
   c. Number of training epochs set to 3 after preliminary experiments and model convergence considerations. It makes sure that the models can learn from the training data adequately without falling into overfitting.
   d. The primary metric selected for evaluating model performance was Accuracy. This evaluation criteria assists in assessing the model's ability to accurately categorize

Jurnal Info Sains : Informatika dan Sains
Volume 14, Number 03, 2024, DOI 10.54209/infosains.v14i03
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

text as either positive or negative sentiments, which is crucial to the goal of sentiment analysis.

e. Utilizing FP16 Precision (with fp16 = true enabled) employs mixed-precision training to speed up training on GPUs. Through the utilization of 16-bit floating-point numbers, the model experiences decreased memory utilization and quicker calculation speeds, all while preserving accuracy and stability.

These training parameters were fine-tuned to improve the performance of the models, guaranteeing strong results on sentiment analysis.

6. Train the model

In the training phase, both BERT and DistilBERT were trained with specific arguments to enhance their performance on sentiment analysis. The process of training included sending the data in batches through the models, using a train batch size of 16 to efficiently handle memory and computational tasks. The models were trained for 3 epochs, giving enough time for the algorithms to grasp patterns and enhance accuracy. Assessment was conducted using a bigger batch size of 64, allowing for effective processing and evaluation of model performance on validation data. The accuracy measurement was utilized to assess the efficiency of the models, aiming to pinpoint the top-performing model. In addition, mixed-precision training (fp16) was used to speed up the training on GPUs, decreasing training time while keeping model accuracy. This method guaranteed that both models were effectively and efficiently trained, establishing a strong base for assessing their performance and generalization abilities.

7. Evaluation

In the evaluation phase, the performance of the trained BERT and DistilBERT models was assessed to determine their effectiveness in sentiment analysis. The primary evaluation metric used was accuracy, which measures the proportion of correctly classified sentiments in the test set. This metric provides a clear indication of how well each model performs in distinguishing between positive and negative reviews.

The results of this evaluation will provide insights into the practical trade-offs between model performance and computational efficiency, guiding future decisions on model selection based on specific requirements and resource constraints.

RESULTS AND DISCUSSION

The experimental process included a methodical way of training and assessing the BERT and DistilBERT models. In the beginning, the data set underwent preprocessing, which involved actions like removing stopwords, normalizing text, and tokenizing to get the data ready for analysis. Next, the models were set up with particular hyperparameters: a training batch size of 16, an evaluation batch size of 64, and a training period of 3 epochs. Mixed-precision training (fp16) was utilized to improve computational efficiency on GPUs. Every model underwent training on the training data and was then assessed using the test data to measure their effectiveness. The accuracy measure was utilized to assess how well each model categorized sentiments. The outcomes were documented, with the comparison centered on accuracy and training time between BERT and DistilBERT. This process guaranteed a

Jurnal Info Sains : Informatika dan Sains
Volume 14, Number 03, 2024, DOI 10.54209/infosains.v14i03
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

comprehensive and uniform assessment of the models, offering understanding of their unique advantages and drawbacks.

Preliminary Analysis

In the initial stage of the study, a portion of 100 data samples was utilized to perform an initial assessment of the sentiment analysis models. This small dataset was chosen to promptly assess the models' efficacy and pinpoint any problems before moving on to the complete dataset. The aim of this initial analysis was to confirm the models' ability to be trained effectively and to establish a basic understanding of their performance in terms of accuracy and loss. The results of the initial training include metrics for training loss, validation loss, and accuracy over three epochs. These measurements provide understanding of the models' learning and generalizing abilities from a limited sample, guiding future actions in the experiment. Below is the result of the preliminary analysis:

Table 1. Preliminary Analysis

Epoch	Accuracy
1	0.909500
2	0.909900
3	0.911300

This preliminary experiment required 38 minutes and 25 seconds in total for training. This time period encompasses the entire training process throughout all three epochs for the model. Given that only 100 samples were used for this training, the short duration of the training emphasizes the effectiveness of the process. Employing mixed-precision (fp16) training on GPUs probably played a role in speeding up the process, enabling the model to handle data more quickly while maintaining performance. Efficiency is important to scale up the experiment to the full dataset, as the computational needs will be much greater.

Experiment Result

The result of the experiment using BERT and DistilBERT model are shown in table below:

Table 2. BERT Model Result

Epoch	Accuracy
1	0.891800
2	0.913100
3	0.913900

The BERT model's training duration was 54 minutes and 0 seconds. This length of time represents the duration needed to finish three training cycles with the set parameters, such as batch size and mixed-precision (fp16) optimization. Due to the intricate nature and large scale of the BERT model, this training duration is justifiable and showcases the model's need for computational power. While BERT requires more time for training than smaller models such as DistilBERT, its impressive accuracy makes the extended training time worthwhile. This level of accuracy and ability to generalize in natural language processing tasks is achieved by making a significant time investment in BERT.

Jurnal Info Sains : Informatika dan Sains
Volume 14, Number 03, 2024, DOI 10.54209/infosains.v14i03
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Table 3. DistilBERT Model Result

Epoch	Accuracy
1	0.898900
2	0.917900
3	0.918000

The DistilBERT model's training duration was 39 minutes and 15 seconds in total. The shorter training time of DistilBERT in comparison to BERT emphasizes its effectiveness. Even with less training time, DistilBERT managed to reach similar accuracy levels, highlighting its enhancement in computational efficiency without compromising its strong performance. This is why DistilBERT is especially attractive in situations where quicker training is important, like when dealing with big datasets or restricted computational resources, without sacrificing accuracy much.

According to the experiment findings, BERT and DistilBERT performed well in sentiment analysis by accurately classifying movie reviews. Although BERT obtained a slightly higher accuracy of 91.39% after three epochs, it took a total of 54 minutes for training. On the other hand, DistilBERT achieved a similar accuracy of 91.80% in just 39 minutes and 15 seconds, demonstrating a much faster training time.

CONCLUSION

The results suggest that while BERT offers marginally better accuracy, DistilBERT provides a more efficient alternative, delivering nearly identical performance in a fraction of the training time. This makes DistilBERT a practical choice for scenarios where computational efficiency and speed are crucial, without sacrificing much in terms of accuracy. Overall, the choice between BERT and DistilBERT may depend on the specific requirements of the task, such as the need for maximum accuracy versus the need for faster training and lower resource usage. Significant disparities in accuracy and training time are evident when comparing this experiment's results to R. Talibzade (2023). Talibzade's test with a BERT model based on transformers yielded an outstanding accuracy rate of 98%. Nevertheless, achieving this level of performance took a significant toll on computational time, necessitating 12 hours of training. This prolonged training period is probably due to the utilization of bigger datasets, more intricate fine-tuning, and perhaps increased model capacity in order to reach optimal accuracy. On the other hand, the latest test resulted in an accuracy of up to 91.80% using DistilBERT in a mere 39 minutes and 15 seconds, compared to 91.39% with BERT in 54 minutes. Despite being less accurate than Talibzade's, the models with significantly shorter training times are much more suitable for situations requiring rapid model creation and implementation. The experiment shows that Both BERT and DistilBERT, especially DistilBERT, are efficient for tasks that require a balance between accuracy and training time.

FUTURE WORK

After analyzing the experiment and its outcomes, there are multiple potential areas for future research that could improve the effectiveness and practicality of sentiment analysis models.

Future research can concentrate on further fine-tuning BERT and DistilBERT, possibly using methods like learning rate scheduling, early stopping, or more extensive hyperparameter tuning. This may aid in closing the accuracy difference seen in models such as the one employed in R. Talibzade's (2023) research, while also keeping training times efficient. Although this study focused on BERT and DistilBERT, upcoming research could evaluate how they perform in comparison to other transformer-based models like RoBERTa, various GPT models, and ALBERT. These models might provide varying balances between accuracy, speed, and computational efficiency.

REFERENCE

T. Ahmed Khan, R. Sadiq, Z. Shahid, M. M. Alam, and M. Mohd Su'ud, "Sentiment Analysis using Support Vector Machine and Random Forest," *J. Informatics Web Eng.*, 2024, doi: 10.33093/jiwe.2024.3.1.5.

S. K. Singh and N. Singla, "Sentiment Analysis on IMDB Review Dataset," *J. Comput. Mech. Manag.*, 2023, doi: 10.57159/gadl.jcmm.2.6.230108.

R. Talibzade. "Sentiment Analysis of IMDb Movie Reviews Using Traditional Machine Learning Techniques and Transformers." 2024. DOI: 10.13140/RG.2.2.29464.16644.

S. Ouyang, "Deep learning for sentiment analysis on IMDB movie reviews using N-gram features," Appl. Comput. Eng., 2024, doi: 10.54254/2755-2721/35/20230361.

G. Cahyani, W. Widayani, S. D. Anggita, Y. Pristyanto, I. Ikmah, and A. Sidauruk, "Klasifikasi Data Review IMDb Berdasarkan Analisis Sentimen Menggunakan Algoritma Support Vector Machine," *J. MEDIA Inform. BUDIDARMA*, 2022, doi: 10.30865/mib.v6i3.4023.

N. G. Ramadhan and T. I. Ramadhan, "Analysis Sentiment Based on IMDB Aspects from Movie Reviews using SVM," *Sinkron*, 2022, doi: 10.33395/sinkron.v7i1.11204.

S. M. Qaisar, "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory," 2020, doi: 10.1109/ICCIS49240.2020.9257657.

Amulya, K. et al. "Sentiment Analysis on IMDB Movie Reviews Using Machine Learning and Deep Learning Algorithms." 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT). Piscataway: IEEE, 2022. 814–819. Web.

M. R. Haque, S. Akter Lima and S. Z. Mishu, "Performance Analysis of Different Neural Networks for Sentiment Analysis on IMDb Movie Reviews," 2019 3rd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE), Rajshahi, Bangladesh, 2019, pp. 161-164, doi: 10.1109/ICECTE48615.2019.9303573.

M. Yasen and S. Tedmori, "Movies Reviews Sentiment Analysis and Classification," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 2019, pp. 860-865, doi: 10.1109/JEEIT.2019.8717422.

IMDB Dataset of 50K Movie Reviews. (n.d.). IMDB Dataset of 50K Movie Reviews | Kaggle. https:///datasets/lakshmi25npathi/imdb-dataset-of-50k-movie reviews

G. Nkhata, "Movie Reviews Sentiment Analysis using BERT." Order No. 29997252, University of Arkansas, United States -- Arkansas, 2022.

Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Jurnal Info Sains : Informatika dan Sains
Volume 14, Number 03, 2024, DOI 10.54209/infosains.v14i03
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

T. P. Sahu and S. Ahuja, "Sentiment analysis of movie reviews: A study on feature selection & classification algorithms," 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), Durgapur, India, 2016, pp. 1-6, doi: 10.1109/MicroCom.2016.7522583.

Baid, Palak & Gupta, Apoorva & Chaplot, Neelam. (2017). Sentiment Analysis of Movie Reviews using Machine Learning Techniques. International Journal of Computer Applications. 179. 45-49. 10.5120/ijca2017916005.