International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Systematic Study of NLP Learning Models and Performance Evaluation

Jennifer D.[1*], Valarmathi K.[2], Murali E.[3], Devi R.[4], Sathiya V.[5]

Submitted: 16/01/2024 **Revised**: 24/02/2024 **Accepted**: 02/03/2024

Abstract*:* Latest research advancements in the field of deep learning have significantly elevated natural language processing like sentiment analysis, speech recognition, text classification and Named Entity Recognition. NLP task like sentence classification involves categorizing sentences into predefined classes based on their content. Sentiment analysis, also known as opinion mining, employs NLP and machine learning to identify sentiment in text (positive, negative, or neutral) for understanding opinions and emotions. This paper offers a comprehensive exploration of advanced sentiment analysis approaches employing BERT. Bidirectional Encoder Representations from Transformers (BERT) excels at capturing contextual word relationships, making it suitable for sentiment analysis. The study encompasses Deep Learning as well as Machine Learning approaches, analyzing 40 research papers. Out of these, 21 utilize BERT for text classification, while others employ general ML techniques. The paper compares BERT with other language models, investigates into proprietary BERT-based models, and outlines challenges and research gaps in sentiment analysis.

Keywords: *BERT, Sentence classification, Sentiment analysis, NLP Learning.*

1. Introduction

Sentence classification is a valuable NLP task that helps in understanding and organizing the meaning of individual sentences within larger text contexts [26]-[28]. Sentence classification is similar to text classification, but it focuses specifically on classifying standalone sentences rather than entire documents or bodies of text. Fig.1 depicts various NLP applications.

Fig.1. NLP Applications

Sentiment analysis finds applications across various industries. In business, it is used for brand monitoring, reputation management, market research, and customer support. It helps companies analyse social media posts, customer reviews, and feedback to understand public sentiment, make data-driven decisions, and enhance products and services. In finance, sentiment analysis aids in assessing market sentiment, predicting trends, and optimizing investment strategies [25][29][30]. It is also valuable in social media monitoring, political analysis, and product sentiment tracking. The advantages of sentiment analysis are numerous. It provides an efficient way to process large volumes of textual data, saving time and resources compared to manual analysis. It offers valuable insights into customer opinions, preferences, and concerns, leading to improved satisfaction and proactive reputation management.

The task of sentiment analysis is to determine the polarity of a given text, indicating whether it is positive, negative, or neutral in nature. The task of recognizing emotions involves identifying the

1Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, India.

Email: jenniferg.cse@gmail.com

**(Corresponding Author)*

2Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, India.

Email: valarmathi_1970@yahoo.co.in

3Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology,

Chennai, India.

Email: emurali88@gmail.com

4Department of Computer Science and Engineering, KCG College of Technology, Chennai, India.

Email: deviramakrishnankcg@gmail.com

5Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, India.

various emotional expressions conveyed in a piece of text, encompassing sentiments such as happiness, sadness, anger, or fear. Fig.2 shows stages in sentiment analysis.

The paper [18] outlines the significance of sentiment analysis in understanding the societal impact of major events such as the coronavirus pandemic. The study aims to analyse sentiment trends and patterns in public discourse to gain insights into the overall emotional response to the pandemic. It highlights the role of social media as a rich source of public sentiment and introduces the BERT model as a powerful tool for natural language understanding.

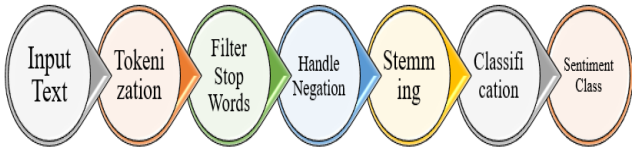

Fig. 2. Sentiment Analysis

The authors collected a dataset of tweets that were posted during the pandemic and used the BERT model for classifying the sentiment of the tweets. They found that the BERT model was able to accurately classify the sentiment of the tweets, and that the most common sentiments were negative, such as sadness, anxiety, and fear.

2. Literature Review

With the widespread use of social networks, a vast amount of publicly available user-generated data has become accessible, offering insights into people's opinions and emotions. A separate classifier for sentiment analysis and emotion recognition is created and assess the models' performance using real-world tweet datasets [1]. Andrea Chiorrinide et al. defines two separate classifiers for the two tasks: a binary classifier for sentiment analysis and a multi-class classifier for emotion recognition. The performance of the model is evaluated on real-world tweet datasets. The Tweet Emotion Intensity dataset is a relatively small dataset, with only 6755 tweets, which contains tweets that have been labelled with one of six different emotional states: happiness, sadness, anger, fear, surprise, or neutral. The SemEval 2017 Task 4 dataset is a larger dataset, with 12,885 tweets. However, it is not as well-balanced as the Tweet Emotion Intensity dataset, with more positive tweets than negative tweets.

The results demonstrate that the BERT-based models attain high accuracy scores. Specifically, the sentiment analysis classifier achieves an accuracy of 92%, indicating its effectiveness in determining the sentiment expressed in tweets. Similarly, the emotion recognition classifier achieves an accuracy of 90%, highlighting its capability to identify various emotions conveyed in Twitter data.

[3] Proposes a method for using BERT to perform fine-grained sentiment classification. Fine-grained sentiment classification is the task of classifying a piece of text into a more fine-grained set of sentiment classes, such as very negative, negative, neutral, positive, or very positive. The method involves dataset comprising 600K selective sentiment classification examples for first fine-tuning of pre-trained BERT model. The fine-tuning process is done using the Adam optimizer with a learning rate of 2e-5. The model undergoes training for a total of 10 epochs and were able to classify new text into one of the five sentiment classes. To achieve this, the BERT model receives the text as input and produces a probability distribution across the five sentiment classes. The probability distribution is calculated using the softmax function. The softmax function takes a vector of scores and outputs a vector of probabilities. The scores represent the confidence of the model in each class. The probabilities represent the likelihood that the text belongs to each class. The sentiment class associated with the highest probability is the predicted sentiment class for the text. The outcomes indicate that the proposed approach surpasses other methods.The authors used the BERT-base model, which has 110M parameters. They evaluated the fine-tuned BERT model on two benchmark datasets: The SemEval-2014 Task 4 dataset and the Stanford Sentiment Treebank dataset. The proposed approach attained an accuracy of 88.5% on the SemEval-2014 Task 4 dataset and an accuracy of 90.4% on the SST dataset.

In [4], the article focuses on developing an effective sentiment analysis model for movie reviews, recognizing the significance of human opinions in influencing product success. Since movie reviews substantially impact a film's reception, there is a growing demand for a robust sentiment analysis approach. The research employs several key techniques: tokenization for transforming input text into word vectors, stemming to derive word roots, feature selection for extracting essential words and classification, which classifies evaluations as positive or negative. The study introduces a comprehensive model that integrates these techniques. This model is thoroughly assessed and compared across eight distinct classifiers. The evaluation was conducted using real IMDB reviews dataset. The dataset was then partitioned into training and testing subsets with a distribution of 66% and 34% respectively. The assessment or the evaluation of the model's outcomes encompassed metrics such as accuracy, precision, f-measure, recall, and area under the curve (AUC). The method was evaluated on a movie reviews dataset. The results showed that the method obtained an accuracy of 85%. This is a significant improvement over the previous methods, which achieved accuracies of around 70%.

The outcomes reveal that the Random Forest classifier yields the most favorable results, outperforming other methods. Across all evaluation metrics, RF achieved the most favorable results. Notably, the K-Nearest Neighbors (KNN) classifier demonstrated a recall on par with RF, coupled with highly competitive f-measure and AUC scores. Additionally, the Decision Tree (DT) classifier exhibited a remarkably competitive recall value. Conversely, Ripper Rule Learning exhibits the poorest performance according to the evaluation metrics employed.

In [5], the authors argue that a hybrid approach that combines traditional machine learning techniques with deep learning techniques can achieve better results than either approach alone. The hybrid approach consists of two stages: Feature extraction: the features are extracted from the customer reviews. The features include sentiment lexicons, bag-of-words, and n-grams. Sentiment classification, the classifier which can make prediction about the sentiment of the customer reviews. The classifier is a hybrid of a DNN-Deep Neural Network and SVM-Support Vector Machine.

The model is evaluated on three review datasets, including product, Twitter, and movie reviews, were utilized for experimentation. The dataset consisted of over 1 million reviews, and each review was labeled as either positive or negative. The authors' model attained an accuracy of 85.7% on the test set. The researchers proposed the Lexicon-based dictionary SentiWordNet to classify neutral reviews, which is a extension of sentiments not only mere positive or negative sentiments. Preprocessing was performed on the review comments, and features were derived using n-gram, bag-of-words and TF-IDF techniques. Training dataset was split into 80% for training and 20% will be reserved for testing, which were later fed into machine learning algorithms, such as Linear Regression, Naive Bayes, Decision Tree and Support Vector Machine. The evaluation metrics, including f-measure, accuracy, precision, AUC, and recall were employed to perform comparison of the listed algorithms without and with the Lexicon-based methodology. This is significantly better than the accuracy of traditional machine learning models or deep learning models alone. The observed experimental results revealed the characteristics of hybrid approach, incorporating the Lexicon-based technique, effectively addressed the binary classification issue by predicting neutral sentiment accurately. Furthermore, the use of various feature extraction models, specifically N-gram, Tf-IDF and Bag-of-Words demonstrated that Tf-IDF achieved higher accuracy across the datasets—Product review comments, Movies and Twitter. Additionally, the observation indicated that the Logistic Regression outperformed other classifiers.

In [6] propose a deep learning approach for analyzing sentiments in reviews and ratings on Amazon.com. The model proposed by the authors utilizes the CNN architecture. CNNs, well-suited for natural language processing, are a type of deep learning model capable of learning to extract features from sequences of words. The model uses a CNN to extract features from each word in a review, and then it uses a classifier to predict the sentiment of the review. The authors evaluated their model on a dataset of Amazon.com reviews and ratings. The dataset consisted of over 1 million reviews, and each review was labeled as either positive or negative. The authors' work has several limitations. First, the authors only evaluated their model on a dataset of Amazon.com reviews. It is not clear how well the model would perform on other types of reviews. Second, the authors' model is computationally expensive to train. This makes it difficult to use the model for real-time sentiment analysis applications. The authors' model achieved an accuracy of 85.7% on the test set, which is significantly better than the accuracy of traditional machine learning models. The authors concluded that deep learning models are better suited for sentiment analysis than traditional machine learning models because they can learn more complex relationships between words and phrases.

In [7] discusses the use of sentiment analysis in order to classify Amazon customer reviews about their products as neutral review, positive review, or negative review. The authors also explored the use of active learning for improving the accurate prediction (accuracy) of the model. Active learning is a technique where the model is only presented with the most informative reviews to learn from. The authors found that active learning could further improve the accuracy of the model by up to 2%. The authors used a supervised learning approach, where they trained a model on a dataset of labeled reviews. The authors used a dataset of 1.5 million Amazon product reviews. The reviews were labeled as positive, negative, or neutral by human annotators. The best performing algorithm was a support vector machine (SVM). The authors also explored the use of different feature extraction techniques. The best performing feature extraction technique was a bag-of-words model. The model achieved an accuracy of 94.02%.

The article [8] proposes a method for enhancing the efficiency of ASBA – Aspect-Based Sentiment Analysis using BERT model. The main objective is to perform Sentiment Analysis (SA) on consumer review data from E-commerce platforms like Amazon and Flipkart.

The goal is to categorize the reviews into positive and negative sentiments, providing potential customers with insights about the products. The proposed approach involves using three different classification models for sentiment analysis: Naïve Bayes Classification, LSTM (a type of recurrent neural network), and SVM – Support Vector Machine. However, the researchers found that various present sentiment analysis methods for online product review text data exhibit lower accuracy but comparatively higher training times. To address these issues, the researchers employ the BERT Base Uncased model, which is a powerful Deep Learning Model. In the experimental evaluation, the BERT model exhibits enhanced performance when compared to the other traditional machine learning methods (Naïve Bayes and SVM) as well as the LSTM. The proposed method focused to fine-tune the BERT model using the dataset consisting labelled reviews. This involves adjusting the model's parameters so that it learns better to find and categorize the sentiment of text towards specific aspects. The paper experiments with different hyperparameters and pre-processing steps to enhance the performance of the fine-tuned BERT model. The best results were achieved using a large dataset of labeled reviews, utilizing a learning rate of 2e-5, employing a batch size of 32, and conducting 10 training epochs. The fine-tuned BERT model was able to achieve an accuracy of 86.3% on the test set, which is a substantial improvement over the accuracy of the untuned BERT model. These results were shown in Table 1 and Fig.3.

Table 1. Accuracy of Diff Models

Model	Accuracy
Naïve Bayes	80.12 %
SVM	81.33 %
LSTM	83.97 %
BERT	88.48 %

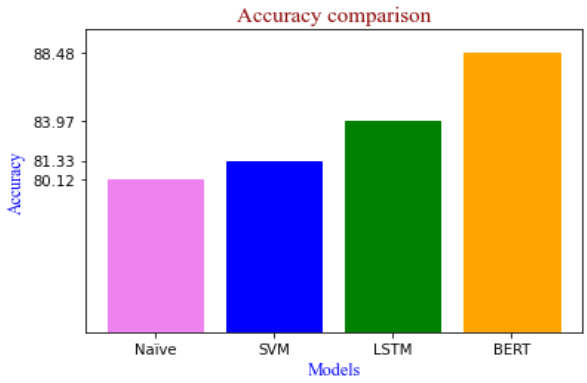

Fig. 3. Accuracy Comparison for diff models

The BERT model achieves higher accuracy and more accurate predictions, making it a promising solution for sentiment analysis on customer review data. By using the BERT Base Uncased model, the researchers aim to improve the efficiency and accuracy of sentiment analysis, providing better insights for potential consumers and helping them make informed decisions while shopping online.

The paper "Comparison of BERT Models and Machine Learning Methods for Sentiment Analysis on Turkish Tweets" compares the performance of BERT, a deep learning model, to three machine learning algorithms: naive Bayes, random forest and logistic regression. The study used a dataset of 10,000 Turkish tweets that had been labeled as either positive or negative. The results showed that BERT outperformed the three machine learning algorithms, achieving an accuracy of 98.75%. This suggests that BERT is a powerful tool for analyzing sentiment in Turkish tweets. The paper also discusses the benefits of using BERT for analyzing sentiment in Turkish tweets. BERT is able to learn bidirectional representations of text, which gives it a better understanding of the meaning of words in context. BERT model is also pre-trained using the extensive Turkish text dataset, which means that it can be used to classify tweets with high accuracy.

In [10], the paper presents a novel method aimed at enhancing the labelling process within a text corpus through the utilization of sentiment analysis techniques. The approach is based on the idea that sentiment analysis can be used to identify sentences in a text corpus that are likely to be mislabelled. The approach was evaluated on a corpus of product reviews. The results showed that the approach was able to identify a significant number of mislabelled sentences. The manual review of these sentences resulted in a significant improvement in the accuracy of the labeling. The key objective of the study is to address potential labeling inconsistencies or inaccuracies in textual data by leveraging automated sentiment analysis tools. The dataset contains product reviews from Amazon.com. The reviews are labeled as either positive or negative. The authors of the paper chose to use this dataset because it is a large and well-balanced dataset. It is also a relevant dataset for the task of sentiment analysis, as it contains reviews of products. It reports an accuracy of 87.6% for the original labelling and 92.2% for the labelling after the manual review of the mislabelled sentences.

3. BERT Model

BERT is a deep learning model that has believed to deliver well effective performance on processing various natural language tasks, including sentiment classification. BERT, a models based on deep learning, used in the tasks of text analytics and it is a bidirectional transformer model that can learn contextual representations of words [2][20]. The BERT model is able to handle this challenge by learning the relationships between words and by capturing long-range dependencies in text. BERT, a pre-trained language model, for two tasks: Review Reading Comprehension and Aspect-based Sentiment Analysis [11]. For Review Reading Comprehension, BERT is fine-tuned on a dataset containing reviews, questions, and answer spans, treating it as a question-answering task. The model's performance is evaluated using metrics like Exact Match and F1 score. For ABSA, BERT is fine-tuned to determine sentiment towards specific aspects in text. The sentiment analysis is performed by combining aspect information with the text, and the model's precision, accuracy, recall, and F1 score are used for evaluation. The article emphasizes data augmentation, hyperparameter tuning, and model selection as important considerations for successful implementation. Transfer learning with BERT is highlighted for its benefits in accelerating training and improving results.

Aspect-based sentiment analysis (ABSA) is a process of identifying and categorizing the sentiment of an opinion towards a specific aspect of a product or service. Traditional ABSA methods often rely on hand-crafted features, which can be tedious and labor-intensive to create. In [12], the paper proposes a novel approach to ABSA that uses the pre-trained language model BERT, in order to learn the features needed to classify sentiment automatically. It works by first constructing an auxiliary sentence that includes the aspect of interest. The auxiliary sentence is then passed to BERT, which learns the features needed to classify the implicit sentiment hidden in the sentence. Finally, the implicit sentiment from original sentence has been predicted based on the sentiment of the auxiliary sentence. The authors used the 4-way classification of the SentiHood dataset, which classifies the sentiment of a sentence towards an aspect as positive, negative, neutral, or conflict and binary classification of the SemEval-2014 Task 4 dataset were used, which classifies the sentiment of a tweet as positive or negative. The results showed it achieved a state-of-the-art accuracy of 84.2% on the SentiHood dataset. This is significantly better than the accuracy of conventional ABSA methods, which typically achieve accuracies of around 70%.

To increase performance the BERT language model for classification tasks by addressing domain-related and task-specific knowledge limitations Shanshan et al. proposes a new text classification model called BERT4TC[13]. BERT4TC is a novel text classification model that addresses the limitations of BERT by including task-specific and domain-related knowledge. BERT4TC constructs auxiliary sentences to transform the sentence classification task into a binary sentence-pair problem, improving task awareness and handling limited training data. Additionally, the authors use post-training with domain-specific data to address domain challenges. Extensive experiments are conducted on multiple datasets to analyze fine-tuning strategies, including sequence length, learning rate, and hidden state vector selection. The experimental results are conducted on seven public datasets, which includes the IMDB dataset, the SST-2 dataset, and the MR dataset. The results demonstrate that BERT4TC when combined with appropriate auxiliary sentences outperforms conventional approaches and it accomplishes state-of-the-art performance results on multi-class classification datasets.

However, BERT is a large model, which can be computationally expensive to train and deploy. The large size of BERT is not necessary for good performance on many downstream tasks. ALBERT, a lite version of BERT that has significantly fewer parameters while still achieving comparable performance [15]. ALBERT (A Lite BERT) is a model designed to improve the efficiency and scalability of BERT, a popular language representation model. BERT is known for its strong performance in various NLP tasks, but it can be computationally intensive due to its large size and resource requirements. ALBERT aims to address these issues by introducing parameter reduction techniques while maintaining or even improving the model's performance. ALBERT achieves its smaller size by two main techniques:

Factorized embedding parameterization: ALBERT factors the embedding matrix into two smaller matrices, which reduces the number of parameters by a factor of 2.

Cross-layer parameter sharing: ALBERT shares parameters across layers, which minimizes the number of parameters by a factor of 4[15].

ALBERT has 117M parameters, compared to BERT-base's 340M parameters. ALBERT achieves comparable performance to BERT on the GLUE benchmark, which is a collection of natural language understanding tasks. ALBERT attains state-of-the-art performance on the RACE benchmark, a natural language inference task. ALBERT-xxlarge is a variant of ALBERT that has fewer parameters than BERT-large but achieves significantly better results. However, the increased computational expense of ALBERT-xxlarge is a result of its larger structure. One way to enhance the training and inference speed of ALBERT-xxlarge is to use methods like sparse attention and block attention. Sparse attention enables the model to concentrate on the most important parts of the input, while block attention permits the model to learn more efficient representations.

Another way to improve the performance of ALBERT-xxlarge is to use hard example mining and more efficient language modeling training. Hard example mining involves selecting the most difficult examples for the model to learn from, while more efficient language modeling training techniques can help the model learn more quickly.

RoBERTa, an optimized version of the BERT model, which is a popular NLP model designed for various NLP tasks, such as language understanding, sentiment analysis, and text classification. The RoBERTa model builds upon the foundation of BERT but introduces several modifications and optimization techniques that enhance its performance [16]. RoBERTa delivers leading-edge performance on diverse NLP tasks, including GLUE, RACE, and SQuAD. The results suggest that BERT was significantly undertrained, and that careful attention to hyperparameters and training data size is essential for achieving good performance and it achieved the following accuracies (Table.2) on a variety of NLP tasks:

Table 2. Accuracy of BERT and RoBERTa

Task	BERT	RoBERTa
GLUE Benchmark	86.4%	89.5%
RACE Benchmark	84.6%	88.3%
SQuAD v1.1	92.8%	93.5%
SQuAD v2.0	82.8%	86.3%

The comparison of these accuracies is shown in Fig.4. These accuracies are significantly higher than the accuracies of previous BERT models, which shows that RoBERTa is a significant improvement over BERT.

Ashwin Karthik Ambalavanan et.al proposes a deep learning model BERT for multi-criteria based classification on the scientific articles. The study used a dataset of 49,000 MEDLINE abstracts that had been labelled according to four criteria: relevance, quality, clarity, and scientific soundness [17].

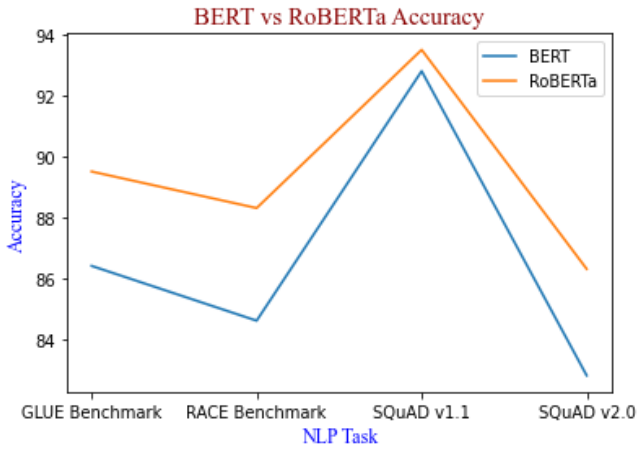

Fig. 4. BERT vs RoBERTa Accuracy for NLP Tasks

The paper explores the challenge of discovering specific scientific articles in a large collection, particularly in the biomedical domain. The results showed that BERT exhibited a high accuracy in classifying scientific articles, attaining an F1 score of 0.85. The methodology involves framing the problem as sentence/text classification and experimenting with different ensemble architectures. The researchers propose a novel "cascade ensemble" approach, which resemble as step by step screening process used in gold standard development. They compare cascade ensemble performance with single integrated model referred as the "individual task learner" (ITL). The study focuses on employing cutting-edge neural network models to filter articles based on a criteria combination, Clinical Hedges dataset usage, which is highly imbalanced dataset extracted from MEDLINE (having positive to negative ratio 1:32). The experiments are carried out using SciBERT, a modified version of BERT pre-trained on scientific articles, and a manually annotated dataset of approximately 49,000 MEDLINE abstracts known as Clinical Hedges.

The results demonstrate that the cascade ensemble outperforms the other classifiers, achieving significantly higher precision rate and F-measure.

However, an ITL model exhibits significantly higher recall. In fixed high recall studies, ITL shows improvements over previous studies, achieving good precision at high recall levels

4. Training Deep Learning Models

Training BERT (Bidirectional Encoder Representations from Transformers) from scratch typically requires significant computational resources and large datasets, which may not be feasible for most individuals or organizations. Therefore, it's more common to fine-tune a pre-trained BERT model on a particular downstream task.

Sun et al. proposes generic solution related to fine-tuning the BERT model to address the text classification tasks [14]. The proposed solution consists of following steps:

1. Further pre-training BERT on within-task training data or in-domain data:

This step is optional, but it can be helpful to increase the model's performance. The within-task training data is the data for the specific task wish to tune BERT for and in-domain data indicates

the data from the targeted task domain.

2. Fine-tuning BERT with multi-task learning:

This step is also optional. Multi-task learning is a technique where we train a model on multiple tasks at the same time. This can help the model to learn more general features that are useful for all of the tasks.

3. Fine-tuning BERT for the target task:

This is final step, where we fine-tune BERT on the data for the specific task that we want to classify.

It's essential to remember, fine-tuning BERT typically requires substantial volume of labeled facts for the downstream task. Many deep learning frameworks, such as TensorFlow and PyTorch, provide pre-trained BERT models and libraries to simplify the fine-tuning process. Fig.5 demonstrate complete Pre & Fine-tuning process.

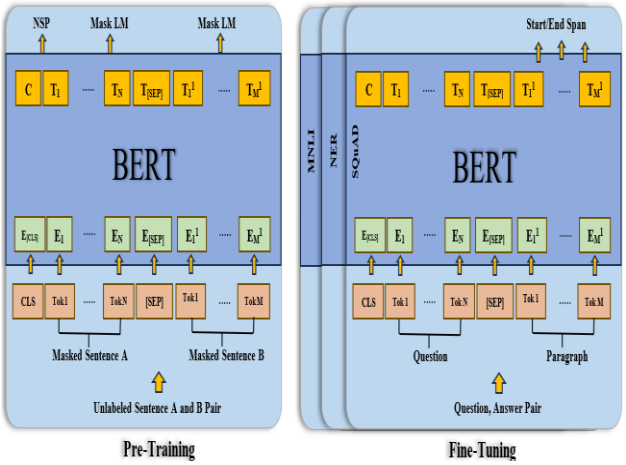

Fig. 5. BERT Pre-Training and Fine-Tuning

You et al. (2019) proposes a new method for training deep learning models called large batch optimization [19]. This method uses a much larger batch size than traditional methods, which permits the model to learn more efficiently. The authors of the paper experimented with large batch optimization on the BERT model, a large language model that is used for natural language understanding tasks. They found that large batch optimization was able to train BERT in 76 minutes, which is absolutely significantly faster than the previous state-of-the-art technique, which took 12 hours. The key idea behind large batch optimization is that it allows the model to learn more from each update. When the batch size is small, the model is only able to learn about a small number of examples at a time. However, when the batch size is large, the model is able to learn about a much larger number of examples at a time. This allows the model to learn more efficiently and converge to a better solution faster. The authors of the paper also found that large batch optimization was able to enhance the performance of BERT on natural language understanding tasks. They found that BERT trained with large batch optimization achieved a score of 93.5 on the GLUE benchmark, which is significantly higher than the score of 92.9 achieved by BERT trained with the previous state-of-the-art method.

5. Training Data Sets

For every model the training dataset is crucial. There are several opens datasets are available for various data science and Machine Learning research and model to be build. These datasets are useful for carrying out research on Natural Language processing, Computer Vision and Deep Learning etc. This study considers various datasets used for building NLP models. Cloud providers like Amazon offer open datasets for carrying out NLP researches. Their product reviews used as a training data set for building models for predicting sentiment analysis. They offered many datasets like SNIPS, ATIS, SLURP (SLU resource package) and their latest dataset is MASSIVE (Multilingual Amazon SLURP for Slot Filling, Intent Classification, and Virtual-Assistant Evaluation).

Unlike other datasets the MASSIVE consists of 1 million realistic parallel text spanning 51 languages in the earth including Tamil, Chinese, German, Spanish, French, Russian, Arab etc. This help to carry out cross language model build.

For performing sentiment analysis twitter has provided their data set. This dataset contains three sentiments, they are positive, negative and neutral. It contains 162980 unique data values out of them 35,509 are tweet text that are negative, 72,249 are positive and rest 55,212 are neutral tweet texts. GokulYenduri used this Twitter Dataset for building their Heuristic-Assisted BERT model for sentiment analysis [22]. [23] Uses the same Twitter Dataset for their work on sentiment analysis Target-Dependent Sentiment Classification with BERT. Zekeriya Anil used Turkish tweets as their training dataset for their work to compare BERT Models with other Machine Learning Methods for sentiment analysis on Turkish Tweets [9].

Multimodal Corpus of Sentiment Intensity (MOSI) is a dataset used for carry out sentiment analysis. This consist of 2199 opinion video clips, containing emotions in the range of (-3,3) indicating happiness, sadness, Anger, Disgust, Surprise and cry [21][24]. And the dataset Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) is one of the large datasets containing 23500 sentence utterance videos from over 1000 YouTube speakers. They are gender balanced and chosen from various topics. Jiaxuan He et al used both MOSI and MODEI for their wok on MF-BERT: Multimodal Fusion in Pre-Trained BERT for Sentiment Analysis.

AG News is another larger dataset build from over a million news articles gathered from over 2000 news sources. This dataset is useful for carrying out NLP based researches.

The dataset Text REtrieval Conference (TREC) consist of 5500 sentences in the form of questions. It contains average sentence length as 10. Size of its vocabulary is 8700. These data are collected from four different sources which includes manually constructed questions as well.

Like these many such datasets are available for carry out machine learning or building NLP models.

6. Gap Identification

Most existing sentiment classification models only consider the sentiment of the entire sentence, ignoring the sentiment of specific target entities. This can lead to inaccurate sentiment classification, as the sentiment of a sentence can be different from

the sentiment of the target entity.

There are a limited number of large-scale datasets available for fine-grained sentiment classification. This makes it difficult to train BERT models that are able to accurately classify text with multiple levels of sentiment.

BERT is trained with an enormous text dataset, but it is not possible to include every word in the English language in this dataset. This means that BERT may not be able to accurately classify text that contains out-of-vocabulary words.

The BERT model is only pre-trained on a single task, namely masked language modeling. This means that the model may not be able to generalize well to other tasks, such as answering the question or natural language inference.

The BERT model does not explicitly model the relationships between words, which may limit its ability to understand complex language structures.

The BERT model is computationally expensive to train and fine-tune.

BERT models have a maximum token limit (e.g., 512 tokens for the original BERT model), which can limit their ability to handle very long documents without truncation or other strategies.

While BERT excels at capturing contextual information, it can still struggle with resolving ambiguous language or situations where multiple interpretations are possible, as it does not have explicit reasoning abilities.

While BERT's transfer learning approach is powerful, it still requires task-specific labeled data for fine-tuning. In cases where obtaining labeled data is expensive or challenging.

Training BERT models from scratch on large datasets can be time-consuming, requiring substantial computational resources.

BERT does not explicitly model coreference resolution, which is the task of finding when words or phrases in a text refer to the same entity. Resolving coreference is important for understanding the full meaning of a text, but BERT relies on contextual embeddings to capture some level of coreference information.

BERT has been trained on a extensive dataset of text in English. However, it is important to evaluate its performance on other languages and dialects to ensure that it is generalizable.

7. Conclusion

This study focused on basic of Natural language processing and sentiment analysis. Deeply examined various research works on Natural Language processing and sentiment analysis. This review considers various techniques and algorithms to build NLP models and their effectiveness in serving their intended purpose. Also caried out assessment on the models by comparing their accuracy in making the prediction on sentiment analysis. The crucial part in all research work is the training data set being used. Deeply analysed various training data set available to use. Based on these reviews understood that for sentiment analysis depends on the outcome the appropriate training dataset has to be used. The dataset used for predicting the is positive or negative then the same dataset cannot be used for predicting the emotions like happy, sad, cry, suspense etc, Each research must use an appropriate dataset depends on their nature of their work.

References

[1] Chiorrini, A., Diamantini, C., Mircoli, A., & Potena, D. (2021). Emotion and sentiment analysis of tweets using BERT. EDBT/ICDTWorkshops.

[2] Koroteev, Mikhail. (2021). BERT: A Review of Applications in Natural Language Processing and Understanding. 10.48550/arXiv.2103.11943.

[3] M. Preetha, N. Anil Kumar, K. Elavarasi, T. Vignesh, V. Nagaraju “A Hybrid Clustering Approach Based Q-Leach in TDMA to Optimize QOS-Parameters”, Journal of Wireless Personal Communications Vol.123,Issue2, pages 1169–1200 (2022): 2 October 2021. ISSN No. 0929-6212 DOI:10.1007/s11277-021-09175-8

[4] Munikar, M., Shakya, S., & Shrestha, A. (2019). Fine-grained sentiment classification using BERT. In arXiv [cs.CL]. http://arxiv.org/abs/1910.03474.

[5] 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.

[6] A.M. Rajeswari, M. Mahalakshmi, R. Nithyashree and G. Nalini, "Sentiment Analysis for Predicting Customer Reviews using a Hybrid Approach," 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), Cochin, India, 2020, pp. 200-205, doi: 10.1109/ACCTHPA49271.2020.9213236.

[7] Shrestha, N., & Nasoz, F. (2019). Deep learning sentiment analysis of Amazon.com reviews and ratings. International Journal on Soft Computing Artificial Intelligence and Applications, 8(1), 01–15. https://doi.org/10.5121/ijscai.2019.8101

[8] M. Preetha, Raja Rao Budaraju, Jackulin. C, P. S. G. Aruna Sri, T. Padmapriya “Deep Learning-Driven Real-Time Multimodal Healthcare Data Synthesis”, International Journal of Intelligent Systems and Applications in Engineering (IJISAE), ISSN:2147-6799, Vol.12, Issue 5, page No:360-369, 2024.

[9] T. U. Haque, N. N. Saber and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," 2018 IEEE International Conference on Innovative Research and Development (ICIRD), Bangkok, Thailand, 2018, pp. 1-6, doi: 10.1109/ICIRD.2018.8376299.

[10] Geetha, M. P., & Karthika Renuka, D. (2021). Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. International Journal of Intelligent Networks, 2, 64–69.

[11] Z. A. Guven, "Comparison of BERT Models and Machine Learning Methods for Sentiment Analysis on Turkish Tweets," 2021 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 2021, pp. 98-101, doi: 10.1109/UBMK52708.2021.9559014.

[12] Santhosh Kumar, B., Geetha, M. P., Padmapriya, G., & Premkumar, M. (2020). An approach for improving the labelling in a text corpora using sentiment analysis. Advances in Mathematics: Scientific Journal, 9(10), 8165–8174. https://doi.org/10.37418/amsj.9.10.46.

[13] Xu, H., Liu, B., Shu, L., & Yu, P. S. (2019). BERT post-training for Review Reading Comprehension and aspect-based

sentiment analysis. In arXiv [cs.CL]. http://arxiv.org/abs/1904.02232.

[14] Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In arXiv [cs.CL]. http://arxiv.org/abs/1903.09588.

[15] S. Yu, J. Su and D. Luo, "Improving BERT-Based Text Classification With Auxiliary Sentence and Domain Knowledge," in IEEE Access, vol. 7, pp. 176600-176612, 2019, doi: 10.1109/ACCESS.2019.2953990.

[16] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In arXiv [cs.CL]. http://arxiv.org/abs/1905.05583.

[17] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. In arXiv [cs.CL]. http://arxiv.org/abs/1909.11942.

[18] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. In arXiv [cs.CL]. http://arxiv.org/abs/1907.11692.

[19] Ambalavanan, A. K., & Devarakonda, M. V. (2020). Using the contextual language model BERT for multi-criteria classification of scientific articles. Journal of Biomedical Informatics, 112(103578), 103578. https://doi.org/10.1016/j.jbi.2020.103578.

[20] Singh, M., Jakhar, A. K., & Pandey, S. (2021). Sentiment analysis on the impact of coronavirus in social life using the BERT model. Social Network Analysis and Mining, 11(1), 1–11. https://doi.org/10.1007/s13278-021-00737-z.

[21] You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., & Hsieh, C.-J. (2019). Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. In arXiv [cs.LG]. http://arxiv.org/abs/1904.00962.

[22] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional Transformers for language understanding. In arXiv [cs.CL]. http://arxiv.org/abs/1810.04805.

[23] J. He and H. Hu, "MF-BERT: Multimodal Fusion in Pre-Trained BERT for Sentiment Analysis," in IEEE Signal Processing Letters, vol. 29, pp. 454-458, 2022, doi: 10.1109/LSP.2021.3139856.

[24] Yenduri, G., Rajakumar, B. R., Praghash, K., & Binu, D. (2021). Heuristic-assisted BERT for Twitter Sentiment Analysis. International Journal of Computational Intelligence and Applications, 20(03). https://doi.org/10.1142/s1469026821500152.

[25] Z. Gao, A. Feng, X. Song and X. Wu, "Target-Dependent Sentiment Classification With BERT," in IEEE Access, vol. 7, pp. 154290-154299, 2019, doi: 10.1109/ACCESS.2019.2946594.

[26] Sun, Z., Sarma, P., Sethares, W., & Liang, Y. (2019). Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In arXiv [cs.LG]. http://arxiv.org/abs/1911.05544.

[27] Li, M., Li, W., Wang, F., Jia, X., & Rui, G. (2021). Applying BERT to analyze investor sentiment in stock market. Neural Computing & Applications, 33(10), 4663–4676. https://doi.org/10.1007/s00521-020-05411-7.

[28] Abdirahman, A. A., Hashi, A. O., Dahir, U. M., Elmi, M. A., & Rodriguez, O. E. R. (2023). Enhancing natural language processing in Somali text classification: A comprehensive framework for stop word removal. International Journal of Engineering Trends and Technology, 71(12), 40–49. https://doi.org/10.14445/22315381/ijett-v71i12p205.

[29] Syed Tanzeel Rabani, Maheswaran K, "Software Cognitive Complexity Metrics for OO Design: A Survey", International Journal of Scientific Research in Science, Engineering and Technology, Vol. 3 , No. 3, pp. 691-698, June 2017

[30] Maheswaran K, Aloysius A, "An Interface based Cognitive Weighted Class Complexity Measure for Object Oriented Design", International Journal of Pure and Applied Mathematics, Vol. 118, No. 18, pp. 2771-2778, 2018.