Itihaas: Analyzing Historical Journals through Sentiment Analysis

Shreyas Kalikar, Priyanshu Prasad, Rohit Yerramsetty, Amol Sinha, Vani Ruchika Pabba
{skalikar,pprasad,rohityerramsetty,amol.sinha,vaniruchikapabba}@ufl.edu

Abstract

This research project aims to perform sentiment analysis on historical journals to gain insights into the emotional and psychological experiences of authors, which are often overlooked in traditional historical accounts. Specifically, we compare the performance of three state-of-the-art models for sentiment analysis: BERT, DistilBERT, and RoBERTa. Our hypothesis is that transformer-based models, such as these, will outperform traditional deep learning models for sentiment analysis due to their ability to capture complex relationships between words in a sentence. We aim to identify which model outperforms the others in terms of accuracy, precision, recall and F1 score. Additionally, we investigate how the performance of these models is influenced by factors such as data pre-processing, hyperparameter tuning, and model architecture. The research findings will provide insights into the effectiveness of these models for sentiment analysis of historical journals, and potentially help inform decisions about which model to use in similar applications.

1 Introduction

Sentiment analysis is a crucial field of natural language processing (NLP) that involves analyzing text data to determine the subjective information, opinions, and emotions expressed within it. With the rise of digital communication, sentiment analysis has gained significant attention due to its applications in various fields such as social media monitoring, customer feedback analysis, and political analysis. However, despite its importance, sentiment analysis of historical texts or old documents has been largely unexplored.

In this research paper, we focus on sentiment analysis of old texts, particularly historical documents from the 19th century. Sentiment analysis of old texts is an emerging research area that has the potential to provide valuable insights into the past. The analysis of old texts can help to identify the cultural, social, and political attitudes of the time, as well as the evolution of language and linguistic structures over time. It can also be used to analyze the impact of historical events on people's emotions and opinions.

Despite the increasing use of deep learning models for sentiment analysis, there is a lack of studies comparing the performance of different models, particularly in the context of historical documents. Previous studies have mainly focused on using deep learning models for sentiment analysis of contemporary text, such as movie reviews and product reviews (Agarwal et al., 2011) (Maas et al., 2011). In the context of historical documents, there has been limited research on sentiment analysis, particularly with deep learning models.

To perform sentiment analysis on historical documents, we collected a dataset comprising historical documents from the 19th century. We labeled the dataset into three classes, namely positive, negative, and neutral, using the VADER sentiment analysis tool. VADER, (Hutto and Gilbert, 2015), is a rule-based sentiment analysis tool that is specifically designed for analyzing social media texts. It uses a lexicon-based approach, combined with grammatical rules and heuristics, to determine the sentiment of a piece of text.

We then evaluated the performance of three state-of-the-art transfer learning models, namely BERT, RoBERTa, and DistilBERT, on sentiment analysis of our labeled historical dataset. Transfer learning is a machine learning technique that involves using pre-trained models to perform a specific task. In our case, we used pre-trained models that were trained on large-scale text datasets to perform sentiment analysis on our historical dataset.

The primary objective of this research paper is to compare the accuracy, precision, recall, and F1-score of the three transfer learning models and determine which model performs best on our labeled historical dataset. The results of this research can

be used to improve sentiment analysis of old texts and historical documents and provide valuable insights into the sentiment and opinions of people in the past.

The organization of the paper is as follows; Section 2 discusses related works; Section 3 talks about the dataset and preprosessing; Section 4 highlights the different models used and their implementation. The model experiments are outlined in Section 5. The results obtained are presented and discussed in Section 6. In Section 7, the conclusion and future works are highlighted.

2 Related Work

Sentiment analysis has been a topic of interest in NLP research for many years, with various techniques and models proposed to tackle this problem.In this section, we review previous studies on sentiment analysis, particularly in the context of historical text, and deep learning models used for this task.

In recent years, deep learning models have shown great promise for sentiment analysis due to their ability to capture complex relationships between words in a sentence. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) were among the earliest deep learning models used for sentiment analysis (Kim, 2014) (Luong et al., 2013). However, with the introduction of transformer-based models, such as BERT, DistilBERT, and RoBERTa, these models have become less popular due to their superior performance. In a study by (Sanh et al., 2020), DistilBERT was introduced as a smaller and faster version of BERT, with comparable performance on several benchmark datasets.

Several studies have applied sentiment analysis to historical texts, including literature and social media data. In a study by (Hou and Frank, 2015), sentiment analysis was applied to Chinese classical poetry and they evaluated their lexicons intrinsically and extrinsically.

There have been several studies that have compared the performance of different deep learning models for sentiment analysis on various datasets. For example, (Reimers and Gurevych, 2019) compared the performance of BERT and other deep learning models for sentiment analysis of movie reviews and found that BERT achieved the highest performance. However, to the best of our knowledge, there has been limited research on sentiment analysis of historical documents, particularly with deep learning models.

3 Dataset and Preprocessing

In this section, the different stages of the sentiment analysis process are explained. The dataset was obtained, processed, and then inputted into several pretrained models. These models were fine-tuned using the dataset before making the final predictions.

3.1 Data Acquisition

The British-literature NLP phrases is a publicly available dataset(Kaggle, 2023) which was taken as the base dataset for this project. The dataset has been aggregated from famous British writers from the 14 to 21 centuries. The data set is labelled by the Name of the writer, Name of books, and Century. The data has been extracted from sentence by sentence from famous British novels, by NLP techniques.

We also chose 'Diary of a Young Girl' by (Frank, 1989) as our target historical journal to better understand and analyze the sentiments contained within its text.

3.2 Data Preprocessing

Before labeling the data, we must preprocess the text data in order to do sentiment analysis. To maintain consistency and get rid of extraneous material that might skew sentiment analysis results, preprocessing entails cleaning and standardizing the text input. To clean and normalize the text data for this specific sentiment analysis assignment, we used a variety of text preparation approaches. To guarantee uniformity, we first changed all text to lowercase. The URLs were then all eliminated from the text data because they don't offer any useful data for sentiment analysis. Then, in order to remove any extraneous data that can influence the sentiment analysis, we deleted any non-alphabetic characters from the text input aside from spaces.

We eliminated all English stopwords from the text data in order to decrease the dimensionality of the feature space. Stopwords are frequently used words in a language, such "the," "and," and "a," that have no major sentimental value. Stopwords can be eliminated so that we can concentrate on words that are more pertinent to the sentiment anal-

ysis task. To standardize the data, we also deleted any excess spaces and replaced any underscores in the text with spaces. To make sure the text data was in a uniform and standardized format, we also deleted any digits, numerals, and other punctuation.

We deleted all duplicate instances and non-English material using these preparation processes, leaving 19311 occurrences. After removing occurrences with three words or fewer, we were left with 15758 instances.

Figure 1: The word cloud of the most occurring words in the dataset shows that “said”, “would”,"one" are few of the most frequently used words in this dataset

The word cloud is a helpful text analysis tool that exhibits the most frequently used words in a text in a visual format by enlarging them and using different colors. The size and boldness of the word in the cloud are proportional to its frequency in the text, whereas the smaller words have less importance. This method provides a representation of the significance of words in the context of the text and their frequency.

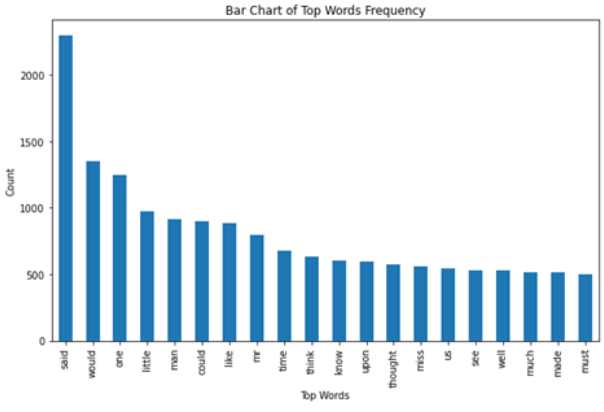

Figure 2: Frequency of Top Words in the Dataset

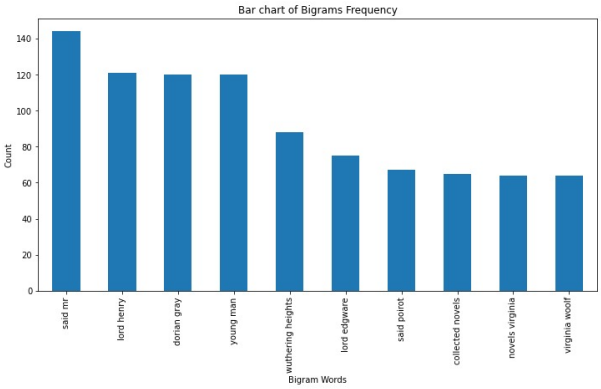

Figure 3: Frequency of Top Bigrams in the Dataset

The bar charts depicted in Figure 2 and Figure 3 represent frequency of the most commonly used words and most commonly used bigrams in this dataset.Further, we employed the VADER sentiment analysis method to label the dataset. A sentiment score is given to a piece of text by VADER, a lexicon- and rule-based sentiment analysis tool, depending on the presence of particular words and rules. Although it is a commonly used technique in sentiment analysis, not all datasets or projects may benefit from its use. In some circumstances, alternative techniques like hand-annotated data or machine learning-based methods may be more suitable. We discovered that the dataset had 6913 positive, 4598 neutral, and 4247 negative labeled statements after classifying it. It is significant to observe that there are more positive than negative phrases in the class distribution, which makes it less balanced. This distribution may have an impact on the sentiment analysis’s accuracy since models developed using unbalanced datasets may be biased toward the majority class. As a result, this must be considered when evaluating the sentiment analysis’s findings.

4 Methodology

After deliberation, we considered implementing the following Sentiment Analyzers that we thought would be a good fit for SA on large sets of literature:

4.1 VADER

We first created a collection of old texts and historical journals in order to perform sentiment analysis on a corpus of British literature. We divided the dataset into three separate categories—positive,

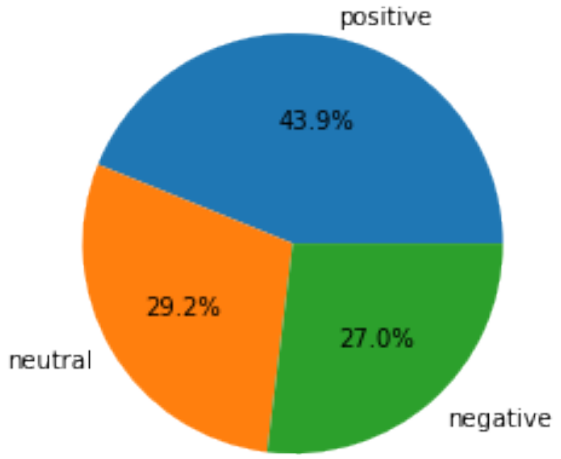

Figure 4: Classification chart of old English text available in the dataset.

negative, and neutral—in order to facilitate categorization. We used VADER, (Hutto and Gilbert, 2015), a sentiment analysis tool designed specifically for content analysis, to do this work. To determine the tone of a text, VADER employs a lexicon-based technique together with grammatical conventions and heuristics. By applying this process, VADER can identify the text's overall emotional tone and categorize it into one of the three groups stated above. We were able to divide the British literature dataset into classifications that were favorable, negative, and neutral by using VADER. As a result, we were able to examine the corpus's general sentiment and draw important conclusions from the information.

4.2 BERT

A cutting-edge pre-trained transfer learning model called BERT (Bidirectional Encoder Representations from Transformers) was created by Google for applications involving natural language processing(Devlin et al., 2019). It is an attention-based neural network design that can recognize the context of a word inside a phrase. It is pre-trained using a masked language model and a task that predicts the next phrase on a significant amount of text input. Using smaller labeled datasets, BERT may subsequently be fine-tuned for certain downstream applications, such as sentiment analysis. Using the Python ktrain module, we develop BERT for sentiment analysis in this work. The data is initially preprocessed by the code using the tokenization and embedding function of BERT. The classification BERT model is then created, and is trained using the preprocessed data.

4.3 DistilBERT

A compact, quick, and light Transformer model named DistilBERT,(Sanh et al., 2020), is brought into light by distilling BERT basis, and is about 40% the size of BERT while being 60% faster. This model stood put when it comes to processing large amounts of data because of it's effectiveness and quickness. Its tokenizer preprocesses the data, and then uses this clean data to train the DistilBERT model.

DistilBERT uses knowledge distillation to train a smaller model to mimic a bigger pre-trained model called BERT. By employing this method, DistilBERT may learn from the same pre-training data as BERT more successfully and gain from its generalization skills.

DistilBERT provides various changes to BERT's architecture and training procedure in addition to knowledge distillation. Included in these adjustments are a smaller transformer design with fewer layers and lower hidden layer sizes, as well as a changed training method that promotes quicker convergence and a smaller memory footprint.

On a number of NLP tasks, DistilBERT has shown performance that is equivalent to the original BERT model while being easier to install and train due to its lower size. We used the "distilbert-base-case" model in this project.

4.4 RoBERTa

RoBERTa was launched in 2019 as an enhanced version of the BERT language model by Facebook AI,(Liu et al., 2019),. RoBERTa was only pre-trained using the masked language modeling (MLM) job with a significantly bigger corpus of text data, in contrast to BERT, which was pre-trained on a variety of tasks with varied aims. As a result, RoBERTa was able to create text representations that were more robust and generic, which enhanced its performance on subsequent NLP challenges.

In addition to the pre-training procedure, RoBERTa makes a number of changes to the BERT architecture and training procedure. These modifications include increasing the batch size during pre-training, dynamic text masking, and training epoch lengths with a wider variety of training data.

It is a well-liked option for many NLP applications thanks to its improved efficiency and stability.

4.5 Fine Tuning of Models

4.5.1 BERT

The ktrain package is used to fine-tune sentiment analysis using the BERT model. The dataset is first prepared by separating it into training and testing sets and transforming the text data into arrays. The class_names option is used to specify the class names, which are positive, negative, and neutral. The maxlen and max_features parameters are used to determine the maximum length of input text and maximum number of features in the preprocessor, respectively. The preprocess_mode option is set to 'bert' to utilize the BERT preprocessor. Using the text_classifier and get_learner functions, we build a BERT text classifier model and a learner object. The validate function with the class_names parameter is used to assess the model's performance on the test data after it has been trained using the fit_onecycle function and a one-cycle learning rate strategy. The algorithm employs a one-cycle learning rate strategy with two epochs, a maximum sequence length of 350, a maximum number of features of 3000, and a batch size of 6. The BERT model is a flexible transfer learning model that can be tailored for different NLP applications, and ktrain offers an intuitive interface for doing so.

4.5.2 RoBERTa and DistilBERT

We first create an instance of the Transformer class from the ktrain library, giving the name of the pretrained RoBERTa or DistilBERT model to use, as well as configuring hyperparameters like maxlen and class_names. The preprocess_train and preprocess_test methods, which encode and tokenize the input text and cap sequence length at 500 characters, are then used to preprocess the training and validation data. Using the get_classifier function of the Transformer class, we build a classification model that encapsulates the pre-trained model. The classification model, the preprocessed training data, and other hyperparameters like batch size are then sent to the get_learner function of the ktrain library. The method employs a cyclic learning rate schedule to maximize model parameters and is used to train the model. For each sentiment class in the labeled dataset, which we describe using class names as a list of numbers [0, 1, 2], the validate method assesses the model using the preprocessed validation data, providing accuracy and other metrics. The important hyperparameters for RoBERTa and DistilBERT in this project are maxlen=500 and class_names=[0, 1, 2], which respectively represent the dataset's three sentiment classes and the maximum input sequence length. Other hyperparameters, such as learning rate and batch size, are either hardcoded in the code or are defaulted to those values.

4.6 Evaluation Metrics

We utilized four evaluation metrics for the sentiment analysis task:

- Accuracy, a simple measure of a classifier's performance, which indicates the proportion of correctly classified instances to the total number of instances.
- Precision, which measures the ratio of true positives to the total number of instances predicted as positive. It assesses the proportion of positive predictions that are genuinely true positives.
- Recall, also known as sensitivity or true positive rate, which measures the ratio of true positives to the total number of actual positive instances in the dataset. It evaluates the proportion of actual positives that the classifier correctly identifies.
- F1-score, which is the harmonic mean of precision and recall, providing an overall evaluation of a classifier's performance. In the case of imbalanced datasets, it is a more equitable measure than accuracy, as it takes both false positive and false negative rates into account.

5 Model Experiments

All the experiments were performed using the GPU hardware accelerator platform of Google Colab.The objective of our first experiment was to fine-tune three pre-trained NLP models on British-literature and compare their performance before and after fine-tuning. Transfer learning is a machine learning method that uses a pre-existing model's knowledge and adjusts it to a particular task. The three models used in this experiment were BERT, RoBERTa and DistilBERT, which are advanced transfer learning models that were pretrained using a self-supervised learning technique

on extensive text data. They are capable of learning intricate language patterns and features that can be tailored to perform several NLP tasks, such as sentiment analysis.The dataset was preprocessed to include only text data and sentiment labels categorized into three classes: negative, neutral, and positive.The experiment was conducted in two stages: before fine-tuning and after fine-tuning. In the first stage, the three pre-trained models were evaluated on the dataset to establish a baseline performance for the task. In the second stage, the models were fine-tuned on the dataset using transfer learning techniques to improve their performance on the sentiment analysis task. This experiment is aimed to provide insights into the effectiveness of fine-tuning pre-trained models on historical text data for the task of sentiment analysis, and to inform future work on sentiment analysis of historical text.

Further, to test the performance of sentiment analysis models on a real-world data, we selected the diary of a young girl by Anne Frank as the dataset. This diary provided a rich source of text to explore and analyze the sentiments expressed by the young girl during her life in hiding.Even after extensive research, we were not able to find a pre-labelled dataset for this book or any other similar book from same era, so the dataset was manually labeled to serve as our ground truth , and pre-trained models, including DistilBERT, BERT, and RoBERTa, were used to perform sentiment analysis on the text.The aim of the experiment was to gain insights into how the young girl felt about her life in hiding during the war.We compared the performance of the three pre-trained models before and after fine-tuning on the selected dataset. By doing so, the experiment sought to identify the best-performing model for sentiment analysis on this type of text data.

6 Results and Discussion

In this section, a comprehensive analysis of the outcomes achieved by each model is presented. The time taken for the completion of each model, its accuracy, performance, and the challenges encountered during the process are explicated. The results presented in Figure 5 clearly demonstrate the importance of fine-tuning pre-trained BERT model for sentiment analysis of journal text. The baseline BERT model without fine-tuning achieved very poor performance, with an accuracy of only 0.41 and an F1 score of 0.31. This indicates that pre-trained models may not be directly applicable to historical texts without additional adaptation to the specific task and domain.

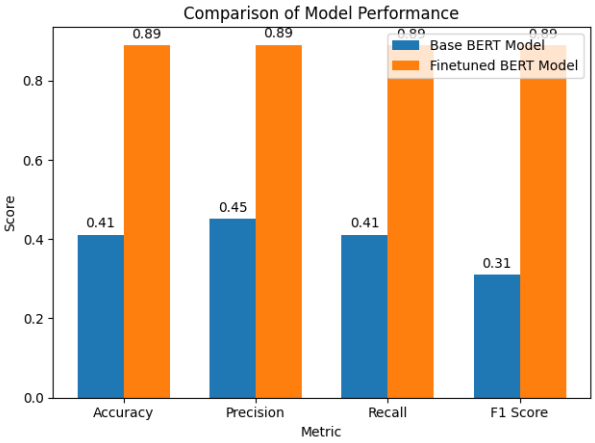

Figure 5: Comparison of Base BERT and Finetuned BERT Model

However, when the same BERT model was fine-tuned on the historical journal, the results significantly improved. The fine-tuned BERT model achieved an accuracy of 0.89 and an F1 score of 0.89, indicating its high effectiveness for sentiment analysis of historical texts. This suggests that fine-tuning the pre-trained models with domain-specific data can significantly enhance their performance.

Therefore, the main finding of this study is that fine-tuning pre-trained models with historical text data can significantly improve their performance for sentiment analysis of old texts or journal texts. This is especially important because historical texts may contain archaic language, cultural references, and different writing styles that are not present in contemporary texts, which can affect the performance of pre-trained models. Fine-tuning allows the model to adapt to the specific language and style of the historical text, resulting in improved accuracy and F1 score.

The results presented in Figure 6 of using the base Roberta model without fine-tuning on the old text or journal text dataset for sentiment analysis are significantly lower than the results obtained after fine-tuning the model on the same dataset. The base Roberta model achieved an accuracy of only 0.27, precision of 0.23, recall of 0.27, and an F1 score of 0.12. This suggests that the base Roberta model was not well-suited for sentiment analysis on this dataset without fine-tuning. However, after fine-tuning the model on the dataset, the accuracy,

precision, recall, and F1 score all improved significantly, achieving an accuracy of 0.88, precision of 0.88, recall of 0.88, and an F1 score of 0.88. These findings indicate that fine-tuning the Roberta model on the old text or journal text dataset improved its performance for sentiment analysis significantly.

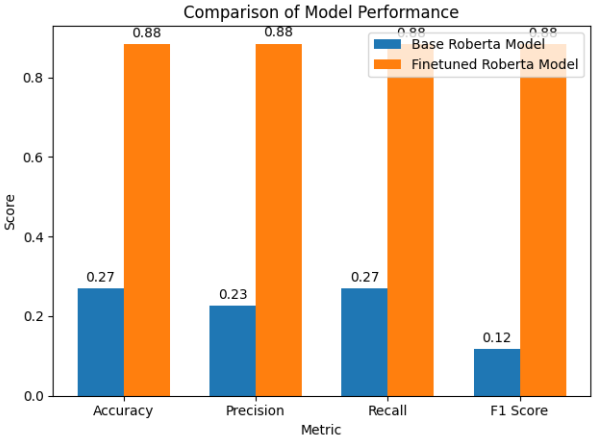

Figure 6: Comparison of Base Roberta and Finetuned Roberta Model

The results of DistilBERT are presented in Figure 7. The results show that the base models of all three models - BERT, Roberta, and DistilBERT - performed poorly with very low accuracy, precision, recall, and F1 score. This is likely due to the fact that the base models were not trained on the specific domain of old texts or historical journals, which can have a unique style and language compared to contemporary text.

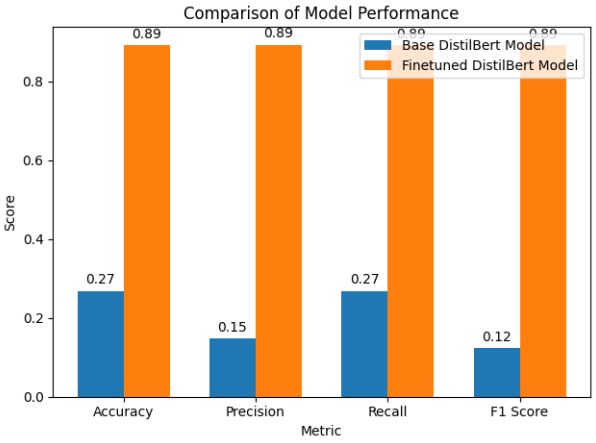

Figure 7: Comparison of Base DistilBert and Finetuned DistilBert Model

However, after fine-tuning the models on the dataset, the results improved significantly with accuracy, precision, recall, and F1 score all reaching 0.88 or above for all three models. This suggests that fine-tuning is a crucial step in achieving good performance for sentiment analysis on old texts or historical journals.

Furthermore, the fact that all three models achieved similar results indicates that they are all suitable for the task and the choice of which model to use may depend on factors such as computational resources or specific requirements of the task.

Table 1: Results Obtained using All Base Models

Model	Accuracy	Precision	Recall	F1 Score
BERT	**0.41**	**0.45**	**0.41**	**0.31**
Roberta	0.27	0.23	0.27	0.12
DistilBERT	0.27	0.15	0.27	0.12

Table 2: Results Obtained using All Finetuned Models

Model	Accuracy	Precision	Recall	F1 Score
BERT	0.89	0.89	0.89	0.89
Roberta	0.88	0.88	0.88	0.88
DistilBERT	0.89	0.89	0.89	0.89

Figure 8: Tables representing the results of base models and fine-tuned models

Table 1 and Table 2 presents the results of base models and fine-tuned models respectively. From the results obtained in the experiments, it can be seen that all three models - BERT, Roberta, and DistilBERT - perform significantly better after fine-tuning on the British-literature dataset than their base versions without fine-tuning. The fine-tuned BERT model achieved the highest accuracy, precision, recall, and F1 score among the three models, with an accuracy of 0.89 and an F1 score of 0.89. The fine-tuned Roberta model achieved the second-highest accuracy and F1 score, with an accuracy of 0.88 and an F1 score of 0.88, followed by the fine-tuned DistilBERT model, which also achieved an accuracy of 0.88 and an F1 score of 0.88. The base versions of all three models, without fine-tuning, performed poorly compared to the fine-tuned models. The base BERT model achieved an accuracy of 0.41 and an F1 score of 0.31, while the base Roberta and DistilBERT models achieved an accuracy of 0.27 and an F1 score of 0.12. Overall, the results show that fine-tuning the pre-trained models on the specific domain dataset can significantly improve the performance of the models for sentiment analysis of old texts or historical journals. The fine-tuned BERT model is the most effective model among the three for this task.

After successfully training and fine-tuning our sentiment analysis models on the chosen dataset and achieving promising results, we decided to test our approach in a more real-world setting. To accomplish this, we selected the diary of a young girl by Anne Frank as a suitable choice. We evaluated the performance of all three pre-trained models, namely DistilBERT, BERT, and RoBERTa and we compared the results before and after fine-tuning the models.

The results we obtained were fascinating, providing a unique glimpse into the mind of a young girl during a tumultuous time in history. We discovered that the diary entries were predominantly neutral in tone, with occasional fluctuations towards more positive or negative sentiments.

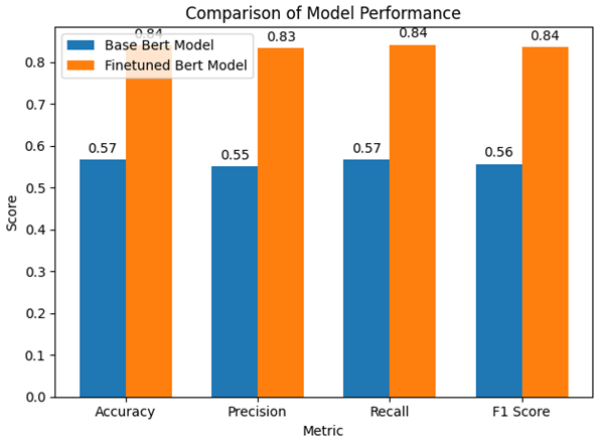

Figure 10: Comparison of Base Bert and Finetuned Bert Model for Diary of a Young Girl

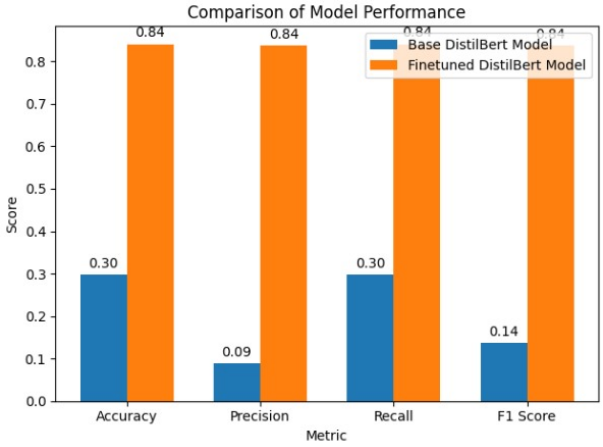

Figure 9: Comparison of Base DistilBert and Finetuned DistilBert Model for Diary of a Young Girl

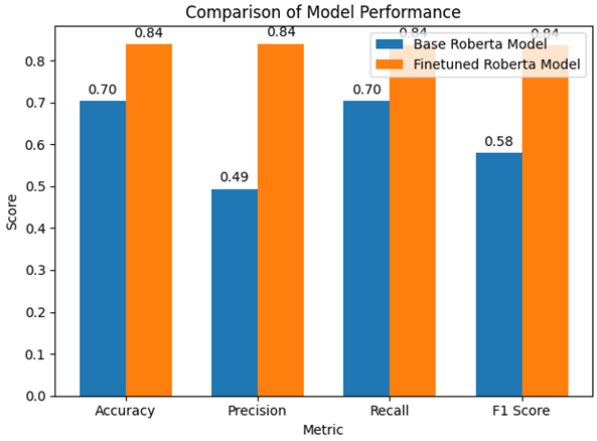

Figure 11: Comparison of Base Roberta and Finetuned Roberta Model for Diary of a Young Girl

- DistilBERT:

Before fine-tuning, the accuracy, precision, recall, and F1 score of DistilBERT were 0.297, 0.088, 0.297, and 0.136, respectively. However, after fine-tuning, the performance of the model improved significantly, with an accuracy of 0.841, precision of 0.838, recall of 0.841, and F1 score of 0.837.

- BERT:

The pre-fine-tuning results of BERT were an accuracy of 0.568, precision of 0.551, recall of 0.568, and F1 score of 0.557. After fine-tuning, the accuracy of the model improved significantly, with an accuracy of 0.841, precision of 0.835, recall of 0.842, and F1 score of 0.836.

- RoBERTa:

Without fine-tuning, the accuracy, precision, recall, and F1 score of RoBERTa were 0.703, 0.494, 0.703, and 0.580, respectively. After fine-tuning, the performance of the model improved, with an accuracy of 0.839, precision of 0.840, recall of 0.838, and F1 score of 0.836.

Overall, we observed that fine-tuning improved the performance of all three transformer models for the sentiment analysis task. RoBERTa and BERT showed similar performance after fine-tuning, whereas DistilBERT had a lower F1 score. However, DistilBERT still achieved an acceptable F1 score after fine-tuning.

7 Conclusion and Future Work

From the results obtained it can be concluded that the accuracy and precision displayed by the fine-

Model	Fine-tuning	Accuracy	Precision	Recall	F1 Score
DistilBERT	Before	0.297	0.088	0.297	0.136
	After	0.841	0.838	0.841	0.837
BERT	Before	0.568	0.551	0.568	0.557
	After	0.841	0.835	0.842	0.836
RoBERTa	Before	0.703	0.494	0.703	0.580
	After	0.839	0.840	0.838	0.836

Figure 12: Comparison of all models pre and post fine-tuning for Diary of a Young Girl

tuned models in this project are mostly similar. However base RoBERTa model displayed a better understanding of context from the texts and displayed a much higher accuracy compared to the other models pre-finetuning when used on a journal.

This experiment demonstrates that sentiment analysis can be effectively applied to literary works, providing valuable insights into the emotions and perspectives of the characters. The results obtained from this analysis have significant implications for the fields of psychology, literature, and history, and they provide a unique opportunity to gain a better understanding of the human experience during times of conflict and adversity. With continued research and development, it is likely that we will see more sophisticated NLP models that could surpass the results produced by present models.

Another area where we can expect to see progress is in the analysis of texts in languages other than English. While NLP techniques are successfully applied to English in historical texts other languages would present their own set of challenges. An essential improvement that could be made to this project is to incorporate sentiment analysis with a time series model. By doing so, we can observe how sentiments evolve over extended historical periods.

Finally, the analysis of historical texts using NLP techniques has potential to provide new insights and culture of different time periods. By analyzing large volumes of historical texts, it is possible to uncover patterns and trends that were not previously visible.

8 Github Link

https://github.com/priyanshu0499/Itihaas

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, Portland, Oregon. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Anne Frank. 1989. The diary of a young girl.

Yufang Hou and Anette Frank. 2015. Analyzing sentiment in classical chinese poetry. pages 15–24.

C.J. Hutto and Eric Gilbert. 2015. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Kaggle. 2023. British literature nlp labelled phrase dataset. https://www.kaggle.com/datasets/ahmadalijamali/british-literature-nlp-labeld-phrase.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.