A Comparative Study of Pretrained Text Classification Models

1. Introduction

This project aims to simplify the selection of NLP models for text classification tasks. With the advent of transformer-based models like BERT, XLNet, XLM-RoBERTa, and DistillBERT, the task of choosing an appropriate model has become challenging due to the sheer number of options. We aim to alleviate this by examining these models across different tasks. Text classification, a key NLP application, involves assigning predefined labels to text. Our project will compare these models across various datasets and tasks, aiming to understand each model's strengths and limitations. Key objectives include training models on different tasks, comparing their performance using metrics, and analyzing each model's strengths and weaknesses. The goal is to identify the most efficient model in terms of accuracy, resource consumption, and implementation complexity for each task.

Some important objectives of this project are:

1. To train various models on different text classification tasks.
2. To compare the performance of these models using various metrics.
3. To analyze the strengths and weaknesses of each model.

Research Goal: Determining the Optimal NLP Models for Various Text Classification Tasks

By conducting this survey, we aim to identify the most efficient model in terms of accuracy, resource consumption, and implementation complexity for each task. We plan to include a visual representation of the training progression of the models, which will assist in determining the most suitable model for larger, similar tasks. The ultimate objective of this project is to offer a valuable resource that can guide researchers and practitioners in the field of NLP, enabling them to make informed decisions when tackling text classification tasks.

2. Methodology

2.1 Sentiment Analysis

2.1.1 Data: The Python Reddit API Wrapper (PRAW) was used to access Reddit's content and features and extract the top comments from various subreddits. The data was cleaned and formatted by removing stopwords, punctuation, and URLs, and tokenizing the text into words, handling sarcasm, gifs, etc. This resulted in a dataset ready for sentiment analysis.

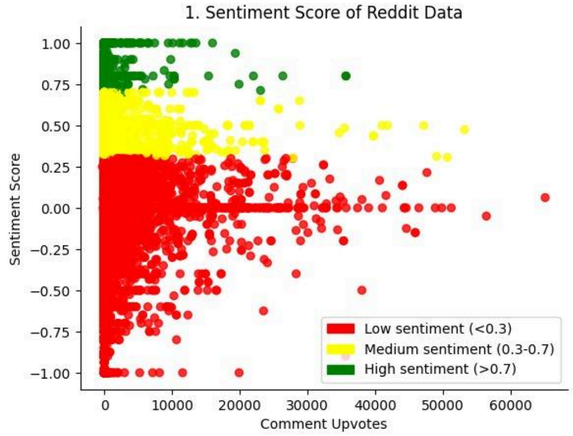

Fig: Distribution of sentiment scores

2.1.2 Modeling:

2.1.1 BERT: The model and the tokenizer were loaded from the Hugging Face library, using the bert-base-uncased version. The data was split into training and testing sets, with 10% of the data reserved for testing. The texts were tokenized and padded to a maximum length of 128, and the labels were converted to one-hot encoding. The model architecture consisted of the BERT model followed by a dense layer with softmax activation. The model was compiled with Adam optimizer, categorical cross-entropy loss, and accuracy metric. The model was trained for two epochs, using a batch size of 16.

2.1.2.2 DistilBert: DistilBERT uses the same architecture and tokenizer as BERT, but with fewer layers and parameters. The data preprocessing and splitting steps were the same as for BERT and RoBERTa. The model was trained for two epochs, using a batch size of 8. The confusion matrices showed that the model had some difficulty in distinguishing between neutral and positive sentiments, as well as between negative and positive sentiments. This could be due to the reduced capacity and generalization ability of DistilBERT compared to BERT and RoBERTa.

2.1.2.3 Roberta: The next model that was used in the project was RoBERTa, which is a variant of BERT that has been optimized for better performance. RoBERTa uses the same architecture and tokenizer as BERT, but with some differences in the training data, hyperparameters, and pre-training objectives. The data preprocessing and splitting steps were the same as for BERT. The model was trained for two epochs, using a batch size of 10.

2.1.2.4 XLNet: The final model that was used in the project was XLNet, which is another variant of BERT that uses a different pretraining objective and a permutation-based language modeling technique. XLNet uses a different architecture and tokenizer than BERT, RoBERTa, and DistilBERT, which are based on the Transformer encoder. XLNet uses a two-stream self-attention mechanism and a Transformer-XL decoder, which can capture long-term dependencies and handle variable-length inputs. The model and the tokenizer were loaded from the Hugging Face library, using the xlnet-base-cased version.

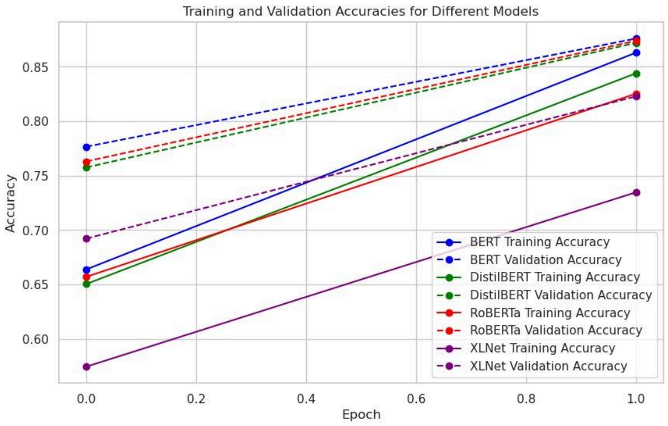

Fig: Performance of the models for the same set of hyperparameters

2.2 Multilingual Classification

2.2.1 Data: In this dataset, the primary focus is on multilingual sentiment analysis with the goal of identifying the most effective model for analyzing sentiment in diverse language data. The dataset consists of movie reviews and sentiment labels from various internet sources in 2020, spanning languages such as Czech (cs), German (de), Spanish (es), French (fr), Polish (pl), and Slovak (sk). With 96,000 movie review comments in the training set and 10,000 in the testing set.

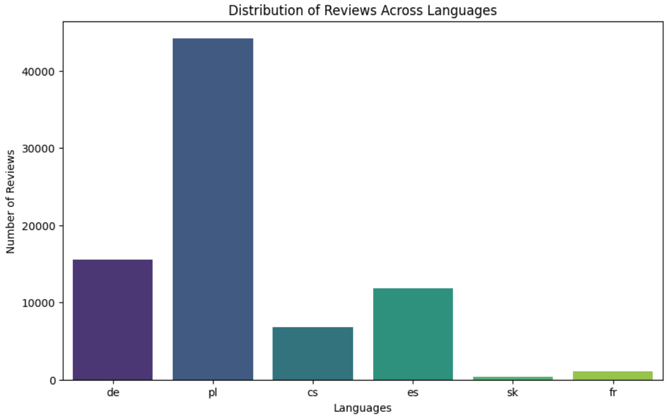

Fig: Distribution of Movie Reviews

2.2.2 Modeling:

2.2.2.1 BERT: We implemented a sentiment analysis model using the BERT architecture 'bert-base-multilingual-cased' from Hugging Face's transformers library. Fine-tuned the labels, the model showed an accuracy of 0.6 on our multilingual movie review dataset, demonstrating its effectiveness in discerning sentiments, including instances labeled as neutral or not assigned.

2.2.2.2 DistilBERT: Despite its similarity to BERT, DistilBERT demonstrated comparable performance with an accuracy of 0.55. Its notable advantage lies in demanding less computational power during training. Employing a batch size of 16, the model efficiently processed multilingual movie review data, making it a resource-efficient choice for sentiment analysis tasks.

2.2.2.3 RoBERTa: This model demonstrated efficiency in training, taking less time than BERT but more than DistilBERT. The model exhibited an accuracy of 0.66, showcasing its efficacy in capturing sentiment nuances, including instances labeled as neutral or not assigned.

2.2.2.4XLNet: This model emerged as the best-performing model with an accuracy of 0.74. In addition to its superior performance, XLNet's unique autoregressive approach sets it apart, enabling it to capture nuanced relationships in the data. Its comprehensive methodology outperformed all other models in our multilingual movie review dataset analysis.

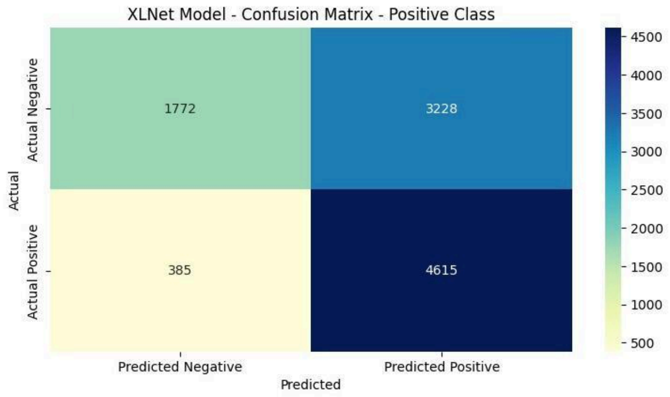

Fig: Confusion Matrix for XLNet Model

2.3 Stance Detection

2.3.1 Data: The stance detection dataset consists of a news article body and headline. The purpose of the task is to identify how the headline is related to the body i.e. whether the headline agrees, disagrees, discusses or is unrelated to the body. The train dataset consists of 49972 entries and the test dataset consists of 25413 entries. The stance column contains values - ‘unrelated’, ‘discuss’, ‘agree’ and ‘disagree’ depending on the relation between the body and the headline.

2.3.2 Modeling:

2.3.2.1 BERT: The database was tokenized using the BertTokenizer. We used attributes max length 256, train batch size 8 and validation batch size 4 and added a dense 768x768 layer with 0.3 dropout and a classifier layer 768x4. The original weights of the pretrained BERT were frozen and the model was trained over 2 epochs. The maximum accuracy obtained was 0.72.

2.3.2.2 DistilBERT: DistilBERT is a smaller model and the time required to run this model was lesser than that of BERT. The accuracy decreased to 0.71 which is still much worse compared to the BERT model.

2.3.2.3 RoBERTa: RoBERTa gave the highest accuracy over one epoch - 0.929. The improvements RoBERTa made over BERT are visible with the higher accuracy score. One more reason for additional accuracy is that the RoBERTa train dataset includes news articles which are similar to the dataset being used for this task.

2.3.2.4XLNet: XLNet also had high performance 0.917. Its performance gains are due to the same reasons as RoBERTa.

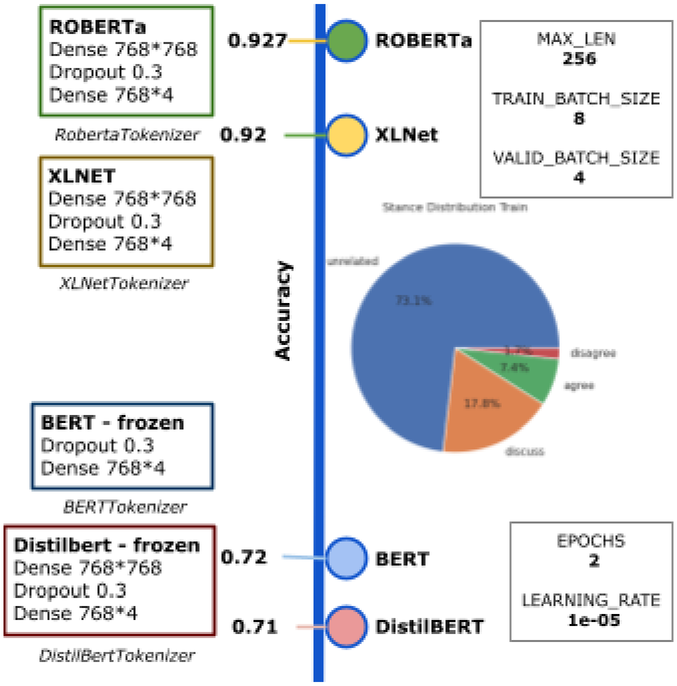

Fig. Stance Detection Dataset and results

2.4 Multiclass Classification

2.4.1 Data: In this wikipedia article dataset, the focus is to identify the genre of the movies from the wikipedia article. This dataset consists of the title, plot, and genres of the movies from 1901 to 2017. There are in total 32432 unique movies and 33869 plots. We then sample roughly an equal number of movie plots from different genres to reduce class imbalance issues. Choosing movie genres based on their frequency.

2.4.2 Modeling:

2.4.2.1 BERT:

Implementing BERT, a transformer-based language model, using the simpletransformers library and a tokenizer. Our model achieved an MCC of 0.48 on our Wikipedia article dataset, which falls slightly below the 0.5 threshold. This result suggests that BERT may be struggling with the highest evaluation loss, hinting at potential overfitting issues. One notable limitation of BERT is its handling of token masking, as it treats predicted tokens as independent entities, which can lead to discrepancies in fine-tuning.

2.4.2.2 DistilBERT:

Through the implementation of DistilBERT, using the simpletransformers library and tokenizer, we achieved an MCC of 0.54 after training. This result surpasses BERT's performance and exceeds the 0.5 threshold, indicating a notably strong model performance. DistilBERT strikes a fine balance between predictive accuracy and generalization.

2.4.2.3 RoBERTa:

RoBERTa, which is also implemented using the simple transformers library and tokenizer and trained on the Wikipedia dataset, achieved an MCC of 0.56 after training. This performance surpasses both BERT and DistilBERT, demonstrating its robustness and strong performance.

2.4.2.4 XLNet:

XLNet achieved the highest MCC score of 0.61 on our Wikipedia dataset, showcasing its exceptional balance between precision and recall. XLNet employs a permutation language modeling approach, enabling it to capture token dependencies more effectively.

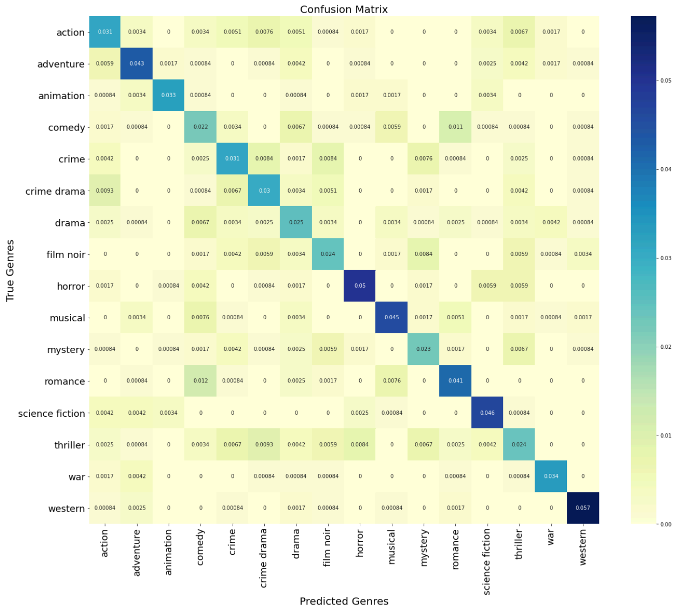

Fig:BERT Confusion Matrix

3. Results and Conclusion

Metrices									
	Models	Accuracy	Precision	F1 Score	Recall	MCC score	Log Loss	Hamming Loss	Speed(Ranking)
Sentiment Analysis	BERT	0.95	0.93	0.89	0.86	0.78	0.91	-	4
	Roberta	0.93	0.88	0.86	0.84	0.66	1.10	-	3
	DistilBERT	0.87	0.78	0.79	0.81	0.59	2.13	-	1
	XLNet	0.90	0.88	0.88	0.87	0.66	1.91	-	2
Winner	BERT	BERT	BERT	BERT	XLNet	BERT	BERT		DistilBert
Multilingual Classification	BERT	0.6	0.56	0.56	0.57	0.411	-	0.4	4
	Roberta	0.66	0.63	0.61	0.60	0.496	-	0.34	3
	DistilBERT	0.55	0.52	0.52	0.51	0.384	-	0.45	2
	XLNet	0.74	0.71	0.70	0.68	0.584	-	0.26	1
Winner	XLNet	XLNet	XLNet	XLNet	XLNet	XLNet		XLNet	XLNet
Stance Detection	BERT	0.72	0.72	0.75	0.72	-	4.77	0.11	4
	Roberta	0.929	0.929	0.935	0.929	-	2.133	0.0324	1
	DistilBERT	0.71	0.71	0.722	0.71	-	9.81	0.137	4
	XLNet	0.92	0.918	0.923	0.919	-	2.13	0.034	2
Winner	Roberta	Roberta	Roberta	Roberta	Roberta			Roberta	Roberta
Multiclass Classification	BERT	0.52	0.80	0.78	0.77	0.48	2.10	0.48	4
	Roberta	0.54	0.81	0.79	0.79	0.56	1.81	0.46	3
	DistilBERT	0.53	0.78	0.77	0.78	0.54	2.15	0.47	1
	XLNet	0.50	0.77	0.74	0.75	0.61	2.23	0.50	2
Winner	Roberta	Roberta	Roberta	Roberta	Roberta	XLNet	XLNet	XLNet	DistilBERT

In conclusion, this research has delved into a comprehensive exploration of state-of-the-art Natural Language Processing (NLP) models, aiming to identify the optimal choices for diverse text classification tasks. Our investigation considered key facets such as accuracy, resource consumption, and implementation complexity, shedding light on the nuanced interplay of these factors.

Sentiment Analysis:

In the domain of sentiment analysis, BERT emerged as the preeminent performer, exhibiting superior accuracy. However, its computational demands rendered it comparatively slower. DistilBERT, despite its efficiency, sacrificed a degree of accuracy. XLNet showcased commendable performance, while RoBERTa delivered results of notable merit.

Multilingual Text Classification:

For multilingual text classification, XLNet demonstrated a remarkable balance between performance and efficiency, outshining both BERT and DistilBERT.

Stance Detection Task:

Our investigation revealed that RoBERTa excelled in the nuanced task of stance detection, boasting optimal accuracy and swift training times. XLNet emerged as a formidable contender with similar accuracy but higher training time, whereas DistilBERT and BERT demonstrated limitations in this specific context.

Multiclass Classification:

In the realm of multiclass classification, RoBERTa asserted its dominance by achieving the highest accuracy. DistilBERT, as the fastest option, presented compromises in accuracy, while BERT maintained a satisfactory standing.

Overall Insights:

The trade-off between accuracy and speed became evident throughout our analysis. BERT consistently demonstrated high accuracy but at the expense of computational time. DistilBERT, being faster, compromised on accuracy in some tasks. XLNet and RoBERTa emerged as versatile contenders, exhibiting strong performance across various text classification tasks.

Recommendations:

The choice of the optimal NLP model depends on the specific requirements of the task. For tasks prioritizing accuracy, XLNet or RoBERTa might be suitable, understanding the associated computational costs. DistilBERT, with its efficiency, is a viable option for scenarios where speed is paramount, and a slight drop in accuracy is acceptable. BERT is a complex model and can be used if enough resources are available.

Future Directions:

As NLP models continue to evolve, future research could explore hybrid approaches or task-specific fine-tuning to further enhance model efficiency without compromising accuracy. Additionally, advancements in model architectures may introduce novel options with improved trade-offs.

References:

[1] Yusuf Arslan, Kevin Allix, Lisa Veiber, Cedric Lothritz, Tegawendé F. Bissyandé, Jacques Klein, and Anne Goujon. 2021. A Comparison of Pre-Trained Language Models for Multi-Class Text Classification in the Financial Domain. In Companion Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 260–268. https://doi.org/10.1145/3442442.3451375

[2]Garrido-Merchan, E. C., Gozalo-Brizuela, R., & Gonzalez-Carvajal, S. . (2023). Comparing BERT Against Traditional Machine Learning Models in Text Classification. Journal of Computational and Cognitive Engineering

[3]Cristóbal Colón-Ruiz, Isabel Segura-Bedmar, Comparing deep learning architectures for sentiment analysis on drug reviews, Journal of Biomedical Informatics, Volume 110, 2020, 103539, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2020.103539

[4]Z. Li et al., "A Unified Understanding of Deep NLP Models for Text Classification," in IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 12, pp. 4980-4994, 1 Dec. 2022, doi: 10.1109/TVCG.2022.3184186

[5]Openai, Alec, et al. Improving Language Understanding by Generative Pre-Training. 2018.

TeamWork:

There were 4 text classification tasks finalized by the team and each member worked on a task each, here is a short summary of their works.

Aryan:Identified popular text classification tasks that can be used for text classification and metrics for them. Created the Dataset for sentiment analysis and did Exploratory Data Analysis. Worked on the models for the Sentiment Analysis task.

Vishnu: Read research papers to identify the models(Transformers) that can be used for text classification. Identified useful datasets for the relevant tasks. Worked on models for Stance detection and fine tuned models to give the best results.

Harsh: Reviewed research papers on multilingual data and understand data with help of Exploratory Data Analysis. Implemented sentiment analysis on multilingual movie review dataset.

Brian: Reviewed research papers that discuss various techniques for enhancing datasets and have also explored different models, including IWV embeddings. Additionally, located pre-trained code that could be valuable. Involved in the development of models for multiclass classification based on Wikipedia movie plots, leveraging both plot and genre information.