A Comparative Study of Pretrained Text Classification Models

Team: Aryan Patil, Brian Lu, Harsh Ramani, Vishnu Raja

Research Goal

Determining the Optimal NLP Models for Various Text Classification Tasks

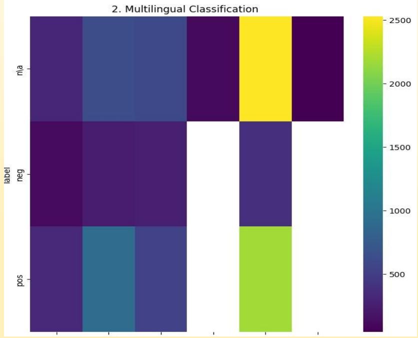
By conducting this survey, we aim to identify the most efficient model in terms of accuracy, resource consumption, and implementation complexity for each task.

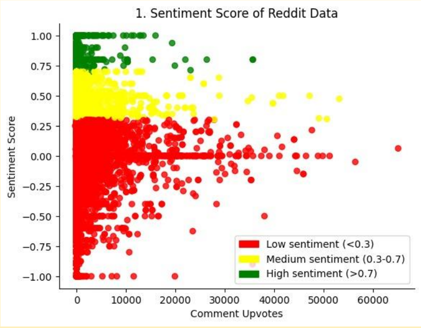
Introduction

In the field of Natural Language Processing (NLP), choosing the right model for a specific task can significantly impact the performance and efficiency of the system. While numerous NLP models exist, understanding which model is best suited for a particular text classification task remains a challenge. In recent years, the landscape of Natural Language Processing (NLP) has been reshaped by the advent of transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers), XLNet, XLM-RoBERTa, (Universal Language Model Fine-tuning), DistillBERT. These models have set new benchmarks in a variety of NLP tasks, including the fundamental task of text classification. However, the sheer number of available models can make the selection of an appropriate model for text classification a daunting task. This project seeks to alleviate this issue by conducting a thorough examination of these cutting-edge pretrained models across different text classification tasks.

Exploratory Data Analysis

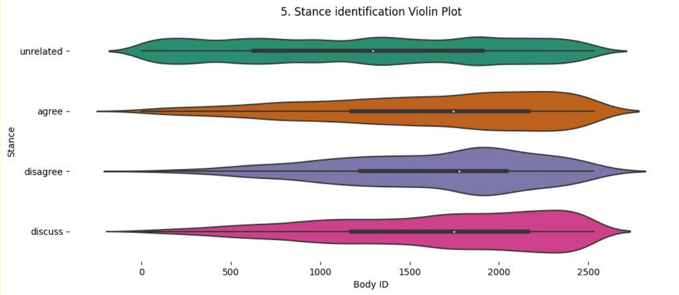
The multilingual dataset comprises movie reviews collected from diverse internet sources in 2020, spanning multiple languages such as Czech (cs), German (de), Spanish (es), French (fr), Polish (pl), and Slovak (sk).


2. Multilingual Classification


1. Sentiment Score of Reddit Data

We utilized PRAW (Python Reddit API Wrapper) to extract headline data from the selected subreddits to extract information including post titles and relevant metadata. Later, employed a sentiment analysis tool (VADER) to categorize headlines into Positive, Negative, and Neutral sentiments.

The stance detection dataset consists of two parts - a body and a headline. The goal of the task is to identify how the headline is related to the task. I.e. whether it agrees, disagrees, discusses or is unrelated to the task.


5. Stance identification Violin Plot

The wikiplot dataset consists of information on 34,886 global movies with columns including Release Year, Title, Origin/Ethnicity, Director(s), main actors/actresses, Genre(s), Wiki Page URL (source of plot description), and a detailed plot description.

Method

In this study, we will be using a variety of modern, pretrained models that have been proven to be highly effective in various NLP tasks. The models include BERT, XLNet, XLM-RoBERTa, , and DistillBERT. Each of these models brings unique strengths to text classification tasks. We will fine tune the models and analyze their usefulness in different text classification tasks.

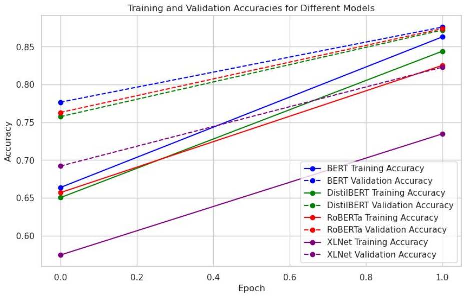
Text Classification Tasks

We will tackle the following text classification tasks and datasets -

- Sentiment analysis - Reddit Comments Dataset:. Identify the sentiments of Reddit comments in top subreddits over various categories.
- Multilingual classification - Multilingual Movie Reviews Dataset: Identify the multilingual movie reviews to be positive or negative
- Multiclass classification - Wikipedia Articles Dataset: Identify the genre of the movies from the Wikipedia article.
- Stance detection - Stance Detection Dataset: Identify whether a comment agrees, disagrees, or is neutral to a statement.

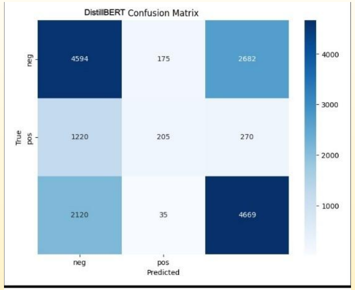
Sentiment Analysis

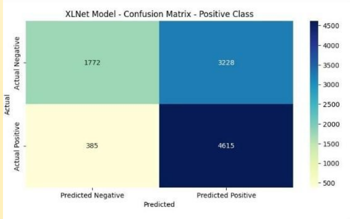
In our sentiment analysis task, BERT demonstrated the highest accuracy at 93%(for a different set of hyperparameters), albeit with increased computational demands. DistilBERT offered a good compromise between accuracy and efficiency, proving to be the fastest among the models. RoBERTa, while maintaining a competitive performance, strikes a balance in terms of both speed and accuracy. exhibited moderate accuracy, with performance influenced by task specifics. XLNet, on the other hand, recorded the lowest accuracy at 72%, highlighting its complexity and resource-intensive nature. These findings underscore the importance of considering trade-offs between model accuracy and computational resources when selecting an appropriate sentiment analysis solution.



Multilingual Classification

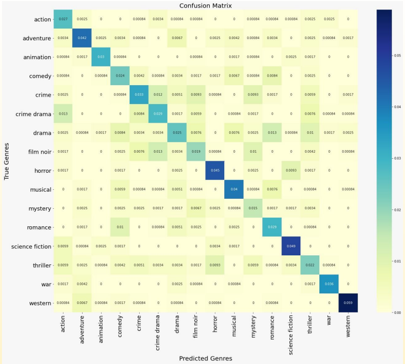
In the evaluation of Transformer models on a multilingual dataset, XLNet demonstrated superior performance with an accuracy of 0.74, despite its lower computational demand. Conversely, DistilBERT, despite its high computational power, performed the worst with an accuracy of 0.55. This highlights potential inefficiencies in its performance. BERT and RoBERTa, also with high computational power, did not yield the best results.



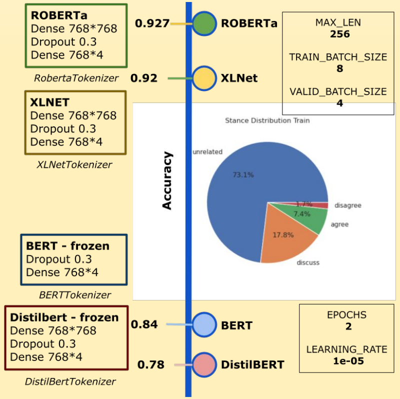


Multiclass Classification

Analyzing the results of the Matthew's Correlation Coefficient (MCC), it becomes evident that XLNet outperforms the other four models, including Roberta. XLNet not only achieves a higher MCC score but also exhibits robust performance, surpassing the 0.5 threshold, indicating its precision and recall balance. DistilBERT also achieves a commendable MCC score, demonstrating its competitiveness. In contrast, BERT struggles with the highest evaluation loss, implying potential overfitting issues and indicating its limitations in this context.



Stance Detection



ROBERTa
Dense 768*768
Dropout 0.3
Dense 768*4

RobertaTokenizer

XLNET
Dense 768*768
Dropout 0.3
Dense 768*4

XLNetTokenizer

BERT - frozen
Dropout 0.3
Dense 768*4

BERTTokenizer

Distilbert - frozen
Dense 768*768
Dropout 0.3
Dense 768*4

DistilBertTokenizer

0.927 **ROBERTa**
0.92 **XLNet**
0.84 **BERT**
0.78 **DistilBERT**

Accuracy

MAX_LEN
256
TRAIN_BATCH_SIZE
8
VALID_BATCH_SIZE
4

EPOCHS
2
LEARNING_RATE
1e-05

The notable performance gap among these models can be attributed to differences in architecture, pre-training strategies, and contextual understanding. RoBERTa and XLNet, leveraging advancements in model architecture and pre-training techniques, tend to capture more nuanced contextual information.

Conclusion

We conducted a comprehensive comparison of the pretrained models BERT, DistilBERT, XLNet, and RoBERTa, evaluating their performance across diverse databases, including two-class, multiclass, multilingual, and text relation datasets. The results give information aiding in the selection of the most appropriate one based on the specific requirements of a given task.

	BERT	DistilBERT	XLNet	Roberta
Sentiment_Anaylsis	0.95	0.87	0.90	0.93
Multilingual_Class	0.6	0.55	0.74	0.66
Multi-Class	0.48 (MCC)	0.54 (MCC)	0.61 (MCC)	0.56 (MCC)
Stance_Detection	0.84	0.78	0.92	0.927

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.