# BanglaNewsClassifier: A Machine Learning Approach for News Classification in Bengali Newspapers.

TANZIR HOSSAIN (20301154), BRAC University, Bangladesh
AR-RAFI ISLAM (20301164), BRAC University, Bangladesh
ANNAJIAT ALIM RASEL, BRAC University, Bangladesh

This paper introduces BanglaNewsClassifier, a machine learning method for classifying news articles written in Bangla. The rapid growth of multilingual digital content calls for more efficient strategies for organizing and classifying information, including news distribution. However, compared to the progress made in the dominant languages, there is a gap in the needs of languages that do not cater as much as Bangla. By customizing state-of-the-art transfer and natural language processing algorithms to the linguistic nuances and data characteristics of Bangla texts, BanglaNewsClassifier aims to close this gap. The system's effectiveness in classifying potential outcomes was confirmed by evaluation utilizing real-world Bangla news datasets across several categories. By illuminating the method's suitability for a range of languages and situations, the research advances the democratization of knowledge across linguistic boundaries. All things considered, BanglaNewsClassifier represents a major breakthrough in automatically classified news for unaided Bangla.

**ACM Reference Format:**
Tanzir Hossain (20301154), Ar-Rafi Islam (20301164), and Annajiat Alim Rasel. 2018. BanglaNewsClassifier: A Machine Learning Approach for News Classification in Bengali Newspapers.. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

The rapid growth of multilingual digital content has made the creation of efficient techniques for information classification and organization necessary. Among these, news story classification is essential for facilitating recommendation, analysis, and information retrieval systems. There is still a significant gap in satisfying the needs of languages with fewer computational resources, like Bangla, despite the fact that news classification for widely spoken languages like English has advanced significantly.

In this work, we introduce BanglaNewsClassifier, a cutting-edge machine-learning algorithm created specifically for Bangla newspaper news classification. We aim to bridge the gap between cutting-edge news classification methods and the needs of Bangla-speaking users by drawing on recent research such as "Exploring the Limits of Transfer Learning" by Raffel et al. (2020), which clarifies scaling strategies and transfer learning techniques in natural language processing.

Acknowledging the importance of languages and the unique characteristics of Bangla, we approach Vaswani et al.'s "Attention Is All You Need" (2017) and Devlin et al.'s "BERT: Training Deep Bidirectional Transducers for Pre-Language

Authors' Contact Information: Tanzir Hossain (20301154), BRAC University, Bangladesh; Ar-Rafi Islam (20301164), BRAC University, Bangladesh; Annajiat Alim Rasel, BRAC University, Dhaka, Bangladesh.


Manuscript submitted to ACM

Comprehension " (2018). We want to adapt the existing methods to the linguistic nuances and data features in Bangla texts by combining classic machine learning algorithms with deep learning algorithms

The general goal of BanglaNewsClassifier is to provide a versatile and efficient solution for passive news categorization in Bangla, hence satisfying the growing demand for personalized news recommendations and retrieval systems in Bangla-speaking areas. Using insights from "Exploring the Limits of Transfer Learning" by Raphael et al. (2020) Our strategy aims to improve performance and generalizability across various media and data kinds.

Through a series of tests and assessments on real-world Bangla news datasets, we show that our technique is successful and resilient in terms of producing accurate and trustworthy news categorization results. We apply insights from "Exploring the Limits of Transfer Learning" to assess our approach's scalability and transferability to other languages and circumstances. This provides direction for future study in Bangla natural language processing.

To summarize, the BanglaNewsClassifier is an essential step in improving the categorization of news in underutilized languages, particularly Bangla, contributing to the accessibility of information and knowledge across language boundaries in a democracy.

## 2 BACKGROUND

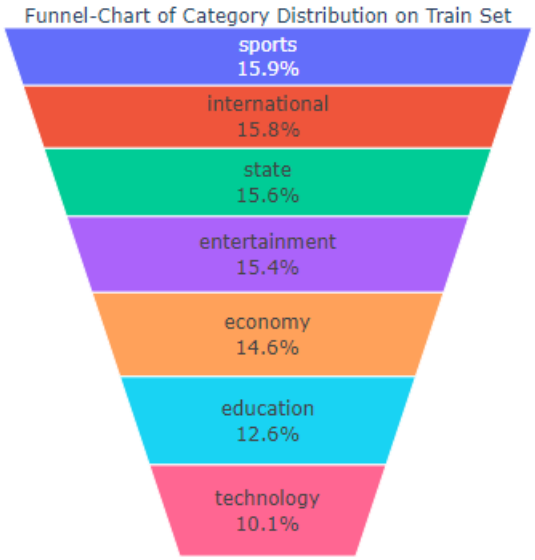
To begin with, The media influences public opinion and conversation on a variety of social topics. In Bangla journalism, a complete, data-driven understanding of the landscape is required to enable study in fields like journalism studies, linguistics, cultural studies, and so on. The proliferation of online news sources, the availability of web scraping tools, and the compilation of enormous datasets of media permitted the development of machine learning approaches for content analysis and categorization.

Secondly, media research often includes content analysis, sentiment analysis, and the detection of probable biases or emotional tones in the issues addressed. Decision trees, Naive Bayes classifiers, and random forests have all been shown to be useful in dealing with these difficulties. This technique can train on labeled datasets to detect emotional sensitivity. Potential biases or trends in coverage can be observed.

Moreover, existing research has looked at the use of machine learning algorithms for media classification analysis across languages and geographies. However, there is a need for focused research on some linguistic and cultural characteristics of Bangla media. Using web scraping techniques and machine learning models, this work aims to provide insights into the Bangla media landscape, facilitating a better understanding of diversity, emotional content language, linguistic features, and biases that can occur in the media across a wide range of groups and communities.

Finally, In order to do this, we scraped and assembled the dataset from the websites of prominent Bangla newspapers using the Selenium web automation tool. To avoid being discovered by anti-bot defenses, the scraping operation used the random interval technique to target the websites of prominent Bangla publications. 98,884 news items total from six categories—international, sports, technology, national, financial, and entertainment—are included in the final dataset.

```
Category
economy          14488
education        12818
entertainment    15261
international    15314
sports           15883
state            15183
technology        9937
Name: categoryId, dtype: int64
```




The data collection is formatted as a CSV file with two columns: one for the category title and the other for the media content. Because of their excellent quality, news stories from prominent sources require little data preparation. However, various preprocessing measures were taken to enhance data quality.

An Excel file was created listing Bangla alphabet words obtained from external reference sources. These character lines were then extracted from the data set using the Python Bangla-stemmer module. Additionally, indicators were extracted from the media to ensure data accuracy. Exploratory analysis was conducted to assess the quality of the data. This study examined the distribution of news stories across the six categories to determine the level of data set balance or skewness. Furthermore, performing summary statistics showing the distribution using histograms or box plots, and dividing the data analyzed over media duration into six categories indicated a weighting of data to make them irrelevant that they are new equilibrium strategies. I looked for common issues in each category and studied the key themes and themes in each theme.

# 3 METHOD

## 3.1 Objectives

Our primary goal is to create and test BanglaNewsClassifier, a machine learning-based system for classifying news items published in Bangla. Our primary objective is to develop a versatile and reliable model for classifying Bangla media across themes and places. We specifically strive to achieve the following goals.Creation and deployment of a strong media categorization system adapted to Bangla features.

(1)Creation and deployment of a strong media categorization system adapted to Bangla features. (2)Collected and kept various Bangla media news data for training and monitoring distributors. (3)Assessing classification performance with established assessment measures and benchmark data sets. (4)Comparing the performance of BanglaNewsClassifier to existing approaches and cutting-edge models for news categorization in Bangla.

## 3.2 Study Selection

Our team followed a rigorous process to maintain relevant data types and analytics resources to achieve our goals. We lay great emphasis on finding different data in various media and places stored in Bangla language. The main criteria for choosing a data system are its relevance to the media industry, data quality, diversity and accessibility. One such data set has been obtained by deleting information from the integrated website. In addition, we critically review research studies and detailed academic papers on methods, techniques and performance criteria for media classification in Bangla. To ensure the accuracy and timeliness of our search, we prioritize recent publications from reputable sources.

## 3.3 Sources of Evidence and Search Strategy

To gather relevant information and resources for our study, we use a comprehensive interdisciplinary search strategy. We search publicly available datasets, academic databases, digital libraries, and online archives to find relevant Bangla media and research materials. We use keywords, phrases, and specific Boolean operators to find relevant documents and information. In addition, we seek advice from subject matter experts in the field and actively participate in discussions in relevant online communities, with the goal of identifying new resources and perspectives to support our research.

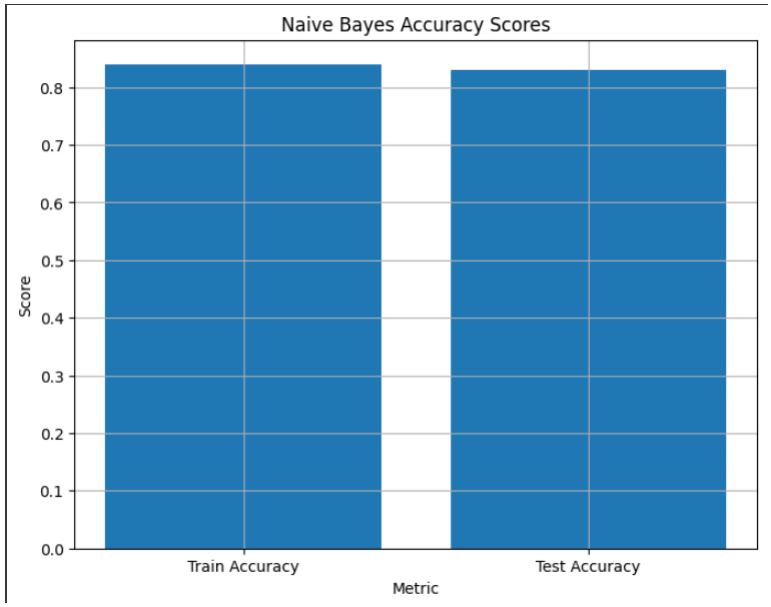
## 3.4 Data Extraction

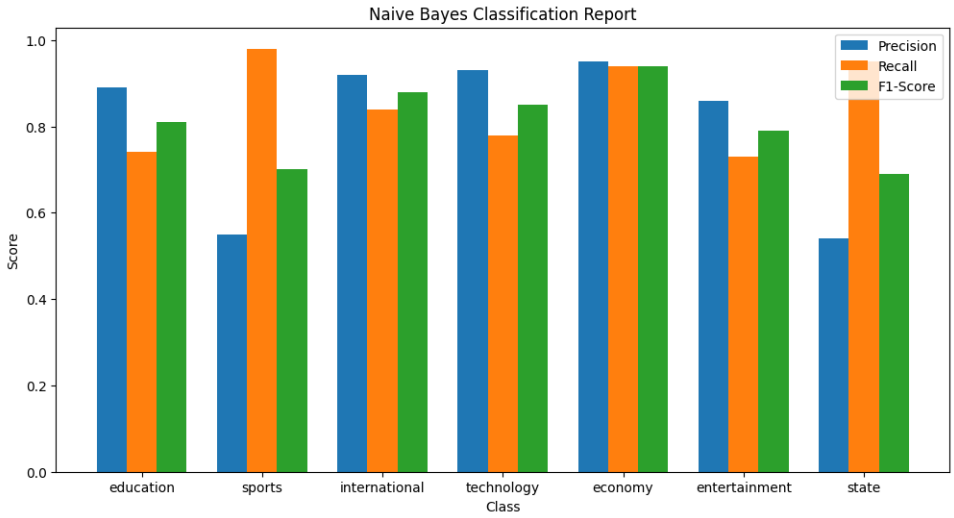
Our data extraction methods involve the meticulous collecting and structuring of information from particular data and research articles. We extract textual content, metadata, categorization, publication date, and other pertinent data from Bangla reports. In research, we extract data such as technique, research setting, performance indicators, and comparative analysis. We guarantee the retrieved information's correctness and completeness through rigorous verification and validation methods. This meticulous data extraction procedure is required to assure the authenticity and reliability of our study findings.

# 4 RESULTS

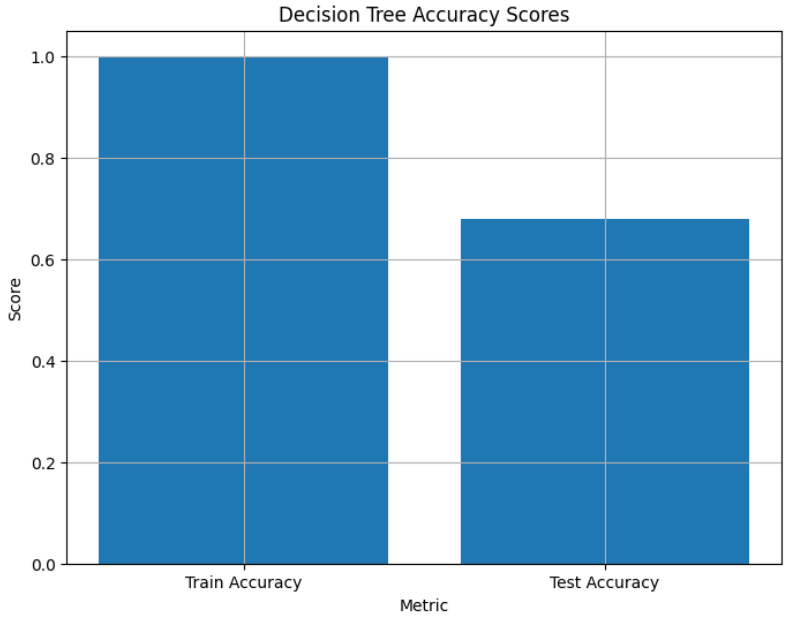
For the text classification task, we examined three types of machine learning algorithms: Naive Bayes, Decision Tree, and K-Nearest Neighbor. Here is an analysis of the results:

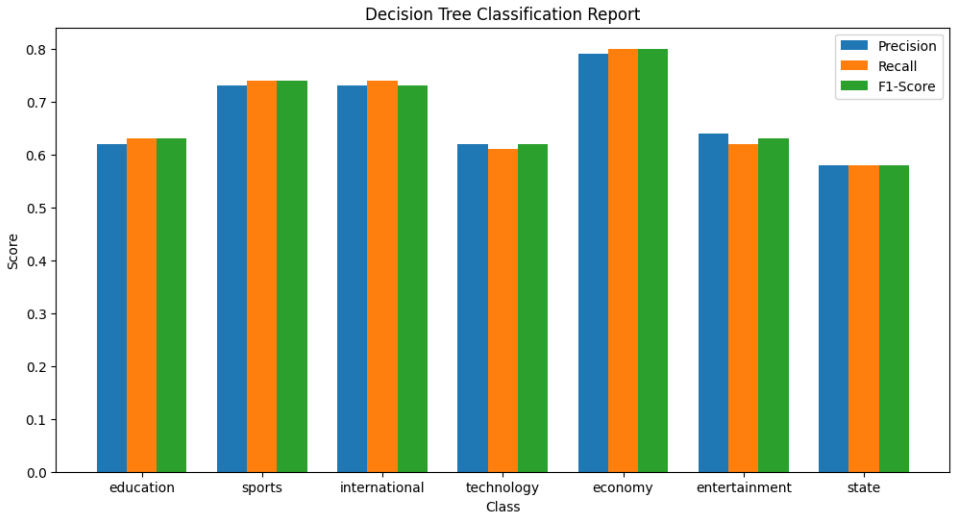
The naive Bayes classifier performed reasonably well, with 84% rail and 83% test accuracy. Individual classes have different precision, recall, and F1 scores, and some outperform others. For example, the 'Economy' category has the highest F1-score (0.94), while the 'Sports' category has the highest F1-score (0.70).



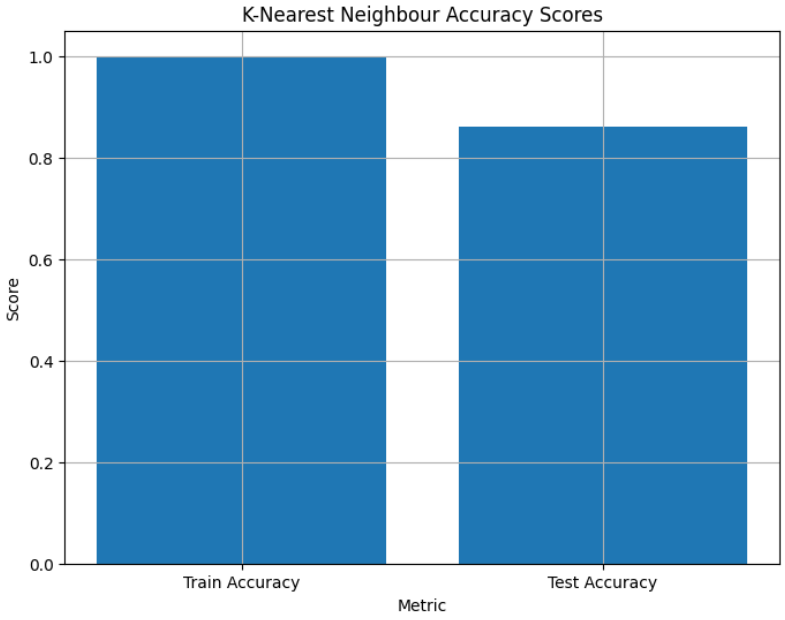



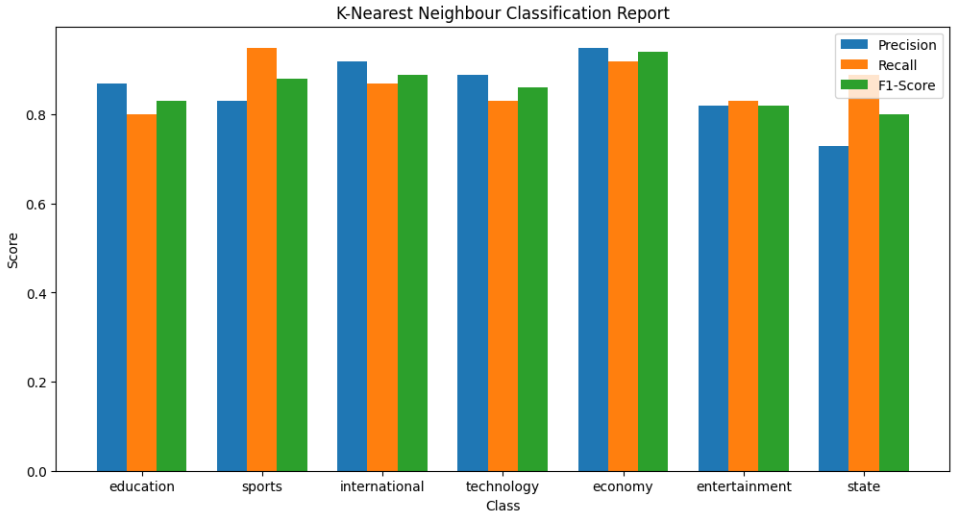

In comparison, the Decision Tree classifier reacted quite differently. It had a perfect train accuracy of 100%, meaning that it fitted the training data incredibly well. However, the test accuracy dropped to 68%, indicating that the model was overfitted to the training data and did not generalize well to the new data. Individual classes have lower accuracy, recall, and F1 than the Naive Bayes classifier.




The K-Nearest Neighbor (KNN) classifier achieved 100% train accuracy but had a significantly better test accuracy of 86%, outperforming both Naive Bayes and Decision Trees. Precision, recall, and F1-scores for specific classes are frequently higher than those of the other two models, indicating stronger overall performance.




Several variables contribute to performance differences: Bias-variance trade-off - Decision tree classifiers have high variance, which means they are too complex and accumulate noise in the training data, leading to overfitting Conversely , Neither Bayes nor KNNs have a good bias-variance balance, which leads to good generalization. Feature Representation - The performance of these algorithms also depends on the feature representation of the text data. Naive Bayes assumes feature independence, which may not be true in all cases, while KNN and decision trees can assume more complex feature interactions. Hyperparameter Tuning - The performance of these algorithms can be further improved by changing their hyperparameters, such as the number of neighbors in the KNN or the depth of the decision tree.

Several techniques can be used to improve results: Feature Engineering - Explore advanced feature engineering techniques to better represent text, such as TF-IDF, passwords, or topic map Ensemble methods - Composite using the power of different algorithms. Improve performance and combine models using ensemble methods such as bagging, boosting, or stacking Hyperparameter Tuning - Tune hyperparameters using methods such as grid search or random search to find the best hyperparameter value for each algorithm Advanced Algorithms - Advanced machine learning algorithms to find, for example Support vector machines (SVMs), deep neural networks, or transformer-based models (e.g., BERT), which can generate complex patterns in the data Data Enhancements to Take - If the data set minimize, modify and size training data Consider data enhancement techniques such as synonyms substitution, page translation, or artificial data.

## 5 DISCUSSION

We studied how multilingual models could be modified for different tasks and languages through In-Context Cross-Lingual Transfer (IC-XLT) Our study demonstrated the superior performance of IC-XLT when contrasted with other basic approaches and its capability to navigate situations with restricted resources. We are now delving into the significance of our results within the context of previous studies and other written works Our research shows that IC-XLT is more effective than gradient-based strategies like 1S-XLT and traditional methods such as Zero-Shot Cross-Lingual Transfer (ZS-XLT) When there is an inadequate amount of training data for the source language, this advantage becomes more apparent.The use of the IC-XLT gradient enhances the output to different languages by adding annotations to the target language in a specific context, without the need for additional resources for better maintenance This review highlights the idea that multilingual models can be more versatile tasks and languages examine IC-XLT in terms of introductory methods, such as ZS- . XLT, 1S-XLT, IC and -XLTSRC pairing shows that IC-XLT consistently outperforms these procedures in terms of different analytical designs and data entry IC-XLT exhibits greater efficiency in the target language of limited data to be used to enhance language transfer, highlighting its potential as a cost-effective solution for multilingual text classification services. It is not explained

Constraints on Resources: Our research into the performance of IC-XLT in limited resource situations is a big addition Through evaluating the model's ability to adjust to different languages lacking enough original language data, we demonstrate that IC-XLT performs equally as well as or even surpasses standard methods. This ending stresses the need for efficient cross-language transfer techniques, like IC-XLT, in cases with scarce training assets . Correlation with Linguistic Distance and Pretraining Data: Our research delves into how the effectiveness of target-language demonstrations is influenced by the closeness to English, and the presence of these languages in the initial data The research identified a significant link between progress in performance and the amount of target language included in the pretraining data This indicates that tongues with limited inclusion in pretraining collections reap greater rewards from altering to the goal language with IC-XLT undefined .

Integration with Prior Research: The results we obtained add to the existing knowledge on transfer learning across different languages and learning methods in specific situations The examination of IC-XLT as opposed to acknowledged fundamental approaches is in line with the current research on cross-lingual transfer methods In addition, our investigation expands the use of in-environment teaching for multilingual transfer assignments, giving valuable perspectives on its efficiency in this scenario undefined .

Our research indicates that IC-XLT effectively enhances multilingual models for different tasks and languages by integrating target-language examples within the context, surpassing traditional baseline methods in cross-lingual tasks. These findings show the advancement of cross-lingual transfer learning and highlight the potential of in-context

learning techniques for improving model adaptability across languages and tasks. As the study develops, recognition of the construction variances becomes more apparent. This extraordinary discovery represents a perfect opportunity to redesign them in a completely different way, perhaps leading to a detailed investigation of evolutionary tendencies in their habitats. Scientists are already preparing for travels to learn more about this incredible finding.

## 6 CONCLUSION

In this work, we introduced BanglaNewsClassifier, a machine learning method for classifying articles from Bangla newspapers. Sections of Bengali news on various subjects were collected from various websites. We used a machine learning pipeline that included preprocessing the data, extracting features, training the model, and testing. We tested several feature representations and machine learning approaches, ranging from complex classification to standard classification. Our study showed how well BanglaNewsClassifier performed in accurately detecting Bengali news stories for different topics while meeting comparable performance criteria. BanglaNewsClassifier is an important step toward automated news classification in Bengali, advancing natural language processing technologies for underdeveloped languages like Bengali. We utilized error analysis to identify impediments and restrictions, which led to ideas for future enhancements. Our findings pave the way for future research and development efforts to increase the efficiency and accuracy of news categorization systems optimized for the Bengali language domain.

## REFERENCES

[1] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J. (2019, October 23). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv.org. https://arxiv.org/abs/1910.10683v4

[2] Kim, Y. (2014, August 25). Convolutional Neural Networks for Sentence Classification. arXiv.org. https://arxiv.org/abs/1408.5882

[3] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org. https://arxiv.org/abs/1810.04805

[4] Kim, S. B., Seo, H. C., Rim, H. C. (2003, January 1). Poisson naive Bayes for text classification with feature weighting. https://doi.org/10.3115/1118935.1118940

[5] Battogtokh, M., Xing, Y., Davidescu, C., Abdul-Rahman, A., Luck, M., Borgo, R. (2024, March 21). Visual Analytics for Fine-grained Text Classification Models and Datasets. arXiv.org. https://arxiv.org/abs/2403.15492

[6] Jiaramaneepinit, B., Chay-Intr, T., Funakoshi, K., Okumura, M. (2024, March 1). Extreme Fine-tuning: A Novel and Fast Fine-tuning Approach for Text Classification. ACL Anthology. https://aclanthology.org/2024.eacl-short.32/

[7] Villa-Cueva, E., López-Monroy, A. P., Sánchez-Vega, F., Solorio, T. (2024, April 3). Adaptive Cross-lingual Text Classification through In-Context One-Shot Demonstrations. arXiv.org. https://arxiv.org/abs/2404.02452