DeepBlueAI@DravidianLangTech-RANLP 2023

Zhipeng Luo **Jiahui Wang**
DeepBlue Technology (Shanghai) Co., Ltd, Shanghai, China
{luozp, wangjh}@deepblueai.com

Abstract

This paper presents a study on the language understanding of the Dravidian languages. Three specific tasks related to text classification are focused on in this study, including abusive comment detection, sentiment analysis and fake news detection. The paper provides a detailed description of the tasks, including dataset information and task definitions, as well as the model architectures and training details used to tackle them. Finally, the competition results are presented, demonstrating the effectiveness of the proposed approach for handling these challenging NLP tasks in the context of the Dravidian languages.

1 Introduction

The field of natural language processing (NLP) has made significant progress in recent years, with the development of increasingly powerful models and techniques for understanding and analyzing human language. However, one major challenge that remains is the development of NLP systems that can effectively handle regional and under-resourced languages (Chakravarthi and Raja, 2020), which often lack the resources and support needed for effective NLP research. The Dravidian languages (Kolipakam et al., 2018) are one such family of languages that have received relatively little attention in the NLP community, despite their significant cultural and linguistic importance.

To address this gap, this paper focuses on the language understanding of the Dravidian languages, with a particular emphasis on five specific tasks related to text classification. These tasks include abusive comment detection, sentiment analysis and fake news detection.

As the Transformer model (Vaswani et al., 2017) gained popularity, various pretrained models have been proposed, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and the cross lingual pretrained model XLM-RoBERTa (Conneau et al., 2020), etc. Meanwhile, finetuning of pretrained language models has gradually become the standard approach of various natural language understanding tasks, including text classification, as demonstrated in this paper.

In this paper, we provide a detailed description of these tasks, including the dataset information and task definitions, as well as the model architectures and training details used to tackle them. We also present our competition results, demonstrating the effectiveness of our approach for handling these challenging NLP tasks in the context of the Dravidian languages.

2 Task Description

In this section, we describe the task definition and dataset of the 3 tasks we participated in.

2.1 Abusive Comment Detection in Tamil and Telugu

This task requires to indentify whether the given text is an abusive comment. Table 1 describes the dataset (Priyadharshini et al., 2022) information. The dataset contains 7429 samples in total, each sample contains a comment and a label that represents whether it is an abusive comment. The dataset includes 8 label types in total, including misandry, counter-speech, misogyny, xenophobia, hope-speech, homophobia, transphobic, and none-of-the-above.

2.2 Sentiment Analysis in Tamil and Tulu

Given a Youtube comment, this task requires to classify the comment's emotions. Table 2 & 3 describe the provided datasets (Chakravarthi and Raja, 2020) (Hegde et al., 2022) of Tamil and Tulu languages respectively. Both datasets contain Dravidian texts mixed with English collected from social media. The two datasets contains 37775 and

Third Workshop on Speech and Language Technologies for Dravidian Languages,
pages 171–175, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_023

Table 1: data distribution of Abusive Comment Detection task.

Label	Count	Percentage (%)
None-of-the-above	4632	62.35
Misandry	1048	14.10
Counter-speech	443	5.96
Xenophobia	367	4.94
Hope-Speech	266	3.58
Misogyny	261	3.51
Homophobia	215	2.89
Transphobic	197	2.65

7238 samples respectively and each contains 4 sentiment labels, including positive, neutral, negative and mixed feelings.

Table 2: data distribution of Sentiment Analysis task of Tamil language.

Label	Count	Percentage (%)
positive	22327	59.11
unknown state	6239	16.52
negative	4751	12.58
mixed feelings	4458	11.80

Table 3: data distribution of Sentiment Analysis task of tulu language.

Label	Count	Percentage (%)
positive	3487	48.18
neutral	1921	26.54
negative	736	10.17
mixed feeling	1094	15.11

2.3 Fake News Detection in Dravidian Languages

Given a comment from Youtube, this task requires to indentify whether it comes from original or fake news. Table 4 describes the dataset information. The dataset contains 4072 samples in total, each sample has an input comment, and a corresponding label. Label 0 means it comes from fake news, and label 1 means original news.

3 Models & Training

3.1 Model architecture

In order to ensure consistency and simplicity across the different tasks we participated in, we used the same model architecture for all five tasks. Specifically, we employed the XLM-RoBERTa (Conneau et al., 2020) as the pretrained language model, which has been shown to achieve state-of-the-art performance on a range of natural language understanding tasks.

Table 4: data distribution of Fake News Detection task.

label	count	percentage (%)
original	2067	50.76
fake	2005	49.24

For each task, we extracted the representation of the [CLS] token, which is a special token added to the input sequence by the BERT family of models, and passed it through an additional linear layer. This final layer was used to generate the task-specific predictions, which were then compared to the ground truth labels using the softmax cross entropy loss function.

Adopting a consistent model architecture across all tasks allowed us to focus on the differences in the data and task-specific nuances, rather than spending time optimizing different model architectures for each task. It also enables us to easily analyze the generalizability of our approach.

3.2 Adversarial Training

Adversarial training (Miyato et al., 2015; Goodfellow et al., 2014; Miyato et al., 2016) is a technique used in machine learning to improve the robustness of models against adversarial attacks. Adversarial attacks are inputs that are specifically crafted to deceive the model and cause it to misclassify or produce incorrect outputs. Adversarial training involves training the model on both normal and adversarial examples, with the goal of making the model more resistant to adversarial attacks.

Word embeddings are an important component of Transformer models, but they can be vulnerable to overfitting and instability. To address these issues, Miyato et al. (2016) have proposed adding perturbations to the embedding layer during training. This technique, known as Fast Gradient Method (FGM), has been shown to improve the stability and generalization of word representations, leading to better performance on unseen data. By introducing small random perturbations to the embeddings, the model learns to be more robust to variations in the input data and can better capture the underlying semantic and syntactic relationships

between words.

3.3 Ensembling

Ensembling is a widely-used technique in machine learning competitions, and k-fold cross-validation is a common approach used during training to evaluate and improve model performance.

In k-fold cross-validation, the dataset is randomly split into k parts, with one part used as the validation set and the remaining parts used for training. This process is repeated k times, with each part used once for validation. During testing, the model predictions for the test set are obtained from all k models trained in the cross-validation process, and the predicted label for each sample is determined by selecting the most common label among the k predictions. This ensemble approach has been shown to improve model accuracy and generalization, and is widely used in various deep learning applications.

4 Experiments

This section provides a detailed account of the training settings used in our experiments.

4.1 Abusive Comment Detection in Tamil and Telugu

The task is composed of three subtasks, each with its own unique challenges. To effectively address these challenges, we employed a variety of strategies and techniques throughout our work.

For the Tamil and Tamil-English subtasks, we combined the training data and trained our models based on the F1 score on the validation set. By using this approach, we were able to develop models that performed well on both tasks and effectively leverage the available data. For the Telugu subtask, we solely used the training data provided for the task, as the availability of data for this subtask was more limited.

In addition to our approach to training data, we employed a number of techniques to improve the robustness and generalizability of our models. Specifically, we utilized 10-fold cross-validation for all subtasks, ensuring that our models were validated on unseen data and able to generalize well to new data. Additionally, we employed ensembling techniques to combine the strengths of multiple models and achieve superior performance. These techniques helped to mitigate overfitting and ensure that our models were robust and accurate.

4.2 Sentiment Analysis in Tamil and Tulu

For the sentiment analysis task in Tamil and Tulu, we investigated different training approaches. Initially, we attempted to train 10 models using 10-fold cross validation by merging the two datasets together, resulting in an F1 score of 51.4. Subsequently, we treated the Tamil and Tulu datasets as two seperate tasks, each consisting of training 10 models using 10-fold cross-validation. This approach yielded a slightly higher F1 score of 51.7 and allowed us to better tailor our models to the specific characteristics and nuances of each language. To further improve the performance and robustness of our models, we also utilized ensembling techniques.

4.3 Fake News Detection in Dravidian Languages

Since the dataset for this task only consisted of Malayalam language, we faced the challenge of limited data availability. To address this challenge, we adopted a strategy of training multiple models using 10-fold cross-validation. This approach allowed us to effectively leverage the available data and improve the robustness and generalizability of our models.

Specifically, we divided the dataset into 10 subsets and trained 10 models, each using a different subset for validation, while the remaining subsets were used for training. By doing this, we were able to train our models on the entire dataset while also ensuring that our models were validated on unseen data. This approach helped us to overcome the challenge of limited data availability and ensured that our models were able to generalize well to new data.

Overall, our strategy of training multiple models using 10-fold cross-validation proved to be effective in leveraging the limited data available for this task and improving the generalizability and robustness of our models.

5 Competition Results

Our team participated in this competition and achieved promising results, earning 4 first place rankings as well as 1 second place, as shown in Table 5, 6, 7. These outstanding results demonstrate the effectiveness of our innovative methods. We attribute our success to the utilization of various cutting-edge techniques, such as adversarial training, 10-fold cross-validation, and ensembling. Our

use of adversarial training enabled our model to better handle noisy and adversarial inputs, while 10-fold cross-validation helped us to improve the generalizability of our model. Additionally, ensembling multiple models allowed us to combine the strengths of different models and achieve superior performance.These techniques allowed us to develop a robust and accurate model that performed exceptionally well on the competition tasks.

Table 5: F1 Scores and Rankings for Abusive Comment Detection in Tamil and Telugu

Dataset	F1-score (macro)	Rank
Tamil	0.26	7
Tamil-English	0.55	1
Telugu-English	0.7318	2

Table 6: F1 Scores and Rankings for Sentiment Analysis in Tamil and Tulu

Dataset	F1-score (macro)	Rank
Tamil	0.32	1
Tulu	0.542	1

Table 7: F1 Scores and Rankings for Fake News Detection in Dravidian Languages

Dataset	F1-score (macro)	Rank
Malayalam	0.9	1

6 Conclusion

To summarize, this paper focused on five specific text classification tasks related to the Dravidian languages, including abusive comment detection, sentiment analysis and fake news detection.

With the increasing popularity of Transformer-based models such as BERT, RoBERTa, and XLM-RoBERTa, fine-tuning of pre-trained language models has become a standard approach to various natural language understanding tasks.

The paper provided detailed descriptions of the tasks, dataset information and definitions, as well as the model architectures and training details. Our team achieved impressive results in the competition. Our innovative approaches, such as adversarial training, 10-fold cross-validation, and ensembling, played a significant role in our success.

Overall, our findings demonstrate the potential of natural language processing in addressing challenging tasks in the context of the Dravidian languages.

References

Asoka Chakravarthi and Bharathi Raja. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed Tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

Vishnupriya Kolipakam, Fiona M Jordan, Michael Dunn, Simon J Greenhill, Remco Bouckaert, Russell D Gray, and Annemarie Verkerk. 2018. A bayesian phylogenetic study of the dravidian language family. *Royal Society open science*, 5(3):171504.

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. *arXiv preprint arXiv:1907.11692*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. 2015. Distributional smoothing with virtual adversarial training. *arXiv preprint arXiv:1507.00677*.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.