Simon @ DravidianLangTech-EACL2021: Detecting Offensive Content in Kannada Language

Qinyu Que
School of Information Science and Engineering,
Yunnan University, Yunnan, P.R. China
1309487642@qq.com

Abstract

This article introduces the system for the shared task of Offensive Language Identification in Dravidian Languages-EACL 2021. The world's information technology develops at a high speed. People are used to expressing their views and opinions on social media. This leads to a lot of offensive language on social media. As people become more dependent on social media, the detection of offensive language becomes more and more necessary. This shared task is in three languages: Tamil, Malayalam, and Kannada. Our team takes part in the Kannada language task. To accomplish the task, we use the XLM-Roberta model for pre-training. But the capabilities of the XLM-Roberta model do not satisfy us in terms of statement information collection. So we made some tweaks to the output of this model. In this paper, we describe the models and experiments for accomplishing the task of the Kannada language.

1 Introduction

The network platform builds a brand new living and cultural space and promotes communication among netizens. This leads to an exponential growth of information and speech in cyberspace. The popularity of the Internet facilitates the spread of offensive remarks, and also brings some negative social effects (Chakravarthi et al., 2020c). The people who communicate on the Internet come from different countries (Thavareesan and Mahesan, 2019, 2020a,b). Due to the differences in history and culture of different countries, people in different regions have great differences in their understanding of Offensive speech (Mandl et al., 2020). So it's easy for people to inadvertently make offensive remarks about other people. Offensive language is not a recent phenomenon, but its impact is growing because of its rapid spread on social media. A large amount of offensive language on the Internet can cause serious social problems. Therefore, detection of offensive content is very necessary. This is not only conducive to purifying the network environment but also conducive to promoting the positive development of society (Chakravarthi, 2020a; Chakravarthi and Muralidaran, 2021).

Kannada is a Dravidian language spoken mainly by Karnataka residents in India's south western region (Chakravarthi, 2020b). After AD 600, Kannada developed from the old Tamil script. Tamil has the most ancient Indian non-Sanskritic literature of Indian languages. The Kappe Arabhatta record from AD 700 is the oldest extant record of Kannada poetry in Tripadi meter. The data set used in the experiment is provided by the competition organization(Chakravarthi et al., 2021, 2020a,b; Hande et al., 2020). The data was collected from various social media platforms. In the following content, we introduce recent relevant work, the models we use, our experimental processes, and results.

2 Related Work

Online social media is now one of the main channels of communication. But the growth of offensive content is troubling users and the companies that run social media. The need to identify offensive content has increased dramatically. The purpose of the identification of offensive content is to reduce offensive content in social media and thus improve users' communication experience on social media. Some scholars have done a lot of work on the classification of offensive language. Offensive language attacks people for a variety of reasons. People can be targeted with offensive language because of their gender, skin color, or nationality(Reddy and Vasu, 2002; Gatehouse et al., 2017; Erjavec and Kovali, 2012). In the field of detecting offen-

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 160–163

Label	Train set	validation set
N_o	57.01%	54.38%
O_T_I_O	1.98%	2.06%
O_T_I_I	7.83%	8.49%
O_T_I_G	5.29%	5.79%
n_K	24.48%	24.58%
O_U	3.41%	4.25%

Table 1: Label distribution of Kannada language subtask

sive language, many tasks related to it have been completed by scholars. These tasks include the detection of bullying language(Van Hee et al., 2015) and hate language(Davidson et al., 2017) on social platforms. People use various methods to achieve the classification of text. LSTM(Surhone et al., 2010) was proposed by Hochreiter et al. LSTM plays a great role in the field of natural language processing.

3 Methodology and Corpus

3.1 Data Description

In the Kannada language subtask, the data set contains six tags:

- Not_offensive(N_o): This tag indicates that the comment is Not offensive.
- Offensive_Targeted_Insult_Other(O_T_I_O): This tag indicates that a comment is an offensive text which is neither targetted towards an individual or group.
- Offensive_Targeted_Insult_Individual(O_T_I_I): This tag indicates that a comment is offensive to a person.
- Offensive_Targeted_Insult_Group(O_T_I_G): This tag indicates that a comment is offensive to a group.
- not_Kannada(n_K): This tag indicates that this comment is not in the Kannada language. Item Offensive_Untargetede(O_U): The tag indicates that the comment is offensive, but does not have a specific target.

We can see that this is a six-category classification task. The data set for this task is classified in a very detailed way. This also makes Classification difficult. Table 1 gives the details of the data set.

3.2 Data Preparation

Before training the data set, we preprocessed the text. The main purpose of preprocessing is to standardize the text collected from social media and reduce unnecessary words in the text. Here's what we do to the text:

- Removing punctuation marks and special characters
- Removing the URL
- Removing the emoticons

3.3 Model Description

In the experiment, we used XLM-Roberta(Conneau et al., 2020) as our preprocessing model. XLM-Roberta was released by Facebook in 2019. XLM-Roberta is an improvement based on BERT(Devlin et al., 2019). The researchers found that Bert was undertrained. So they made these improvements to XLM-Roberta on the basis of Bert: training the sequence with relatively large length, deleting the prediction for the next sentence, using dynamic changes in the masking model of training data, and increasing the training capacity of the model. XLM-Roberta is a multilingual model capable of handling text in 100 different languages. In the experiment, we are not satisfied with the statement information collection ability of the XLM-Roberta model. So our team came up with the model shown in Figure 1. This model mainly combines the last three hidden layers' outputs of XLM-Roberta with LSTM. We input the last three hidden layers of XLM-Roberta into the LSTM to get the output of the LSTM. Finally, the output of LSTM and the Pooler output of XLM-Roberta(P_O) are connected together and put into the classifier.

4 Experiment and Results

We combined the training set and validation set provided by the contest organizer, and then used Stratified K-fold Cross Validation to get the new training set and validation set. Cross-validation divides the data set multiple times. The common method of randomly dividing the data set will make the result have a contingency. Cross-validation reduces the chance that comes with a random partition. The model can complete the training of various types of data so as to improve the generalization ability of the model. The scale relationship of the Label in each layer of data is saved.

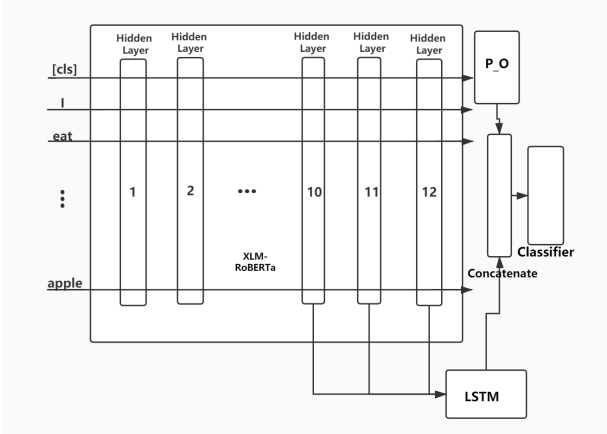

Figure 1: Model description: custom XLM-Roberta architecture that acts as a feature extractor to provide useful information from the given text to LSTM.

Hyper-parameter	Value
dropout	0.5
learning rate	2e-5
epoch	4
per gpu train batch size	32
gradient accumulation steps	8

Table 2: The hyper-parameters

We used XLM-Roberta-Base[1] as our preprocessing model. In the experiment, the hyper-parameters we used are shown in Table 2. The task was evaluated by following the Macro Average F1 of Scikit-itLearn. In the Kannada language sub-task, our final F1-score in the official test set is 0.33.

5 Conclusion

Our ranking in this task is not ideal, and our test scores are not satisfactory. We think there are the following reasons: First, our team did not set the hyperparameters reasonably in the experiment, for example, the value setting of the epoch was small. Second, the classification task has a large number of categories, and the nature of the tags is very similar. That makes the task more difficult. Third, the data set is imbalanced. In the future, we will improve our model to accommodate the data imbalance.

References

Bharathi Raja Chakravarthi. 2020a. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

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

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan V, Elizabeth Sherly, and John Philip McCrae. 2021. Findings of the shared task on Offensive Language Identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop*

[1] https://huggingface.co/xlm-roberta-base

on Speech and Language Technologies for Dravidian Languages. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. 2020c. Overview of the Track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 21–24, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

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.

Karmen Erjavec and Melita Poler Kovali. 2012. "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication Society*, 15(6):899–920.

Cally Gatehouse, Matthew Wood, Jo Briggs, James Pickles, and Shaun Lawson. 2017. Troubling vulnerability: Designing with lgbt young people's ambivalence towards hate crime reporting. In *Conference on Human Factors in Computing Systems*.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63, Barcelona, Spain (Online). Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 29–32, New York, NY, USA. Association for Computing Machinery.

Reddy and Vasu. 2002. Perverts and sodomites: homophobia as hate speech in africa. *Southern African Linguistics Applied Language Studies*, 20(3):163–175.

Lambert M. Surhone, Mariam T. Tennoe, and Susan F. Henssonow. 2010. Long short term memory. *Betascript Publishing*.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based Part of Speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.

Cynthia Van Hee, Ben Verhoeven, Els Lefever, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Guidelines for the fine-grained analysis of cyberbullying, version 1.0.