Simon @ DravidianLangTech-EACL2021: Meme Classification for Tamil with BERT

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

Abstract

In this paper, we introduce the system for the task of meme classification for Tamil, submitted by our team. In today's society, social media has become an important platform for people to communicate. We use social media to share information about ourselves and express our views on things. It has gradually developed a unique form of emotional expression on social media – meme. The meme is an expression that is often ironic. This also gives the meme a unique sense of humor. But it's not just positive content on social media. There's also a lot of offensive content. Meme's unique expression makes it often used by some users to post offensive content. Therefore, it is very urgent to detect the offensive content of the meme. Our team uses the natural language processing method to classify the offensive content of the meme. Our team combines the BERT model with the CNN to improve the model's ability to collect statement information. Finally, the F1-score of our team in the official test set is 0.49, and our method ranks 5th.

1 Introduction

Everything has two sides, and information technology is no exception. On the one hand, the rapid development of information technology makes it easy for people to communicate and discuss on social platforms (Thavareesan and Mahesan, 2019, 2020a,b). On the other hand, the rapid development of information technology has also led to the proliferation of offensive content on the Internet (Chakravarthi et al., 2021; Chakravarthi and Muralidaran, 2021; Chakravarthi et al., 2020a; Mandl et al., 2020). Meme makes people like it because of its unique sense of humor. This also allows meme to spread rapidly on social media. But these memes don't just contain positive content. Some memes also contain offensive content. With the rapid growth of offensive content, hateful content, and offensive content on the Internet, the detection of offensive content on Meme has become more and more important. The meme is special. Different people have different perceptions of its content. Some memes are only accessible to a certain group of people, and others don't get the true meaning of these memes. This also makes people unclear about the offensive content in Meme, and also increases the difficulty of detecting the offensive content in Meme. The meme usually consists of pictures and text. Our team uses a natural language processing method to detect offensive content in the meme. We do the classification for Memes in Tamil language.

Tamil is a Dravidian language spoken by the Tamil people of South Asia as their first language (Chakravarthi et al., 2018). For over 2600 years, Tamil literature has been recorded. Tamil has the oldest non-Sanskritic Indian literature in any Indian language. Tamil was the first Indian classical language to be listed, and it is one of the world's oldest classical languages. There are 12 vowels, 18 consonants, and one special character, the aytam, in the present Tamil script. The vowels and consonants merge to form 216 compound characters, for a total of 247 (12 + 18 + 1 + (12 x 18) characters (Chakravarthi et al., 2020b). In Tamil, suffixes are used to denote noun class, number, and case, as well as verb tense and other grammatical categories. In comparison to Sanskrit, which is the standard for most Aryan languages, Tamil's standard metalinguistic terminology and scholarly vocabulary is Tamil. Tamil is under-resourced languages (Chakravarthi, 2020).

2 Related Work

Many researchers have made efforts to detect offensive language. As people communicate more and

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 287–290
April 20, 2021 ©2021 Association for Computational Linguistics

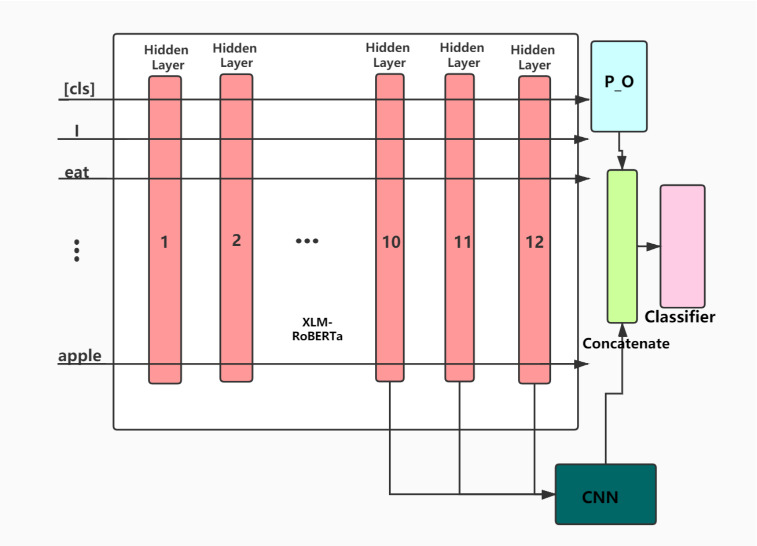

Figure 1: The description of the model used for this task

more online, the detection of hateful content and offensive language has become necessary. Content published in traditional media such as newspapers and television is heavily censored. But people are often free to express their opinions on the new social media platforms. Troll memes disrupt the user's communication experience on these platforms. Nascimento et al. (Reddy and Vasu, 2002) believe that people can communicate without face to face in social networks. So people don't know each other's point of view and background. People abuse anonymous nature of the internet by using offensive language. This greatly accelerates the spread of offensive language. Nascimento et al. classified the Brazilian grape language text into hate languages. They used data sets from the Brazilian 55chan Imageboard. They used three classifiers for classification, with the best F1-score being 0.955. The meme is a regional language, and the expression in communication is that only some people can know the specific meaning of a certain meme. This characteristic of meme makes its offensive language have a specific target group. There has been an increase in Islamophobic comments on social media platforms. It ostracized Muslims in the community. B. Vidgen' team built a model that could distinguish between Islamophobia (Ross et al., 2017). This model can divide content into strongly Islamophobic content, weakly Islamophobic content, and non-Islamophobic content. The targeted groups are not only Muslims, but also immigrant groups and Jewish groups (Vidgen and Yasseri, 2020; Bilewicz et al., 2013; Finkelstein et al., 2018; Djuric et al., 2015). In recent years, developments in the field of detecting offensive content have contributed to many problems (Bohra et al., 2018; Jurgens et al., 2019; Caselli et al., 2020; Fortuna et al., 2019).

Label	Train set	Test set
troll	1282	395
Not_troll	1018	272

Table 1: Class distribution for TamilMemes dataset

3 Methodology and Corpus

3.1 Data Description

Competition organizers provides dataset with text and image as features(Suryawanshi et al., 2020; Suryawanshi and Chakravarthi, 2021). Because our team uses natural language processing to identify offensive content. We used the text portion of the dataset. The text portion of the dataset contains two types of tags, one troll, and the other Not_troll. The tag troll represents a meme that has provocative or offensive content. The Not_troll tag indicates that the meme has no offensive or provocative content. The detailed contents of the dataset are shown in Table 1.

3.2 Model Description

We reviewed the results of recent competitions and found that participants who used pre-trained language models scored highly on many of the tasks. For example, in the Hate Speech and Offensive Content Identification in Indo-European

Hyper-parameter	**Value**
dropout	0.5
learning rate	1e-5
epoch	10
per gpu train batch size	32
gradient accumulation steps	8

Table 2: The hyper-parameters settings used to train the XLM-Roberta on the TamilMemes dataset

Languages(HASOC[1]) competition, many contestants using the BERT model rank high. And in the competition, many players adopted the fine-tuned BERT(Devlin et al., 2019) model method to complete the task. In our model, we use XLM-Roberta(Conneau et al., 2020) as our preprocessing model. The XLM-Roberta model is a multilingual processing model. It can handle text in 100 languages. Compared with the BERT model, the XLM-Roberta model greatly increases the amount of data to be trained. XLM-Roberta has no prediction for the next sentence. Our team fine-tuned the superstructure of the XLM-Roberta model to enhance the semantic collection capability of XLM-Roberta. The specific model is shown in Figure 1. In the first step, we get the pooler output of XLM-Roberta (P_O). For classification tasks, the output of XLM-Roberta(pooler output) is obtained by its last layer hidden state of the first token of the sequence (CLS token) further processed by a linear layer and a Tanh activation function. Second, we combine the output of the last three hidden layers of XLM-Roberta with CNN (Simonyan and Zisserman, 2014). Finally, we concatenate the CNN output with the pooler output of XLM-Roberta and put it into the classifier.

4 Experiment and Results

Our team uses the stratified 5-fold cross-validation to divide the training sets provided by the competition organizer into new training sets and validation sets. Our pre-training model is from XLM Roberta - Base [2]. We put the hyper-parameters used in the experiment in Table 2. Finally, Our method rankings 5th in this task, and the F1-score is 0.49.

5 Conclusion

This paper describes the model and results used by the Simon team in the meme classification task. This model is a fine-tuned XLM-Roberta model, which improves the ability of the model to obtain sentence information. The datasets used for this competition are slightly unbalanced. This also affects the final result of the model. Our team will continue to improve the model to achieve better results. The code is available online.[3]

References

Michal Bilewicz, Miko?Aj Winiewski, Miros?Aw Kofta, and Adrian Wójcik. 2013. Harmful ideas, the structure and consequences of anti-semitic beliefs in poland. *Political Psychology*, 34(6):821–839.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed S. Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*.

Tommaso Caselli, Valerio Basile, Jelena Mitrovi, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*.

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

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2018. Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global Wordnet Conference*, pages 77–86, Nanyang Technological University (NTU), Singapore. Global Wordnet 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, 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 on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

[1] https://competitions.codalab.org/competitions/26027

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

[3] https://https://github.com/queqinyu/EACL

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. 2020a. 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.

Bharathi Raja Chakravarthi, Navaneethan Rajasekaran, Mihael Arcan, Kevin McGuinness, Noel E. O'Connor, and John P. McCrae. 2020b. Bilingual lexicon induction across orthographically-distinct under-resourced Dravidian languages. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 57–69, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

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*.

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.

Nemanja Djuric, Zhou Jing, Robin Morris, Mihajlo Grbovic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *the 24th International Conference*.

Joel Finkelstein, Savvas Zannettou, Barry Bradlyn, and Jeremy Blackburn. 2018. A quantitative approach to understanding online antisemitism.

Paula Fortuna, Joo Rocha Da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*.

David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. 2019. A just and comprehensive strategy for using nlp to address online abuse.

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.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *Computer Science*.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on Troll Meme Classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

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.

Bertie Vidgen and Taha Yasseri. 2020. Detecting weak and strong islamophobic hate speech on social media. *Journal of E-Government*, 17(1):66–78.