HUB@DravidianLangTech-EACL2021: Meme Classification for Tamil Text-Image Fusion

Bo Huang
School of Information Science
and Engineering Yunnan University,
Yunnan, P.R. China
hublucashb@gmail.com

Yang Bai
School of Information Science
and Engineering Yunnan University,
Yunnan, P.R. China
baiyang.top@gmail.com

Abstract

This article describes our system for task DravidianLangTech - EACL2021: Meme classification for Tamil. In recent years, we have witnessed the rapid development of the Internet and social media. Compared with traditional TV and radio media platforms, there are not so many restrictions on the use of online social media for individuals and many functions of online social media platforms are free. Based on this feature of social media, it is difficult for people's posts/comments on social media to be strictly and effectively controlled like TV and radio content. Therefore, the detection of negative information in social media has attracted attention from academic and industrial fields in recent years. The task of classifying memes is also driven by offensive posts/comments prevalent on social media. The data of the meme classification task is the fusion data of text and image information. To identify the content expressed by the meme, we develop a system that combines BiGRU and CNN. It can fuse visual features and text features to achieve the purpose of using multi-modal information from memetic data. In this article, we discuss our methods, models, experiments, and results.

1 Introduction and Background

The meme is a kind of multimedia document based on image level, which is a picture containing text. Netizens can express their feelings, opinions, interests, and so on through it (Yus, 2018a). The meme is an image with some kind of subtitle text embedded in the image pixels. In the past few years, emoticons have become very popular and used in many different contexts, especially young people (Oriol Sabat et al., 2019). However, this form is also used to produce and spread hate speech in the form of black humor. The meme is a term proposed by biologist Richard Dawkins. It was used to describe the flow, flow, mutation, and evolution of culture and is a means of cultural resistance to genes (Milner, 2012). But the meaning of this term has changed in our public life. In social media, Meme is a kind of self-media work, which is produced and spread by the majority of Internet users through social media networks (Castaño Díaz, 2013). Negative information detection on the Internet has become a core social challenge. Nowadays, the detection of negative information in social media has become more and more intelligent (Chatzakou et al., 2017; Pfeffer et al., 2014).

However, due to the multimodal nature of memes, it is difficult to be intelligently identified whether it is aggressive (Suryawanshi et al., 2020a). Especially in India, some recent memes containing hate messages have threatened people's lives several times (Suryawanshi et al., 2020b). Due to the large population and the mixing of multiple languages, a large number of Indian memes are difficult to monitor due to the lack of intelligent systems for specific languages. This adds another serious challenge to the problem of memetic classification. In this work, we participated in the shared task of memetic classification in Tamil (Suryawanshi and Chakravarthi, 2021). The Indian state of Tamil Nadu, as well as two independent countries, Singapore and Sri Lanka, speak Tamil as their official language. Tamil was the first Indian classical language to be listed as such, and it is still one of the world's oldest classical languages. Dravidian civilisations are believed to have flourished in the Indus Valley civilization (3,300–1,900 BCE), which was situated in the Northwestern Indian subcontinent, this period is considered as second Sangam period in Tamil. Tamil is India's oldest language. Tamil, Pali, and Prakrit all added words, texts, and grammar to Sanskrit. We are committed to making our contribution to the identification of Tamil memetic classification.

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 210–215
April 20, 2021 ©2021 Association for Computational Linguistics

2 Related Work

In theory, a meme is an example of a large number of humorous words on the Internet, copied or modified, and then spread to other users. But sometimes some information in Internet memes can bring a negative impact (Yus, 2018b). Compared with the propagation speed of memes on the Internet, the research progress of memes is much slower. The work of Wang et al. showed us that the combination of text and vision can help identify popular memetic descriptions (Wang and Wen, 2015). The automatic meme generation system implemented by Vyalla et al. based on the transformer model can allow users to generate memes of their choice (Vyalla and Udandarao, 2020). Compared with the text field, there is less research on sentiment analysis on memetic data. In recent years, deep learning technology has attracted the attention of researchers, especially in sentiment analysis tasks that are significantly better than traditional methods. Zhang et al. used deep learning techniques for sentiment analysis on text datasets (Zhang et al., 2018) and Poria et al. used deep learning models for sentiment analysis on image and video datasets (Poria et al., 2017a). The work of Sabat found that visual modalities provide more information in hate memetic detection than language modalities (Sabat et al., 2019). Suryawanshi and others shared with us their data sets and methods for detecting offensive content in multimodal memetic data sets (Suryawanshi et al., 2020a). Kumar et al. (Kumar et al., 2020) used a multimodal approach to determine the sentiment of a meme.

The method of using both text and visual information to improve performance has also proven to be effective (Guillaumin et al., 2010; Zahavy et al., 2016). Generally speaking, the fusion strategy of text information and visual information can be divided into two methods (Corchs et al., 2019). On the one hand, some models use feature-level fusion methods, where text input sources or image input sources are processed to extract a set of features. Then, the feature set is put together for the final decision (Atrey et al., 2007; Wang et al., 2018). On the other hand, other models perform complete processing of each input source and perform fusion at the decision-making level (Atreya V et al., 2013; Poria et al., 2017b). The result we submitted was predicted using the second strategy.

3 Data And Methods

3.1 Data Description and Analysis

The training set provided by the task organizer is mainly divided into two types of data, one is a text data set, and the other is a picture data set. There are 2300 pieces of text data in the text data set. The proportion of "Not Troll" text data and "Troll" text data to the total data volume are 44% and 56%, respectively. The text content of some of the text data is "No Captions". This situation means that no text annotations in the Tamil language appear on the meme picture corresponding to this text data. Such meme pictures without text annotations are not many in the dataset. The other text data is the text appearing on the corresponding meme picture. We present the text data set in the training set provided by the task organizer in a word cloud diagram. It is not difficult to see that in these text data, the most frequent words are mainly demonstrative pronouns and some modal particles.

The work of Suryawanshi et al. on the task data set shows us the source of the training set of 2969 meme images that we used in the training phase (Suryawanshi et al., 2020b; Suryawanshi and Chakravarthi, 2021). The proportions of the number of pictures marked as "Not Troll" and the number of pictures marked as "Troll" in the total number of meme picture datasets are 0.66 and 0.34 respectively. Regardless of whether there are texts in the meme pictures, the content of these meme pictures can also express the information of "Troll" and "Not Troll". These meme pictures mainly come from many popular social media platforms (such as YouTube, Facebook, WhatsApp, Instagram, etc.). So the size and style of the pictures are also different. From the comparison between the entire text data and the meme picture data, the relationship between them is not one-to-one. It is not difficult to find that their numbers are not equal. This data distribution feature is not good information for us.

3.2 Methods

Combine our analysis and understanding of task description and data set. Our system must process text data while also processing meme image data. Therefore, we choose to use the BiGRU artificial neural network and CNN artificial neural network capable of processing text data as the basic components of our system model. BiGRU can learn the contextual semantic information in the

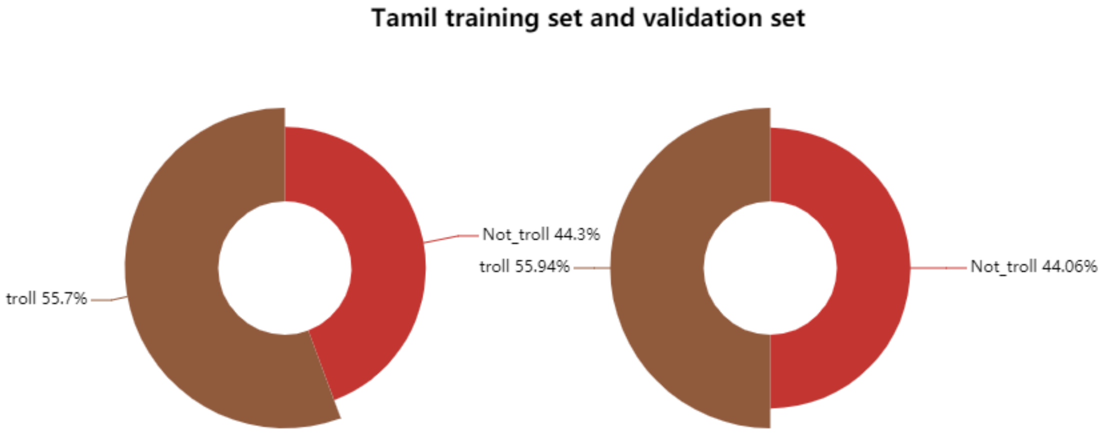


Figure 1: Labels distribution of Tamil training set and validation set. In the training set, Troll: 55.7%, Not troll: 44.3%. In the validation set, Troll: 55.94%, Not troll:44.06%.

Figure 2: No text annotations in the Tamil language appear on the meme picture.

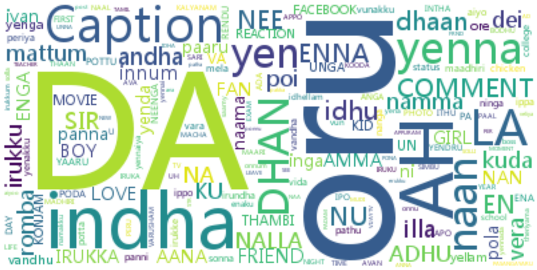


Figure 3: The word cloud image of the text training set is provided by the task organizer. The word "Caption" can not be used as reference information, because it mainly comes from the annotation of the text data, not the text content that appears in the meme picture.

text through the encoded-word vectors. CNN can learn the information in the picture through convolution operation and pooling operation. We use stacking to combine BiGRU and CNN blocks to form the overall architecture of our system.

The CNN block is mainly composed of three two-dimensional convolutional layers and three maximum pooling layers. The size of the convolution kernel is selected as 2. After getting the image processed by the third convolutional layer(Conv2d_2) and the third pooling layer(MaxPooling2D_2), the multi-dimensional tensor is converted into a low-latitude tensor through a straightening(Flatten) operation. Then, the result obtained in the previous step is used as the input of the two dense layers(Dense_0, Dense_1). Finally, the output result of the CNN block is obtained.

Text data We use the Tamil pre-training language vector provided by fasttext[1] to encode the text data (Grave et al., 2018). Then input the encoded text vector into the BiGRU network to obtain the output result of the BiGRU network. Next, input the BiGRU result from the previous step into the Dense layer to get the final output of the BiGRU block. Connect the output result of the CNN block and the output result of the BiGRU block to obtain a new tensor. Finally, input this new tensor into the Dense layer to get the final result of the model. The final effect that our system model needs to achieve is to merge text and image data information. We

[1] https://fasttext.cc/docs/en/crawl-vectors.html

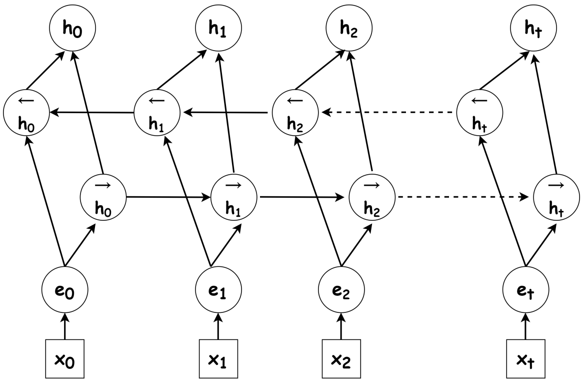
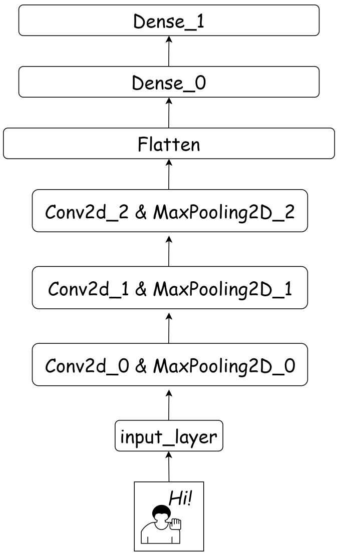


Figure 4: BiGRU structure and data flow.



Dense_1
Dense_0
Flatten
Conv2d_2 & MaxPooling2D_2
Conv2d_1 & MaxPooling2D_1
Conv2d_0 & MaxPooling2D_0
input_layer
Hi!

Figure 5: The CNN block in our system.

provide the code implementation of our system[2].

4 Experiment and Results

4.1 Data Preprocessing

The work in the data preprocessing stage is mainly for text data. We split the text training set released by the task organizer into a new text training set and a text verification set under the premise of ensuring the same data distribution. The text data is divided by spaces to get each word, and then encoded using the fasttext we mentioned in the Methods section. Use the results predicted by the model on the validation set to evaluate our model system and adjust the parameters of the model system. For meme image data, we set their size uniformly to (300, 300).

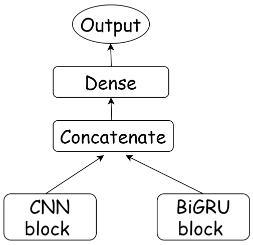


Figure 6: The CNN block, BiGRU block, and Dense layer together constitute the main part of our system.

Language	F1 Score	Precision	Recall
Top1 results	0.55	0.57	0.6
Our results	0.4	0.5	0.54
Validation set	0.50	0.52	0.56

Table 1: The result score of the top1 team on the test set. We submit the test set prediction result score. The score of our system on the validation set.

4.2 Experiment setting

In our experiment, the optimizer uniformly uses Adam optimizer. Because Adam uses momentum and adaptive learning rate to speed up convergence. The number of BiGRU layers is set to 2, and the word embedding vector uses 300 dimensions. epoch, learning rate, and batch seize are set to 10, 0.003, and 32 respectively. The activation function used between the three layers in the CNN block is Relu. The activation function used by the dense layer in BiGRU is Softmax. Connect the results in the CNN block and the BiGRU block as the input of the classifier (Dense layer). Use the classifier to get the output result. Figure 6 shows the architecture of our system.

4.3 Analysis of Results

The leaderboard results announced by the task organizer are ranked using the weighted average F1 score. At the same time, the Precision and Recall scores of all participants' submission results will be announced on the leaderboard. Table 1 shows the final result of our system model on the test set, the result score of the top1 team's system on the test set, and the result score of our model on the validation set. Comparing the scores in the table, the results of my system on the validation set are quite different from the results on the test set. There is also a gap between the result score of my system on the test set and the top1 result score. Our team solution ranked 9th in the final leaderboard. The number of data sets we can use is not large, and we

[2] https://github.com/Hub-Lucas/hub-at-meme

randomly select a part of them as the verification set. The result of this is that there are fewer data sets that can be used to train the model. There is no restriction on overfitting in our model. These are the shortcomings of our system.

5 Conclusion

This article introduces the method and model system used by our team in the Meme classification for the Tamil shared task. We use a text and image fusion scheme to detect memetic categories in the Tamil language environment. We analyzed the deficiencies of our system. These shortcomings are what we need to improve in our future work. We also have a lot of room for improvement in methods and systems. For example, in image processing, we can try some other network models such as MobileNet (Howard et al., 2017), ResNext(He et al., 2016), etc. Some pre-trained language models in text processing.

Considering that meme pictures can quickly spread on Internet social media and can express negative information across languages, we determined that dealing with similar issues in social media is very valuable and meaningful. Similar issues that exist in social media in some small-language communities should also deserve our attention and study. In addition to improving our models and methods in future work, we will continue to pay attention to the research and development of memetic analysis.

References

Pradeep K Atrey, Mohan S Kankanhalli, and John B Oommen. 2007. Goal-oriented optimal subset selection of correlated multimedia streams. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 3(1):2–es.

Arjun Atreya V, Yogesh Kakde, Pushpak Bhattacharyya, and Ganesh Ramakrishnan. 2013. Structure cognizant pseudo relevance feedback. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 982–986, Nagoya, Japan. Asian Federation of Natural Language Processing.

Carlos Mauricio Castaño Díaz. 2013. Defining and characterizing the concept of internet meme. *Ces Psicología*, 6(2):82–104.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22.

Silvia Corchs, Elisabetta Fersini, and Francesca Gasparini. 2019. Ensemble learning on visual and textual data for social image emotion classification. *International Journal of Machine Learning and Cybernetics*, 10(8):2057–2070.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. 2010. Multimodal semi-supervised learning for image classification. In *2010 IEEE Computer society conference on computer vision and pattern recognition*, pages 902–909. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Akshi Kumar, Kathiravan Srinivasan, Wen-Huang Cheng, and Albert Y Zomaya. 2020. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing & Management*, 57(1):102141.

Ryan M Milner. 2012. *The world made meme: Discourse and identity in participatory media*. Ph.D. thesis, University of Kansas.

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *arXiv e-prints*, pages arXiv–1910.

Jürgen Pfeffer, Thomas Zorbach, and Kathleen M Carley. 2014. Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications*, 20(1-2):117–128.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017a. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017b. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *arXiv preprint arXiv:1910.02334*.

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, Mihael Arcan, and Paul Buitelaar. 2020a. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020b. 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).

Suryatej Reddy Vyalla and Vishaal Udandarao. 2020. Memeify: A large-scale meme generation system. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 307–311.

William Yang Wang and Miaomiao Wen. 2015. I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 355–365.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.

Francisco Yus. 2018a. Identity-related issues in meme communication. *Internet Pragmatics*, 1(1):113–133.

Francisco Yus. 2018b. Identity-related issues in meme communication. *Internet Pragmatics*, 1(1):113–133.

Tom Zahavy, Alessandro Magnani, Abhinandan Krishnan, and Shie Mannor. 2016. Is a picture worth a thousand words? a deep multi-modal fusion architecture for product classification in e-commerce. *arXiv preprint arXiv:1611.09534*.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.