TRAC-1 Shared Task on Aggression Identification: IIT(ISM)@COLING'18

Ritesh Kumar
Department of CSE
IIT(ISM) Dhanbad
India, 826004
ritesh4rmrvs@gmail.com

Guggilla Bhanodai
Department of CSE
IIT(ISM) Dhanbad
India, 826004
bhanodaig@gmail.com

Rajendra Pamula
Department of CSE
IIT(ISM) Dhanbad
India, 826004
rajendra@iitism.ac.in

M. R. Chennuru
Department of CSE
IIT(ISM) Dhanbad
India, 826004
cmr.mahesh@gmail.com

Abstract

This paper describes the work that our team **bhanodaig** did at Indian Institute of Technology (ISM) towards TRAC-1 Shared Task on Aggression Identification in Social Media for COLING 2018. In this paper we label aggression identification into three categories: Overtly Aggressive, Covertly Aggressive and Non-aggressive. We train a model to differentiate between these categories and then analyze the results in order to better understand how we can distinguish between them. We participated in two different tasks named as *English (Facebook) task* and *English (Social Media) task*. For *English (Facebook) task* **System 05** was our best run (i.e. 0.3572) above the random Baseline (i.e. 0.3535). For *English (Social Media) task* our **system 02** got the value (i.e. 0.1960) below the Random Bseline (i.e. 0.3477). For all of our runs we used Long Short-Term Memory model. Overall, our performance is not satisfactory. However, as new entrant to the field, our scores are encouraging enough to work for better results in future.

1 Introduction

In the last few years, there is exponential growth in social media and user generated contents. The online platforms like blogs, Q&A forums, online discussion forum and so on helps user to post their comments and to reply other user's comment. These comments may be of various forms like lovable, aggressive, hate speech, offensive languages etc. As the number of people and this interaction over the web has increased, incidents of aggression and related activities like trolling, cyberbullying, flaming, hate speech, etc. have also increased manifold across the globe. Thus, incidents of online aggressive behaviour have become a major source of social conflict, with a potential of forming criminal activity.

A key challenge for aggression identification on social media is to classify it from offensive or vitriolic languages. Some of the task has been performed in this area but still it is a hot topic among researchers. Keeping it in mind, we develop a system to discriminate between Overtly Aggressive (OAG), Covertly Aggressive (CAG), and Non-aggressive (NAG) content in texts. In the paper (Kumar et al., 2018a) the relevant task description is defined in detail.

In this paper, we used the fact that emojis serve as a proxy for the emotional contents of a text. Therefore, pre-training on the classification task of predicting which emoji were initially part of a text can improve performance on the target task. We used word embeddings that were trained on the classification task. Then we used simple transfer learning approach, *chain-thaw* that sequentially unfreezes and fine-tunes a single layer at a time. This approach increases accuracy on the target task at the expense of extra computational power needed for the fine-tuning. By training each layer separately the model is able to adjust the individual patterns across the network with a reduced risk of overfitting.

Organization of rest of the paper is as follows. We describe Related Work in Section 2. Section 3 describes Methodology and Dataset and Section 4 analyse our Results. Finally, we conclude in Section 5 with directions for future work.

Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pages 58–65
Santa Fe, USA, August 25, 2018.

2 Related Work

The interest in identifying trolling, aggression, cyber-bullying and hate speech, particularly on social media, has been growing in recent years. This topic has attracted attention from researchers interested in linguistic and sociological features of aggression, and from engineers interested in developing tools to deal with aggression on social media platforms. In this section we review a number of studies and briefly discuss their findings. For a recent and more comprehensive survey on hate speech detection we recommend (Schmidt and Wiegand, 2017).

Prior work has used theories of emotion such as Ekmans six basic emotions and Plutchiks eight basic emotions(Suttles and Ide, 2013) and (Mohammad, 2012). The heterogeneous NLP tasks are bounded by dearth of manually annotated data. Therefore, emotional expression plays significant role to gauge the mood of users. (Deriu et al., 2016) and (Tang et al., 2014) tried to see the effect of positive/negative emoticons for training their models. Similarly, (Mohammad, 2012) mapped hashtags such as #anger, #joy, #happytweet, #ugh, #yuck and #fml into emotional categories for emotion analysis.

(Davidson et al., 2017) used crowd-sourced hate speech lexicon to collect tweets containing hate speech keywords. They used crowd-sourcing to label a sample of these tweets into three categories: those containing hate speech, only offensive language, and those with neither. They found that Tweets without explicit hate keywords were also more difficult to classify.

(Xu et al., 2012) proposed sentiment analysis to detect bullying roles in tweets. For this they used Latent Dirichlet Allocation to find out relevant topics in bullying texts.They classified the texts either they are bullying or not.

(Dadvar et al., 2013) presented the results of a study on the detection of cyberbullying in YouTube comments. They used a combination of content-based, cyberbullying-specific and user-based features. their results showed that incorporation of context in the form of users activity histories improves cyber-bullying detection accuracy.

(Kwok and Wang, 2013) detected tweets against blacks. They used unigram model and supervised learning for their approach. (Djuric et al., 2015) used binary classification to detect hate speech. For this they used word embeddings that performed better than bag-of-words model.

(Burnap and Williams, 2015) studied cyber hate speech in twitter. They showed that the production of a machine classifier that can be developed into a technical solution for use by policymakers as part of an existing evidence-based decision-making process.

(Nobata et al., 2016) studied the abusive language detection in online user content. They used different syntactic features as well as different embedding features. They found that combining these features with the features of NLP can boost F-score.

(Waseem and Hovy, 2016) presented a list of criteria based in critical race theory to identify racist and sexist slurs. They used n-grams model for their approach. The dataset footnote[1] used for this criteria is freely available for users.

presenting the Hate Speech Detection dataset used in (Malmasi and Zampieri, 2017) and a few other recent papers on the topic. A proposal of typology of abusive language sub-tasks is presented in (Waseem et al., 2017). A recent discussion on the challenges of identifying profanity vs. hate speech can be found in (Malmasi and Zampieri, 2018).

Most of the study,including ours, to discriminate between hate speech and abusive language are in english due to its well annotation. However, few recent studies have also been published in some other languages. Examples are (Mubarak et al., 2017) studied abusive language detection on Arabic Language. (Su et al., 2017) rephrased profanity in Chinese language. (Tulkens et al., 2016) studied racism detection in Dutch social media.

The results demonstrated that it can be hard to distinguish between overt and covert aggression in social media.This is a key motivating factor for this shared task and a highly relevant discussion to include.

[1] https://github.com/zeerakw/hatespeech

Table 1: Results for the English (Facebook) task.

System	F1 (weighted)
Random Baseline	0.3535
LSTM-01	0.3160
LSTM-04	0.3160
LSTM-05	**0.3572**

3 Data

The data collection methods used to compile the dataset used in the shared task is described in (Kumar et al., 2018b). The *Aggression Identification* dataset is composed of 15,000 aggression-annotated Facebook Posts and Comments each in Hindi (in both Roman and Devanagari script) and English for training and validation. The users were asked to develop a system that could make three-way classification in between Overtly Aggressive, Covertly Aggressive and Non-aggressive text data.

4 Methodology

4.1 Preprocessing

We preprocessed the texts firstly. Punctuation symbols, unicode, urls and emoji were removed from data. All the tokens were also lowercased. For generalization proper tokenization is important. We tokenized all tweets on the basis of word-by-word. Words that contain two or more repeated characters were shortened to the same token (e.g. *fool* and *fooool* were tokenized such a way that could be treated as same). After the tokenization, we included these in training set i.e. in the FB/Tweet that contains 1 token (not a punctuation symbol), special token [2].

For all of our runs we used Long-Short-Term-Memory (LSTM) model that performed successfully in many NLP tasks (Sutskever et al., 2014) and (Hochreiter and Schmidhuber, 1997). Moreover, for the word representation we have used deepmoji and *chain-thaw* method similar manner as in (Felbo et al., 2017) to sequentially unfreezes and fine-tunes a single layer at a time. Please see Figure1 for illustration.

4.2 Model

Our model uses an embedding layer of 256 dimensions to project each word into a vector space. We used pretrained embeddings from deepmoji. To capture context of each word we use a bidirectional LSTM layer with 50 hidden units each layer. Then to capture higher level features we used one-dimensional global max pooling. Then in next layer we use 50 hidden units with ReLU as an activation function. In the next layer to avoid over-fitting we set dropout to 0.1. Finally, we used softmax layer with 3 units, each corresponding to one of three given classes. We used categorical cross-entropy as loss function and Adam optimization algorithm for optimization. Please refer Figure2 for illustration.

Our model is implemented using Theano (Team et al., 2016). We also implemented easy-to-use version available in Keras (Chollet, 2016).

Table 2: Results for the English (Social Media) task.

System	F1 (weighted)
Random Baseline	0.3477
LSTM-02	0.1960

5 Results

The scores obtained by us are shown in Table 1 and Table2. The official evaluation measures is *F1* measure. We used random baseline for sake of comparison. Table 1 contains results for English (facebook) task. For English (Facebook) task we submitted three runs. In this paper, we have just named

[2] https://github.com/bfelbo/DeepMoji

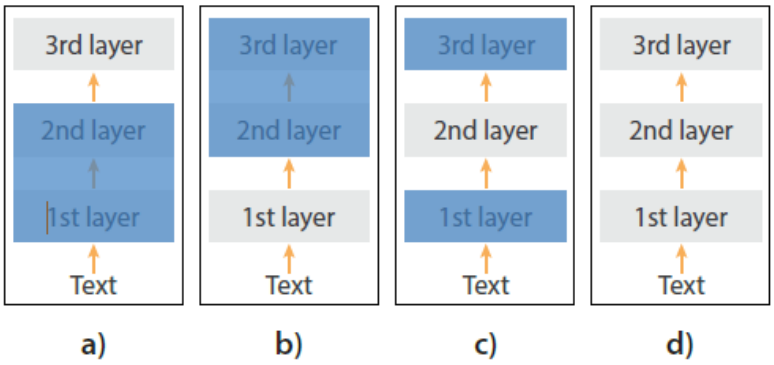

Figure 1: Illustration of the chain-thaw transfer learning approach, where each layer is fine-tuned separately. Layers covered with a blue rectangle are frozen. Step a) tunes any new layers, b) then tunes the 1st layer and c) the next layer until all layers have been fine-tuned individually. Lastly, in step d) all layers are fine-tuned together.

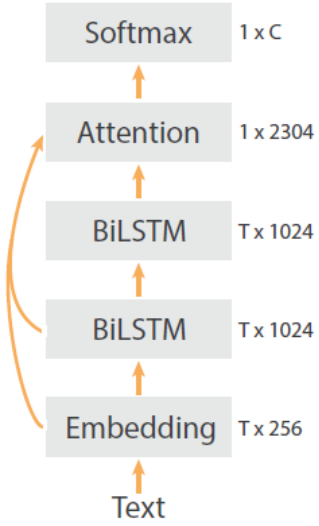

Figure 2: Illustration of the DeepMoji model with T being text length and C the number of classes.

our run as LSTM-01, LSTM-02, LSTM-03, LSTM-04 and LSTM-05. Here, **LSTM-05** performs better than LSTM-01 and LSTM-04. We just obtained slightly better score (i.e 0.3572) comparative to random baseline (i.e. 0.3535). Table2 contains the results for English (Social Media) task. Here, we get the score (i.e. 0.1960) for LSTM-02 that is much lower than random baseline.

The confusion matrix for EN-FB task and EN-TW task is shown in Figure3 and Figure4 respectively. In Figure3 we see that maximum number of 'NAG' comment is correctly predicted. 'CAG' comment is predicted after 'NAG'. The lowest prediction is of 'OAG' Comment. 'OAG' class is little bit confused with 'CAG' class.

Similarly, in Figure4 we see that 71 number of 'NAG' comment is correctly predicted. There is large difference between the class 'OAG' and 'CAG'.

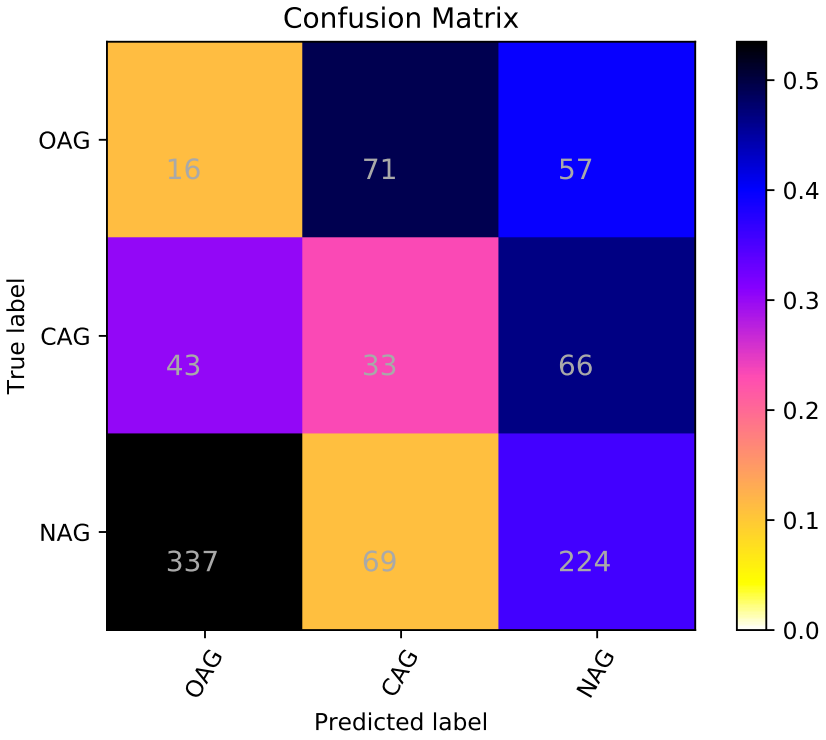

Figure 3: Confusion Matrix for EN-FB task for our 3-classes,

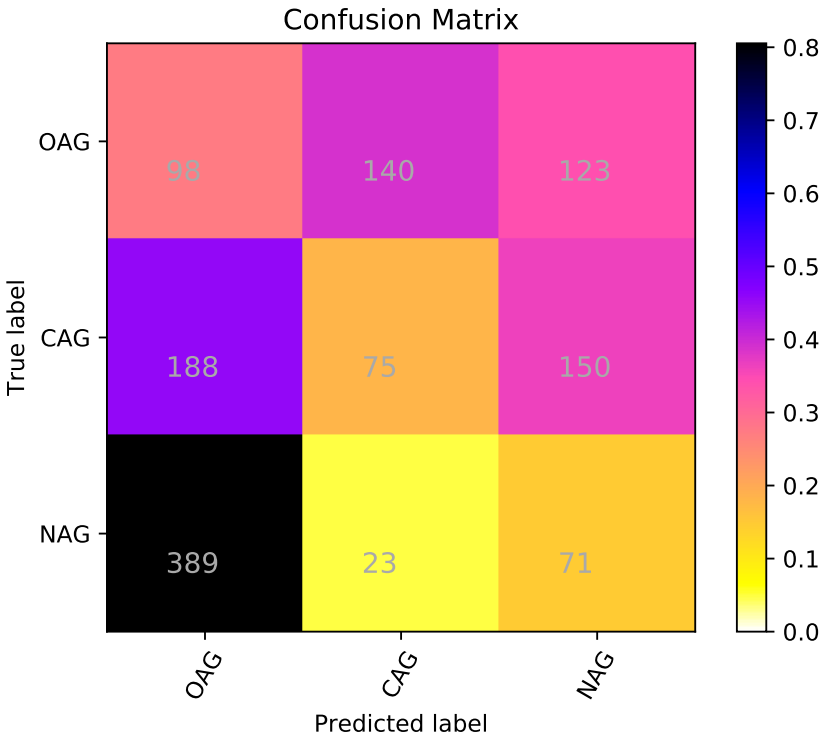

Figure 4: Confusion Matrix for EN-TW task for our 3-classes

6 Conclusion

This year we participated in TRAC-1 Shared Task on Aggression Identification in Social Media. We tried to develop 3-way classification in between 'OAG', 'CAG' and 'NAG'. Our system performs better on English (Facebook) task in comparison with English (Social Media) task. We have experimented using Deepmoji word embeddings which couldn't capture sub categories of aggression like 'OAG' or 'CAG'. While there can be no denial of the fact that our overall performance is dismal, initial results are suggestive as to what should be done next. We need to consult another model like SVM model, Ensemble model etc. for classification. Also, it will be interesting to use char CNN to extract character level features and then to use Random forest as a classifier. We shall explore some of these models in the coming days.

References

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An ap-plication of machine classification and statistical modeling for policy and decision making. Policy & Internet, 7(2):223–242.

Franc¸ois Chollet. 2016. Xception: Deep learning with depthwise separable convolutions. arXiv preprint.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In Advances in Information Retrieval, pages 693–696. Springer.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In Proceedings of ICWSM.

Jan Deriu, Maurice Gonzenbach, Fatih Uzdilli, Aurelien Lucchi, Valeria De Luca, and Martin Jaggi. 2016. Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In Proceedings of the 10th International Workshop on Semantic Evaluation, number EPFL-CONF-229234, pages 1124–1128.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In Proceedings of the 24th International Conference on World Wide Web Companion, pages 29–30. International World Wide Web Con-ferences Steering Committee.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprint arXiv:1708.00524.

Sepp Hochreiter and Jurgen¨ Schmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735–1780.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018a. Benchmarking Aggression Identification in Social Media. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC), Santa Fe, USA.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018b. Aggression-annotated Corpus of Hindi-English Code-mixed Data. In Proceedings of the 11th Language Re-sources and Evaluation Conference (LREC), Miyazaki, Japan.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting Tweets Against Blacks. In Twenty-Seventh AAAI Conference on Artificial Intelligence.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. In Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP), pages 467–472.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profan-ity from Hate Speech. Journal of Experimental & Theoretical Artificial Intelligence, 30:1–16.

Saif M Mohammad. 2012. # emotional tweets. In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pages 246–255. Associa-tion for Computational Linguistics.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In Proceedings of the First Workshop on Abusive Language Online, pages 52–56.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abu-sive Language Detection in Online User Content. In Proceedings of the 25th International Conference on World Wide Web, pages 145–153. International World Wide Web Conferences Steering Committee.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection Using Natural Language Processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics, pages 1–10, Valencia, Spain.

Hui-Po Su, Zhen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. 2017. Rephrasing profanity in chinese text. In Proceedings of the First Workshop on Abusive Language Online, pages 18–24.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pages 3104–3112.

Jared Suttles and Nancy Ide. 2013. Distant supervision for emotion classification with dis-crete binary values. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 121–136. Springer.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1555–1565.

The Theano Development Team, Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermueller, Dzmitry Bahdanau, Nicolas Ballas, Fred´eric´ Bastien, Justin Bayer, Anatoly Belikov, et al. 2016. Theano: A python framework for fast computation of mathematical expressions. arXiv preprint arXiv:1605.02688.

Stephan´ Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A Dictionary-based Approach to Racism Detection in Dutch Social Media. In Proceedings of the Workshop Text Analytics for Cybersecurity and Online Safety (TA-COS), Portoroz, Slovenia.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop, pages 88–93.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understand-ing Abuse: A Typology of Abusive Language Detection Subtasks. In Proceedings of the First Workshop on Abusive Langauge Online.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies, pages 656–666. Association for Computational Linguistics.