Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 83–86
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

LaSTUS/TALN at TRAC - 2020 Trolling, Aggression and Cyberbullying

Lütfiye Seda Mut Altın, Àlex Bravo, Horacio Saggion
Large Scale Text Understanding Systems Lab / TALN Research Group
Department of Information and Communication Technologies (DTIC)
Universitat Pompeu Fabra
Tanger 122, Barcelona (08018), Spain
lutfiyeseda.mut01@estudiant.upf.edu,{alex.bravo,horacio.saggion}@upf.edu

Abstract

This paper presents the participation of the LaSTUS/TALN team at TRAC-2020 Trolling, Aggression and Cyberbullying shared task. The aim of the task is to determine whether a given text is aggressive and contains misogynistic content. Our approach is based on a bidirectional Long Short Term Memory network (bi-LSTM). Our system performed well at sub-task A, aggression detection; however underachieved at sub-task B, misogyny detection.

1. Introduction

With millions of users contributing every day, the amount of user-generated text content forms a great amount of data, making the moderation of unwanted content highly difficult. Problematic areas of unwanted text content includes not only aggression but also trolling activities, misogyny and cyberbullying. This type of content has a proven harmful impact especially on mental health of vulnerable groups such as children and youngsters (Kwan et al., 2020). Therefore, systems that can automatically identify inappropriate content gain a lot of interest.

TRAC 2020: Second Workshop on Trolling, Aggression and Cyberbullying (TRAC – 2) shared task aims at identification of aggression and misogyny in text. It is composed of 2 sub-tasks as follows:

Sub-task A: Aggression Identification Shared Task with the classes and labels given below:

- Overtly Aggressive (OAG),
- Covertly Aggressive (CAG),
- Non-aggressive (NAG)

Sub-task B: Misogynistic Aggression Identification Shared Task with the classes and labels given below:

- Gendered (GEN),
- Non-gendered (NGEN)

The shared task is held in three Languages: English, Hindi, Bangla. With our approach, we participated in both sub-tasks only for the English language and submitted three different runs for each sub-task.
The methodology used to create this dataset is described in (Bhattacharya et al., 2020). Example instances from the dataset can be seen below:

–"Homosexuality is against nature. Thats all!" (OAG, GEN)

–"worst video" (CAG, NGEN)

–"That's the truth" (NAG, NGEN)

In this paper, we describe a neural network for text classification for aggression and misogyny identification. The rest of the paper is organized as follows: In section 2, we provide an overview of relevant research for identification of aggression and various related text classification tasks based on the relevant classes. In Section 3 we describe our model structure and specific differences of each run submitted for each sub-task. In Section 4 we provide the results and discuss the performance of the system. In Section 5 we introduce our conclusions.

2. Related Work

Many platforms such as social media sites, forums, blogs, comment and review sections of many web pages and mobile applications are heavily composed of user-generated content. As the way we communicate being substantially transformed into computer mediated communication, the need to filter out detrimental text content such as aggression and hate speech increases.

As a solution to this problem, machine learning and deep learning approaches have been utilised to classify text accordingly. Surveys reviewing previous researches indicated that instead of particular features for hate speech; generic features such as n-grams, part of speech, bag of words or embeddings are mainly used and result in reasonable performance. Moreover, character-level approaches work better than token-level approaches. In addition, lexical resources do not seem to be effective unless combined with other features (Schmidt and Wiegand, 2017), (Fortuna and Nunes, 2018). (Zampieri et al., 2019) emphasized the challenges of distinguishing profanity and threatening language which may not actually contain any swearword or profane language overtly.

Misogyny is defined as hatred, dislike, or mistrust of women, or prejudice against women [1]. One example of online misogyny is observed in the gender-biased job ads. Although, researches claim that gender discrimination in jobs ads tend to decrease (Tang et al., 2017), with the exponential increase in social media content, the need for

[1] https://www.dictionary.com/browse/misogyny?s=t

an automated identification mechanism in user generated content continues to increase.

(Cardiff and Shushkevich, 2019) reviewed previous research on automatic misogyny detection and pointed out that classical machine learning models, especially ensembles allow to achieve higher results than the models based on neural networks in some cases however these experiments were executed on relatively small datasets, therefore it is not certain that the results will be the same with an expanded dataset. Additionally, there has been shared tasks organized within this scope including identification of misogyny and also the particular groups such as stereotyping, discredit, dominance, sexual harassment and threats of violence (Fersini et al., 2018b) (IberLEF-2018), (Fersini et al., 2018a) (EVALITA-2018).

3. Methodology and Data

In our approach, we utilized the same architecture as used in SemEval-2019 Task 6: Identification and Categorization of Offensive Language in Social Media (Altin et al., 2019). This model is composed of a bidirectional Long Short-Term Memory Networks (biLSTM) model with an Attention layer on top. Within the scope of this model, for pre-processing, the instances were tokenized removing punctuation marks and keeping emojis and full hashtags as they can contribute to define the meaning of text.

Then, an embedding layer transforms each element in the tokenized text such as words, emojis and hashtags into a low-dimension vector. The embedding layer, composed of the vocabulary of the task, was randomly initialized from a uniform distribution (between -0.8 and 0.8 values and with 300 dimensions). The initialized embedding layer was updated with the word vectors included in a pre-trained model based on all the tokens, emojis and hashtags from 20M English tweets (Barbieri et al., 2016).

The dataset for English language given by the shared task organizers contains two separate files prepared for training and test. The training dataset contains around 4,000 instances (Bhattacharya et al., 2020) with two given labels for each classification type for aggression and misogyny.

For the agression sub-task we submitted 3 different runs. For the first run we used only the training data provided by the organizers. For the second run we used the additional dataset published with the same task of last year, TRAC-1 dataset (Kumar et al., 2018). For the last run, we used additional dataset from TRAC-1 and changed the optimizer to RmsProp from Adam.

Likewise, for the misogyny sub-task we submitted 3 different runs. For the first run, again we used only the training data provided by the organizers. For the second run, we used only the training dataset and changed optimizer to Nadam. For the last run we used an additional misogyny dataset (Lynn et al., 2019).

4. Results

Our system ranked 6th in sub-task A and 12th in sub-task B. We have submitted 3 different runs for each sub-task.

For sub-task A, we obtained the best result with the system which used an aditional dataset and RmsProp optimizer instead of Adam. However, the results of all runs were very close to each other. F1 (weighted) scores and accuracies obtained for each run are given in Table 1. Confusion matrix for our best performed submission for sub-task A can be seen in Figure 1. The highest recall belongs to NAG class with 92% whereas recall of other classes are 47% (CAG) and 50% (OAG). With regards to precision, NAG and CAG are similar (both around 78%) where precision of CAG is 52%.

For sub-task B, we obtained the best result with the system which used the basic dataset given and Nadam optimizer instead of Adam. The results of all runs were very close to each other. For sub-task B, F1 (weighted) scores and accuracies obtained for each run are given in Table 2. Confusion matrix for our best performed submission for sub-task B can be seen in Figure 2. Both precision and recall is higher for NGEN class (89% precision and 90% recall) whereas it is much lower for GEN class (38% precision and 34% recall).

Overall, for both sub-tasks, changes in the model for each run did not result in significant difference indicating that different optimizers and additional data did not have much effect on the results. Another point is that the main training dataset is quite unbalanced for both tasks being around 80% of the data labeled as non-Aggressive and around 70% is labeled as non-Gendered. On the other hand, although additional TRAC-1 dataset is more balanced (around 40% labeled as non-Aggressive) that did not improve the result substantially, either.

System	F1 (weighted)	Accuracy
run1	0.7100	0.7308
run2	0.7230	0.7392
run3	**0.7246**	**0.7375**

Table 1: Results for our 3 different submissions for Subtask A.

System	F1 (weighted)	Accuracy
run1	0.8137	0.8242
run2	**0.8199**	**0.8242**
run3	0.8146	0.8217

Table 2: Results for our 3 different submissions for Subtask B.

5. Conclusion

In this paper, we describe the participation of LaSTUS/TALN team to TRAC - 2020 shared task focusing on identification of aggression and misogyny in text. We utilized an architecture based on a bidirectional Long Short

	sub-task A	sub-task B
Best Performer F1 (weighted)	0.8029	0.8716
LaSTUS/TALN F1 (weighted)	0.7246	0.8199
LaSTUS/TALN Ranking / Submissions	6th / 16	12th / 15

Table 3: Comparison of the results with the best performer and rankings

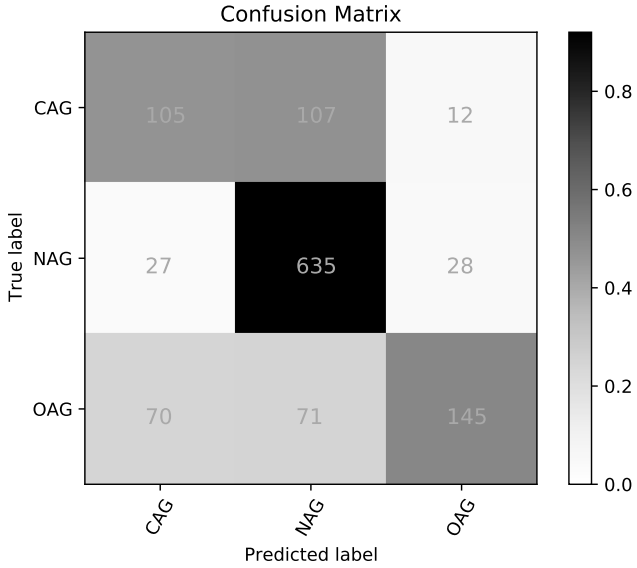

Figure 1: Confusion matrix of our best performed model (Run3) for Sub-task A

Term Memory network (biLSTM) model with an Attention layer on top. Our model performed well in the first task; however the performance was quite poor in the second task indicating that we need to improve our system for future work. Additionally, for future work, data augmentation procedures for a more balanced data can be considered.

6. Bibliographical References

Altin, L. S. M., Serrano, À. B., and Saggion, H. (2019). Lastus/taln at semeval-2019 task 6: Identification and categorization of offensive language in social media with attention-based bi-lstm model. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 672–677.

Barbieri, F., Kruszewski, G., Ronzano, F., and Saggion, H. (2016). How cosmopolitan are emojis? exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 531–535.

Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., Lahiri, B., and Ojha, A. K. (2020). Developing a multilingual annotated corpus of misogyny and aggression.

Cardiff, J. and Shushkevich, E. (2019). Automatic misogyny detection in social media: a survey.

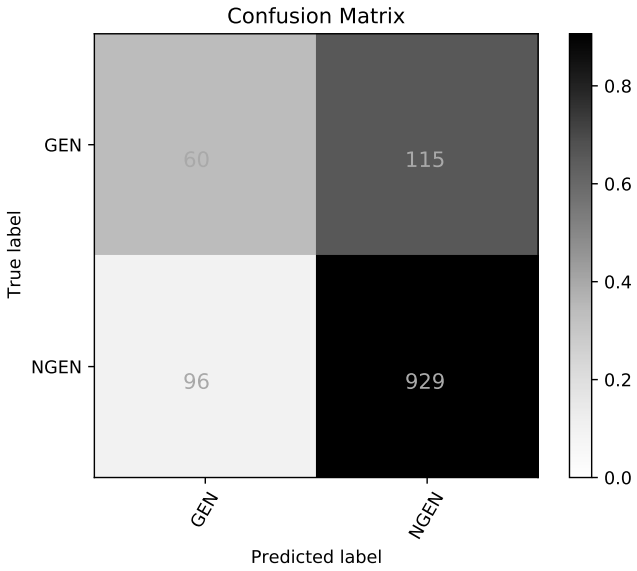

Figure 2: Confusion matrix of our best performed model (Run2) for Sub-task B

Fersini, E., Nozza, D., and Rosso, P. (2018a). Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:59.

Fersini, E., Rosso, P., and Anzovino, M. (2018b). Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@ SEPLN*, pages 214–228.

Fortuna, P. and Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, Santa Fe, USA.

Kwan, I., Dickson, K., Richardson, M., MacDowall, W., Burchett, H., Stansfield, C., Brunton, G., Sutcliffe, K., and Thomas, J. (2020). Cyberbullying and children and young people's mental health: a systematic map of systematic reviews. *Cyberpsychology, Behavior, and Social Networking*, 23(2):72–82.

Lynn, T., Endo, P. T., Rosati, P., Silva, I., Santos, G. L., and Ging, D. (2019). Data set for automatic detection of online misogynistic speech. *Data in brief*, 26:104223.

Schmidt, A. and Wiegand, M. (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.

Tang, S., Zhang, X., Cryan, J., Metzger, M. J., Zheng, H., and Zhao, B. Y. (2017). Gender bias in the job market: A longitudinal analysis. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW), December.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). SemEval-2019 Task 6: Iden-

tifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval)*.