TheNorth at HASOC 2019: Hate Speech Detection in Social Media Data

Pedro Alonso, Rajkumar Saini, and György Kovács

Luleå University of Technology, Sweden
{pedro.alonso, rajkumar.saini, gyorgy.kovacs}@ltu.se

Abstract. The detection of hate speech in social media is a crucial task. The uncontrolled spread of hate speech can be detrimental to maintaining the peace and harmony in society. Particularly when hate speech is spread with the intention to defame people, or spoil the image of a person, a community, or a nation. A major ground for spreading hate speech is that of social media. This significantly contributes to the difficulty of the task, as social media posts not only include paralinguistic tools (e.g. emoticons, and hashtags), their linguistic content contains plenty of poorly written text that does not adhere to grammar rules. With the recent development in Natural Language Processing (NLP), particularly with deep architecture, it is now possible to anlayze unstructured composite natural language text. For this reason, we propose a deep NLP model for the detection of automatic hate speech in social media data. We have applied our model on the HASOC2019 hate speech corpus, and attained a macro F1 score of 0.63 in the detection of hate speech.

1 Introduction

In the course of our lifetime, we have experienced an increase in social media usage [3]. Social media, when used with care, can be beneficial for its users, but it can also be a hotbed for bullying, online harassment, and the spread of hate speech. All these factors can severely impact both individual users and society in a negative way. For this reason, it is becoming more and more important to provide automatic hate speech detection tools, which can help curb its appearance on social media (Twitter in this case). Therefore, it is of the utmost importance to have an ability to monitor the offensive content being published and let the moderators take the steps they deem necessary. This is especially important when trying to protect vulnerable groups of people like immigrants [2], women, members of the LBTQ community, or members of any other group that is the target of hate. While several attempts have been made to detect hate speech in comments, [5], [4], [6]. The models still could use some more fine-tuning, so our model could be considered an addendum to the pool of existing ones aimed at increasing the accuracy of hate speech detection in the wild.

Table 1: Some samples from the English language training data along with their ground truth labels.

Post	**Ground Truth**
New logo for world Cup designed by ICC #ShameOnICC https://t.co/AtFL15Gt9B	NOT/NONE/NONE
@TheRealOJ32 The world will rejoice when you die.	HOF/OFFN/TIN
#DoctorsFightBack we want justice https://t.co/ONUdOhagX3	HOF/HATE/UNT
Just watched this guy spray the crap from his curb to the curb of his next door neighbor. #DickHead	HOF/PRFN/TIN

2 Hate speech detection on twitter

Here, the task to be undertaken is that of "Hate Speech and Offensive Content Identification in Indo-European Languages" challenge, a task inspired by similar prior challenges [9,10]. Ask the task is detailed in the accompanying overview paper [8], here we only discuss it briefly. HASOC2019 data consists of social media posts from Twitter and Facebook in a tab-separated format. The data in the dataset is available in three different languages, namely English, German, and code-mixed Hindi. Here, we exclusively process English language posts. The training data for English consists of 6358 instances. Some examples of the training set (along with their ground truth labels) are shown in Table 1.

The HASOC2019 task description [7,8] defines three sub-tasks in the hate speech detection challenge. These sub-tasks are as follows:
Sub-task A: The task we tackle in our experiments is that of task A. The task here is a more general binary classification of social media posts into two categories, specifically the "Hate and Offensive" category (HOF) and the "Non-Hate and offensive" category (NOT).

- NOT: These are posts without sentences considered to be hate speech or offensive in content.
- HOF: These posts are considered to contain hateful, offensive or profane language.

Sub-task B: This task is concerned with a more detailed classification for the post in the previous category HOF, this time divided into three categories.

- Hate speech (HATE): Posts that belong in this class contain hate speech sentences. These include, description of negative attributes or ascribing deficiencies to individuals because they belong to a certain group (e.g. poor people are dumb). Can also comprise hateful comments geared to certain groups of people based on their race, political opinion, sexual orientation, gender, social status, or health condition.
- Offensive (OFFN): Posts that belong in this class contain offensive content. This means posts that degrade, dehumanize, or insult an individual. Posts that threaten individuals with violent acts re also categorized into this class.

- Profane (PRFN): Posts that belong in this class contain profane words, or unacceptable language but without directed insults or abuse. This class typically concerns the use of swearwords (e.g. shit, fuck), and cursing.

Sub-task C: A fine-level classification of social media posts in the HOF category from a different perspective. Here, differentiation between hateful posts are made on the ground of the post containing directed hate, or hate/offensive language in general (e.g. Who the fuck voted for a no deal?")

- Targeted Insult (TIN): Posts deemed insulting or threatening towards an individual, group, or others.
- Un-targeted (UNT): Posts deemed to no be targeted towards a specific individual or group, but still contain unacceptable language.

Table 2: The name of the Indian prime minister used in pop-cultural references

Post	**Ground Truth**
Modi Ji will never give you up Modi ji will never give you down	NOT
Modi Ji knows Coca Cola's secret ingredient	NOT
Modi Ji knows why is Gamora	HOF
Modi Ji knows who let the dogs out	HOF

2.1 Difficulties

One degree of difficulty emerges from the nature of social media posts. Namely, that textual content shared on social media is rarely well-formed, and often contains paralinguistic elements, such as URLs, emoticons, and other special characters. Another degree of difficulty is due to the inherent unbalanced nature of hate speech detection. As the majority of social media posts contain no hate speech or profanity. Lastly, a third degree of difficulty emerges from the subjective nature of hate speech labeling. For example, the training set of HASOC2019 contains a serious of pop-cultural references that include the prime minister of India, Narendra Modi (Modi Ji). And while on the surface these tweets (see Table 2) are all innocuous, some are classified as hateful, while others are classified as non-hateful, without any clear logic. Another example to the subjective nature of decisions about hate speech and offensive content is how people react very differently to the use of the word "fuck" when it is used as part of a hashtag, as opposed to when it is used without a hashtag. In the training set of HASOC2019 the number of tweets that contain the word fuck with, and without a hashtag is 1159 and 215 respectively. After eliminating those tweets that contain both, these numbers decrease to 1072 and 128. These two categories are very much different, as when the word is used without a hashtag alone in a tweet, more than 97% of the tweets are considered hateful. However, for the hashtagged version, this number is only approximately 41% (while for tweets that do not contain either, this is approximately 38%). This indicated that while the use of the word fuck in and of itself highly increases the probability of a tweet deemed as hateful, tweets with fuck in a hashtag are only slightly more likely to be treated the same way.

3 Experimental setup

In this section we describe the model we applied for the task, and also shortly describe our method for training said model.

4 Hate speech detection model

Now, we present the architectural details of the proposed system. The system architecture is shown in Fig. 1. The following figure shows the deep neural network used in our approach. Our approach is similar to [1] and [11], where they showed that with convolution layers at the beginning, the top could vary and get accurate results. Therefore our model follows the same principle of, stacking a few convolutions at the top, and the varied the intermediate layers, in our case we chose a Bi-LSTM, to contrast with the LSTM and GRU used in the papers. We started with an input layer with the number of batches times the length of the text (in our case fifty), then we used an embedding layer which was self-trained. The next stage is made up of convolutions of sixteen, eight and four to reduce the size of the input as much as we could without losing too much information, we use a max-pooling layer of size 4 at the end. Next stage we, use three bi-directional LSTM with one thousand six hundred neurons for the final classification part. Then we use a combination of dense and dropout layers, where the dropout probability is set to be 0.5. Lastly, we use a soft-max layer with two neurons corresponding to the two classes (NOT, HOF).

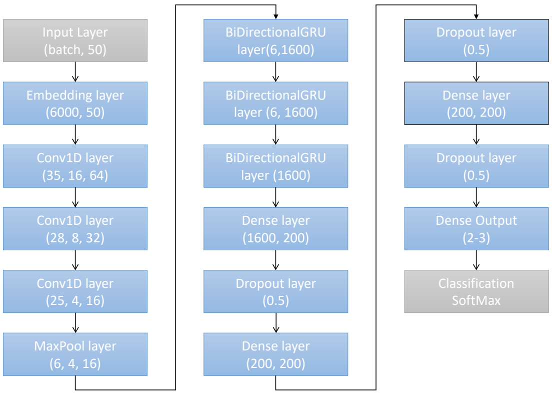

Fig. 1: Architecture of the proposed hate speech detection system.

4.1 Model training

For training our models, we first partitioned the labeled data available into two sets using 10% and 90% of the instances. The former we used for model evaluation (and will reference it in this paper as the evaluation set), while the latter we partitioned again in the same ratio. The bigger set resulting we used for training our models (and will reference it in this paper as the training set), while the smaller set we used for early stopping (and will reference it in this paper as the validation set). Then we trained our model for at most a hundred epochs using the samples from the training set. After each epoch we only kept the changes if there was an improvement in the macro F1 score attained on the validation set. Otherwise we did reset the weights to the result of the last successful epoch, and continued the process of training. If there were three consecutive epochs where the macro F1 score did not improve on the validation set, we stopped the process, and saved our final model.

5 Results and Discussion

For this paper we carried out all our experiments on Sub-task A using only the English language posts. These experiments have been conducted in two runs. Figure 3 shows the statistics for the Sub-task A for both of these runs (run1 and run2). The overall average $F1$ and weighted average $F1$ scores in run1 (run2) as 0.6279 (0.6094), and 0.6963 (0.6779) have been recorded respectively on Sub-task A. As we see in 3 the precision and recall are higher for the NOT class, than for the offensive one.

While in Figure 2, we present the results as a confusion matrix. In this figure we can again see that one weak point of our model is its sensitivity to HOF. This also shows that classifying offensive language is still a difficult task for the algorithm.

6 Conclusion

The detection of hate speech requires more attention in the age of the Internet, as it can now spread faster. Hate speech can cause severe social/moral damage to our society. In this paper, we investigate the HASOC2019 hate speech detection dataset. Alhough the dataset contains three languages (English, German, and code-mixed Hindi), we have worked only with English data. The proposed system works relatively well on Sub-task A. The weighted average $F1$-scores of 0.6963, and 0.6779 have been recorded on Sub-task A in run 1, and run 2 respectively.

Table 3: Results of Sub-task A (run1/run2)

	precision	**recall**	**F1-score**	**support**
HOF	0.41/0.38	0.62/0.61	0.49/0.47	288
NOT	0.85/0.84	0.70/0.67	0.76/0.75	865
accuracy	–	–	0.68/0.66	1153
macro average	0.63/0.61	0.66/0.64	0.63/0.61	1153
weighted average	0.74/0.73	0.68/0.66	0.70/0.68	1153

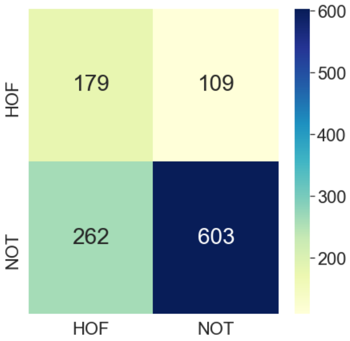

Fig. 2: Confusion matrix of results on the test set, produced by our first model trained for Task A)

In the future, we are planning to try different architectures with varying degrees of complexity to get better performance for the task here described. Also, we shall try to gather more data to be sure our model developed has a sufficient amount of samples to work efficiently.

References

1. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 759–760. WWW '17 Companion (2017)
2. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F.M., Rosso, P., Sanguinetti, M.: SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 54–63. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019), https://www.aclweb.org/anthology/S19-2007
3. Chou, W.y.S., Hunt, Y.M., Beckjord, E.B., Moser, R.P., Hesse, B.W.: Social media use in the united states: Implications for health communication. J Med Internet Res 11(4) (Nov 2009)
4. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Eleventh international AAAI conference on web and social media (2017)
5. Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N.: Hate speech detection with comment embeddings. In: Proceedings of the 24th International Conference on World Wide Web. pp. 29–30. WWW '15 Companion, ACM, New York, NY, USA (2015), http://doi.acm.org/10.1145/2740908.2742760
6. Gambäck, B., Sikdar, U.K.: Using convolutional neural networks to classify hate-speech. In: Proceedings of the First Workshop on Abusive Language Online. pp. 85–90. Association for Computational Linguistics, Vancouver, BC, Canada (Aug 2017), https://www.aclweb.org/anthology/W17-3013
7. Mandl, T., Modha, S., Mandlia, C., Patel, D., Patel, A., Dave, M.: HASOC - hate speech and offensive content identification in indo-european languages. https://hasoc2019.github.io/call_for_participation.html, accessed: 2019-09-20

8. Modha, S., Mandl, T., Majumder, P., Patel, D.: Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (2019)
9. Wiegand, M., Siegel, M., Ruppenhofer, J.: Overview of the germeval 2018 shared task on the identification of offensive language (2018)
10. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 75–86 (2019)
11. Zhang, Z., Robinson, D., Tepper, J.: Detecting hate speech on twitter using a convolution-gru based deep neural network. In: Gangemi, A., Navigli, R., Vidal, M.E., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., Alam, M. (eds.) The Semantic Web. pp. 745–760. Springer International Publishing, Cham (2018)