The Titans at SemEval-2019 Task 5: Detection of hate speech against immigrants and women in Twitter

Avishek Garain
Computer Science and Engineering
Jadavpur University, Kolkata
avishekgarain@gmail.com

Arpan Basu
Computer Science and Engineering
Jadavpur University, Kolkata
arpan0123@gmail.com

Abstract

This system paper is a description of the system submitted to "SemEval-2019 Task 5" Task B for the English language, where we had to primarily detect hate speech and then detect aggressive behaviour and its target audience in Twitter. There were two specific target audiences, immigrants and women. The language of the tweets was English. We were required to first detect whether a tweet is containing hate speech. Thereafter we were required to find whether the tweet was showing aggressive behaviour, and then we had to find whether the targeted audience was an individual or a group of people.

1 Introduction

Hate speech attacks a person or a group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation or gender identity. In the same time, flames (such as rants, taunts, and squalid phrases) are offensive/abusive phrases which might attack or offend the users for a variety of reasons. This is very pertinent due to rise of text messaging through the Internet or cellular phones, which has become a major medium of personal and commercial communication.

Aggression is overt, often harmful, social interaction with the intention of inflicting damage or other unpleasantness upon another individual. It may occur either in retaliation or without provocation. In humans, frustration due to blocked goals can cause aggression. Human aggression can be classified into direct and indirect aggression; whilst the former is characterized by physical or verbal behavior intended to cause harm to someone, the latter is characterized by behavior intended to harm the social relations of an individual or group.

Hate speech and offensive language are pervasive in social media. Online communities, social media platforms, and technology companies have been researching heavily in ways to cope with this phenomena to prevent abusive behavior in social media. This is due to text messaging through the Internet or cellular phones, which has become a major medium of personal and commercial communication.

One of the most effective strategies for tackling this problem is to use computational methods to identify hate speech and aggression in user-generated content (e.g. posts, comments, tweets etc.). This topic has attracted significant attention in recent years of various Natural Language analysts.

The SemEval 2019 task 5 (Basile et al., 2019) was a classification task where we were required to classify a tweet as containing hate speech or otherwise. However, there were some additional challenges presented, which involved automatic detection of aggression, and classification the target audience as an individual or group of people.

To solve the task in hand we built a bidirectional LSTM based neural network for prediction of the three classes present in the provided dataset. In the first subtask our system categorized the instances into `HS` and `NOT`. In the second subtask our system categorized instances into `AGR` and `NOT`. In the third subtask our system categorized instances into `IN` or `GRP`.

The paper has been organized as follows. Section 2 describes a brief survey on the relevant work done in this field. Section 3 describes the data, on which, the task was performed. The methodology followed is described in Section 4. This is followed by the results and concluding remarks in Section 5 and 6 respectively.

2 Related Work

Papers which have been published in the last two years include the surveys by (Schmidt and Wiegand, 2017) and (Fortuna and Nunes, 2018), the

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 494–497
Minneapolis, Minnesota, USA, June 6–7, 2019.

paper by (Davidson et al., 2017) presenting the Hate Speech Detection dataset used in (Malmasi and Zampieri, 2017) and a few other recent papers such as (ElSherief et al., 2018; Gambäck and Sikdar, 2017; Zhang et al., 2018).

We were guided by the work of (Zhang et al., 2018) who used a CNN+GRU based approach for a similar task. We use an approach which was influenced by this work but used an LSTM based approach.

A proposal of typology of abusive language sub-tasks is presented in (Waseem et al., 2017). For studies on languages other than English see (Su et al., 2017) on Chinese and (Fišer et al., 2017) on Slovene. Finally, for recent discussion on identifying profanity vs. hate speech see (Malmasi and Zampieri, 2018). This work highlighted the challenges of distinguishing between profanity, and threatening language which may not actually contain profane language.

Previous editions of related workshops are TA-COS[1], Abusive Language Online[2], and TRAC[3] and related shared tasks are GermEval (Wiegand et al., 2018) and TRAC (Kumar et al., 2018).

3 Data

The dataset that was used to train the model is the HatEval dataset (Basile et al., 2019). It was collected from Twitter; the data being retrieved the data using the Twitter API by searching for keywords and constructions that are often included in aggressive messages.

Label	Meaning
HS	Whether the tweet contains hate speech or not
TR	Whether the tweet containing profanity is targeted against some individual/group/others
AG	Whether the tweet contains aggressive behaviour or not

Table 1: Labels used in the dataset

The dataset provided consisted of tweets in their original form along with the corresponding `HS`, `TR` and `AG` labels. The dataset had 9000 instances of training data and 1000 instances of development data. Our approach was to convert the tweet into a sequence of words and then run a neural-network based algorithm on the processed tweet.

Value	HS	TR	AG
0	5217	2442	2224
1	3783	1341	1559
All	9000	3783	3783

Table 2: Distribution of the labels in the training dataset

Value	HS	TR	AG
0	573	208	223
1	427	219	204
All	1000	427	427

Table 3: Distribution of the labels in the development dataset

The provided training and development data were merged and shuffled to create a bigger training set, and we refer to the same as training data when we discuss our methodology.

Value	HS	TR	AG
0	5790	2650	2447
1	4210	1560	1763
All	10000	4210	4210

Table 4: Distribution of the labels in the combined dataset

4 Methodology

The first stage in our pipeline was to preprocess the tweet. This consisted of the following steps:

1. Removing mentions
2. Removing punctuations
3. Removing URLs
4. Contracting whitespace
5. Extracting words from hashtags

The last step (step 5) consists of taking advantage of the Pascal Casing of hashtags (e.g. `#PascalCasing`). A simple regex can extract all words; we ignore a few errors that arise in this procedure. This extraction results in better performance mainly because words in hashtags, to some extent, may convey sentiments of hate. They play an important role during the model-training stage.

We treat the tweet as a sequence of words with interdependence among various words contribut-

[1] http://ta-cos.org/

[2] https://sites.google.com/site/abusivelanguageworkshop2017/

[3] https://sites.google.com/view/trac1/home

ing to its meaning. Hence we use an bidirectional LSTM based approach to capture information from both the past and future context.

Our model is a neural-network based model. First, the input tweet is passed through an embedding layer which transforms the tweet into a 128 length vector. The embedding layer learns the word embeddings from the input tweets. This is followed by two bidirectional LSTM layers containing 64 units each. This is followed by the final output layer of neurons with softmax activation, each neuron predicting a label as present in the dataset. For subtasks 1, 2 and 3, we train separate models containing 2 neurons for predicting `HS(0/1)`, `TR(0/1)` and `AG(0/1)` respectively. Between the LSTM and output layers, we add dropout with a rate of 0.5 as a regularizer. The model is trained using the Adam optimization algorithm with a learning rate of 0.0005 and using crossentropy as the loss.

We note that the dataset is highly skewed in nature. If trained on the entire training dataset without any validation, the model tends to completely overfit to the class with higher frequency as it leads to a higher accuracy score.

To overcome this problem, we took some measures. Firstly, the training data was split into two parts — one for training and one for validation comprising 70 % and 30 % of the dataset respectively. The training was stopped when two consecutive epochs increased the measured loss function value for the validation set.

Secondly, class weights were assigned to the different classes present in the data. The weights were approximately chosen to be proportional to the inverse of the respective frequencies of the classes. Intuitively, the model now gives equal weight to the skewed classes and this penalizes tendencies to overfit to the data.

5 Results

We participated in English Task B of Semeval 2019 task 5 (HatEval) and our system ranks fourth among the competing participants.

We have included the automatically generated tables with our results. We have also included the provided baselines generated by MFC and SVC classifiers respectively. The SVC baseline is generated by a linear SVM based on a TF-IDF representation. The MFC baseline assigns the most frequent label in the training set to all instances present in the test set. We have used these baselines for comparison.

System	Train (%)	Validation (%)
Without	99.82	66.74
With	99.95	70.31

Table 5: Comparison of development phase accuracies with and without hashtag preprocessing

System	F1 (avg)	EMR
MFC baseline	0.421	0.580
SVC baseline	0.578	0.308
BiLSTM	0.471	0.482

Table 6: Overall Metrics

System	F1	Accuracy
MFC baseline	0.367	0.580
SVC baseline	0.45	0.491
BiLSTM	0.484	0.573

Table 7: HS Metrics

System	F1	Accuracy
MFC baseline	0.452	0.824
SVC baseline	0.697	0.785
BiLSTM	0.464	0.817

Table 8: TR Metrics

System	F1	Accuracy
MFC baseline	0.445	0.802
SVC baseline	0.587	0.692
BiLSTM	0.464	0.763

Table 9: AG Metrics

6 Conclusion

Here we have presented a model which performs satisfactorily in the given tasks. The model is based on a simple architecture. There is scope for improvement by including more features (like those removed in the preprocessing step) to increase performance. Another drawback of the model is that it does not use any external data other than the dataset provided which may lead to poor results based on the modest size of the data. Related domain knowledge may be exploited to obtain better results.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, Paolo Rosso, and Manuela Sanguinetti. 2019. 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 (SemEval-2019)*. Association for Computational Linguistics.

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

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. *arXiv preprint arXiv:1804.04257*.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal Framework, Dataset and Annotation Schema for Socially Unacceptable On-line Discourse Practices in Slovene. In *Proceedings of the Workshop Workshop on Abusive Language Online (ALW)*, Vancouver, Canada.

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

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90.

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

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 Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1–16.

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.

Huei-Po Su, Chen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. 2017. Rephrasing Profanity in Chinese Text. In *Proceedings of the Workshop Workshop on Abusive Language Online (ALW)*, Vancouver, Canada.

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

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Lecture Notes in Computer Science*. Springer Verlag.