KMI−Coling at SemEval-2019 Task 6: Exploring N-grams for Offensive Language detection

Priya Rani
Dr. Bhimrao Ambedkar University
Agra, India
pranijnu@gmail.com

Atul Kr. Ojha
Jawaharlal Nehru University
New Delhi, India
shashwatup9k@gmail.com

Abstract

In this paper, we present the system description of offensive language detection tool which is developed by the KMI−Coling Group under the OffensEval Shared task. The OffensEval Shared Task was conducted in SemEval 2019 workshop. To develop the system, we have explored n-grams up to 8-gram and trained three different systems namely A, B and C system for three different sub tasks within the OffensEval task which achieves the accuracy of 79.76%, 87.91% and 44.37% respectively. The task was completed using the data set provided to us by OffensEval organisers, which was the part of OLID data set. It consists of 13,240 tweets extracted from twitter and were annotated at three levels using crowd sourcing.

1 Introduction

The very first question which arises in one's mind when one starts working in the area of computational sociolinguistics research related to the language usage in social media and networking sites is what is offensive language and its related terms such as hate speech, aggression and all? The second question arises is related to the definition of this terminology. We would suggest that offensive language is still not a very well-defined phenomenon. As we know that the natural language is productive in nature. These aspect of the language use has always existed as part of the speech repertoire of the speaker. In the area of scientific study, we need to move forward with a definition. Therefore we will go by the definition given by Jay and Janschwetiz which states that Offensive language is vulgar, pornography and hateful language (Chen et al., 2012). But even this definition does not incorporate the many more structure which is neither vulgar nor pornography nor hateful but are definitely offensive. Such type of structure is what leads to challenges in the detection of offensive language in the discourse. With the increase in the culture of social media and social networking sites, the use of offensive language has increased rapidly. Moreover, it has also given a very good platform to conduct different research in the given area.

2 Literature review

This section gives a brief outline of the existing literature and approaches that are available for offensive language detection. Lots of research works are being done to detect offensive language and there has been significant progress over time. Lexical Syntactic based framework was used for sentence offensive detection and user offensive detection by Chen et al. (2012). Another study by Xiang et al. (2012) which uses keyword matching technique that performed very well in literature domain. Razavi et al. (2010) uses auxiliary weighted repository by matching the text to its graded entries with the help of both rule-based and statistical pattern to detect flames from the text. Maisto et al. (2017) uses a lexicon-based method for the automatic identification and classification.

3 System Overview

We built three different systems for three sub tasks in the shared task. The system was built using a supervised machine learning approach trained on different classifiers using n-gram model.

3.1 System A

The very first was developed to detect whether the tweets are offensive or not. The system uses unigram and bigram in the feature set and was trained on Linear SVM classifier.

3.2 System B

The second system was developed to detect whether the tweets are targeted or non-targeted

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 668–671
Minneapolis, Minnesota, USA, June 6–7, 2019.

only if the tweets are offensive in nature. This system has also been trained on linear SVM with n-gram language model which consisted of unigram, bigram, trigram and 4-gram.

3.3 System C

The third system was one step ahead to detect whether the tweets are targeted to an individual, group or other. In which the third category 'other' includes a wide range of categories such as entity, organisation, place, country and many more. The system is trained on decision tree with n-gram feature starting from uni-grams to 8-grams.

4 Experiments

In this section, we briefly describe the experimental settings which are used to develop offensive language detection tool.

4.1 Data set

The data we used to train and test the system was provided by SemEval shared task 2019 under task 6 called OffensEval (Zampieri et al., 2019a). The data set consists of 13,240 annotated tweets which were extracted from OLID, Offensive Language Identification Dataset (Zampieri et al., 2019a). The data set was further divided into training and testing set in the ratio 80:20. We have used the same data set to train and test all the three systems developed to participate in the sub task of Task 6 in SemEval 2019.

4.2 Annotation

The data was hierarchically annotated using crowd sourcing. The gold labels were assigned by taking inter-annotator agreement into consideration. No correction has been carried out on the crowd sourcing annotations. The tweets were annotated at three levels. Level A differentiates the tweets between offensive and non-offensive. Level B category the offensive tweets in another level that whether the offensive tweets are targeted or non-targeted insults or threat. Level C category further categorise the targets of the insult into three different categories as an individual, group or other (Zampieri et al., 2019a).

4.3 Development of systems for sub tasks

In the next step, we developed three offensive detection systems to detect offensive tweets, targeted insults and to categorise the targeted insults using n-gram language model.

Training and development of system for sub task A

The systems were trained independently on SVM. To explore the role of n-gram feature in the detection of offensive language we have used the scikit-learn toolkit to experiment with unigram, bigram, trigram and 4-gram. We used the tweets and only Label A to train the system for development of system A.

Training and development of system for sub task B

Like system A, system B is trained on SVM with the scikit-learn toolkit to experiment with unigram, bigram, trigram and 4-gram. The system was trained and tested using tweets, Label A and Label B.

Training and development of system for sub task C

The third system was trained on two different classifier SVM and Decision tree with the same scikit-learn toolkit. The feature set for the system consists of n-gram ranging from unigrams to 8-gram.s In order to train the system, we have used the tweets, Label B, and Label C.

5 Detailed Error Reports of the KMI−Coling System

This section presents a detailed study of the result that is achieved by the developed systems.

System A performs well with the only unigram when trained with SVM. We also weighted the feature set with TF-IDF but that turned out to give very disappointing results and thus was discarded from the final feature set. Similarly, bigram, trigram and 4-gram decrease the accuracy of the system. The final trained system gave the precision of 78.72%, recall of 79.77% with an F1 score of 78.58%. The confusion matrix of the system providing the detail error is given in figure 1, which shows that 122 offensive tweets were called as non-offensive. Whereas 568 tweets were recognised correct by the system. On the other hand, 52 non-offensive tweets were labelled as offensive tweets and 118 offensive tweets were marked correctly by the system (Zampieri et al., 2019b).

System B, unlike system A, performed well in terms of accuracy, but the recall of non-targeted offensive tweets is lowered when trigram and 4-gram are implemented. Weighting the feature set using TF-IDF also did not work well as it de-

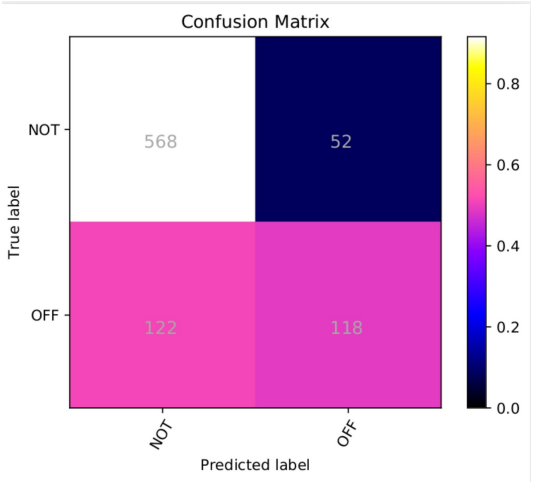

Figure 1: Confusion matrix of sub task A

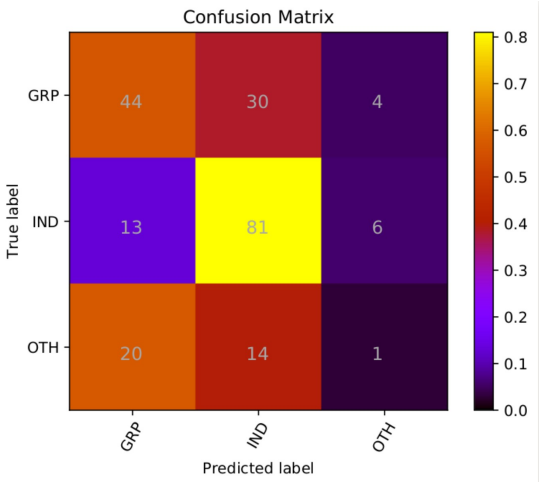

Figure 3: Confusion matrix of sub task C

creases the accuracy of the system. Finally, the system was trained only with unigram and bigram. The overall precision of the system is 84.38%, recall is 87.92% and F1 score is 85.37%. Figure 2 shows the confusion matrix of the system. The matrix gives the error report such that 23 non-targeted tweets were label targeted by the system whereas only 6 targeted tweets were marked as non-targeted.

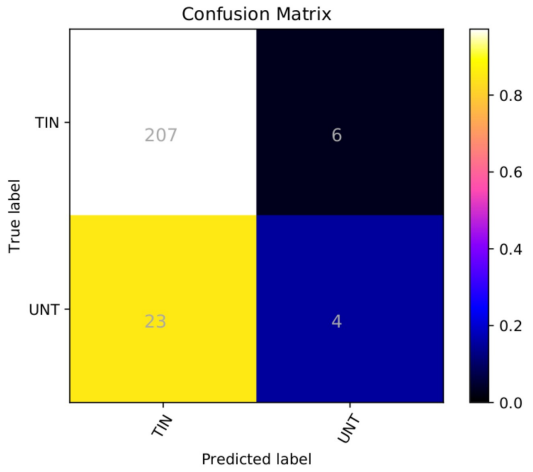

Figure 2: Confusion matrix of sub task B

As we have mentioned above system c was trained on two classifiers SVM and Decision Tree. We will be only reporting the final confusion matrix of the system C which was trained on Decision Tree. The precision of the system is 52.84%, recall is 59.15% and F1 score is 55.31%. We can easily detect and study the error from the confusion matrix given in Figure 3. Twenty tweets which originally belong to other group categorised in GROUP, 14 tweets in INDIVIDUAL. Secondly, 13 tweets which belong to INDIVIDUAL were put into GROUP and 6 of them in OTHER. Thirdly, 30 tweets which were targeted towards a group were labelled in INDIVIDUAL and 4 tweets in OTHER.

6 Conclusion

In this paper, we propose an offensive detection tool which is only based on the n-gram model. We have experimented with n-gram model where n = 1,2,3,4,5,6,7,8 via statistical model. The n-gram model has been shown to perform well in very less amount of time in comparison to other models. The accuracy of system A is 79.76%, system B is 87.91% and of system C is 44.37% in our experiment. In addition to this, it is very easy to implement n-gram and consume very less amount of time. Our system can be further improved with the help of neural network. As we can see that the n-gram model also accommodate the phrase level structure from the given text. Therefore, implementing simple sentence feature would not help in increasing the accuracy. The sentence level feature would work only when there is a language specific feature.

Acknowledgments

We are grateful to the organizers of Offsemeval-2019 for providing us with the tweet Corpus and evaluation scores.

References

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social me-

dia to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 71–80. IEEE.

Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. 2014. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Canadian Conference on Artificial Intelligence*, pages 275–281. Springer.

H Vinutha Divyashree and NS Deepashree. 2016. An effective approach for cyberbullying detection and avoidance. *International Journal of Innovative Research in Computer and Communication Engineering*, 14.

Love Engman. 2016. Automatic detection of cyberbullying on social media.

Altaf Mahmud, Kazi Zubair Ahmed, and Mumit Khan. 2008. Detecting flames and insults in text.

Alessandro Maisto, Serena Pelosi, Simonetta Vietri, Pierluigi Vitale, and Via Giovanni Paolo II. 2017. Mining offensive language on social media. In *Proocedings of CLiC-it 2017 4th Italian Conference on Computational Linguistics*.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

Caitlin Elizabeth Ring. 2013. Hate speech in social media: An exploration of the problem and its proposed solutions.

Semiu Salawu, Yulan He, and Joanna Lumsden. 2017. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, (1):1–1.

Joni Salminen, Hind Almerekhi, Milica Milenkovic, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J Jansen. 2018. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *ICWSM*, pages 330–339.

Sasha Sax. 2016. Flame wars: Automatic insult detection.

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*, pages 1–10.

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *arXiv preprint arXiv:1801.05617*.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984. ACM.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval)*.