CEUR-WS.org/Vol-2517/T3-17.pdf

IRLab@IITBHU at HASOC 2019: Traditional Machine Learning for Hate Speech and Offensive Content Identification

Anita Saroj, Rajesh Kumar Mundotiya, and Sukomal Pal

[1] Indian Institute of Technology (BHU)
India, Varanasi (UP) 221005
[2] {anitas.rs.cse16, rajeshkm.rs.cse16, spal.cse}@iitbhu.ac.in

Abstract. In this paper, the results obtained from the Support Vector Machine, XGBoost method by IRLab@IIT(BHU) on HASOC shared task-organized at FIRE-2019 are reported. The HASOC shared task has three subtasks, namely Hate speech identification, Offensive language identification and Fine-grained classification for the English, Hindi and German languages. The best result for English is obtained after applying Support Vector Machine, XGBoost with a frequency-based feature for hate speech and offensive content identification.

Keywords: Offensive · Hate Speech · Language · Social Media

1 INTRODUCTION

Hate speech is a type of communication of verbal expression to attack a human or group based on characteristics such as caste, religion, ethnic origin, sexual orientation, disability or gender [7]. Hate speech and offensive content in Indo-European languages have become a common phenomenon in the social media. Recent years have seen the spread of offensive language on social media platforms such as Facebook and Twitter. With the freedom of privilege of expression granted to social media users, it became easy to spread disrespect or hatred against individuals or groups. Automated hate language and offensive content detection systems may contain the spread inhibition of toxin textual material. Beyond psychological harm, such toxic online content can give rise to real hate crimes [10]; which justifies the need to automatically detect abusive language and offensive content shared on social media platforms.

The rest of the paper organised as follows. In Sec 2, we do literature survey. Next, we describe the methodology of the paper 3. We discuss the result in Sec 4. Finally we conclude in Sec 5.

[3]

[3] Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). FIRE 2019, 12-15 December 2019, Kolkata, India.

2 RELATED WORK

Over the last few years, several studies on hate speech and offensive content identification have been published. The literature has explored different offensive and abusive language identification problems ranging from aggression to cyber bullying, hate speech, poisonous comments and offensive language. We briefly discuss each of them in this section.

2.1 Aggressive content identification

The first shared task on aggression identification is Trolling, Aggression and Cyberbullying (TRAC-1) at COLING 2018. In this task Aggressive Language Identification on Facebook and Twitter data targeted using word embeddings and sentiment features [13]. Moreover, the best result were obtained through sentiment features with Random Forest (RF) and Support Vector Machine (SVM) with 0.5830, 0.5074 accuracies, respectively. Later, efforts went to develop a classifier that could discriminate between overly aggressive, hidden aggressive, and non-aggressive text. Long short-term memory (LSTM), Convolutional Neural Network (CNN)-LSTM, Bidirectional LSTM with Glove embeddings, the combination of the Passive-Aggressive (PA) and SVM classifiers with character-based n-gram where n is from 1 to 5, TF-IDF as feature representation were used for aggression identification. The best system explained above to achieve a weighted F-score of 0.64 on the Facebook test set entitled as English and Hindi, and the best scores for the surprise set were 0.60 and 0.50 for Hindi and English respectively [8, 16, 1, 18, 19].

2.2 Bullying content identification

Bullying, also known as peer victimization, has been recognized as a serious national health issue by the White House (2011), the American Academy of Pediatrics (2009), and the American Psychological Association (2004) [5, 2, 6]. The growing research into cyberbullying in online social networks have catalyzed by the widespread and profound consequences of abuse. Earlier research works on automatic cyberbullying detection have mainly focused on using (sophisticated) text-based methods [4, 12, 15]. Expanded the text-based identification approach to model the use of hashtags, simultaneously with the emotions the spatio-temporal cyberbullying measures to understand and explore. [17].

2.3 Hate speech identification

Hate speech is a statement of intent to offend another and use cruel or abusive language based on actual or perceived membership to another group [3]. Established a lexical baseline for discriminating between profane and hate speech on the standard dataset this is the main aim of the paper [9]. The authors adopted a linear support vector machine classifier with three groups of extracted features for these tests: word skip-grams, surface n-gram and Brown cluster.

2.4 Offensive language identification

User-generated content on social media platforms such as Twitter often includes a high level of rude, offensive or sometimes hateful language [20]. Increasing vulgarity in online discussions and user comment sections have recently been discussed as relevant issues in society as well as in science [14], and identified offensive tweets with an accuracy of 83.14 %, f1-score 0.7565 on the real test data for the classification of offensive vs non-offensive.

3 METHODOLOGY

In this paper, we focus on hate, offensive, and profane exclusively, for English. We participated in the competition using the team name IRLAB@IITBHU. Figure 1 shows the methodology of the paper.

3.1 Data

The dataset was created from Twitter, Facebook and distributed in tab-separated format. We have participated for all three sub-tasks of English language [11]. The size of training and testing data is 5852 and 1153 posts for the English language, respectively. In Sub-task A the HOF containing Hate, offensive, and profane posts are 288, and NOT not containing any Hate speech, offensive content posts are 865. In Sub-task B, HATE Hate speech posts are 124, and NONE posts are 865. Offensive posts are 71, and Profane posts are 93. In Sub-task C, NONE posts are 865, and TIN (Targeted Insult) posts are 245, and UNT (Untargeted) posts are 43.

3.2 Pre-processing

First the data were cleaned using the tweet preprocessing library[4]. We got the cleaned data after removing the Retweets Symbols (RT), Hashtag, URL's, Twitter Mentions, Emoji's and Smileys. The preprocessed data also exclude the English stop words (available in NLTK[5]) while tokenizing the sentences for the extraction of frequency-based feature extraction. The Hate speech, Offensive and Profane have been predicted through TF-IDF feature.

3.3 Classifier

We use two machine learning classifiers Support Vector Machine (SVM) and XGBoost (XGB) classifying for classification of Hate speech, Offensive and Profane. The input for both the classifier is in the form of TF-IDF feature matrix and output is a label for the categorical result. Both the classifiers give a different score, as classifiers have different specialities.

[4] https://pypi.org/project/tweet-preprocessor/

[5] https://www.nltk.org/

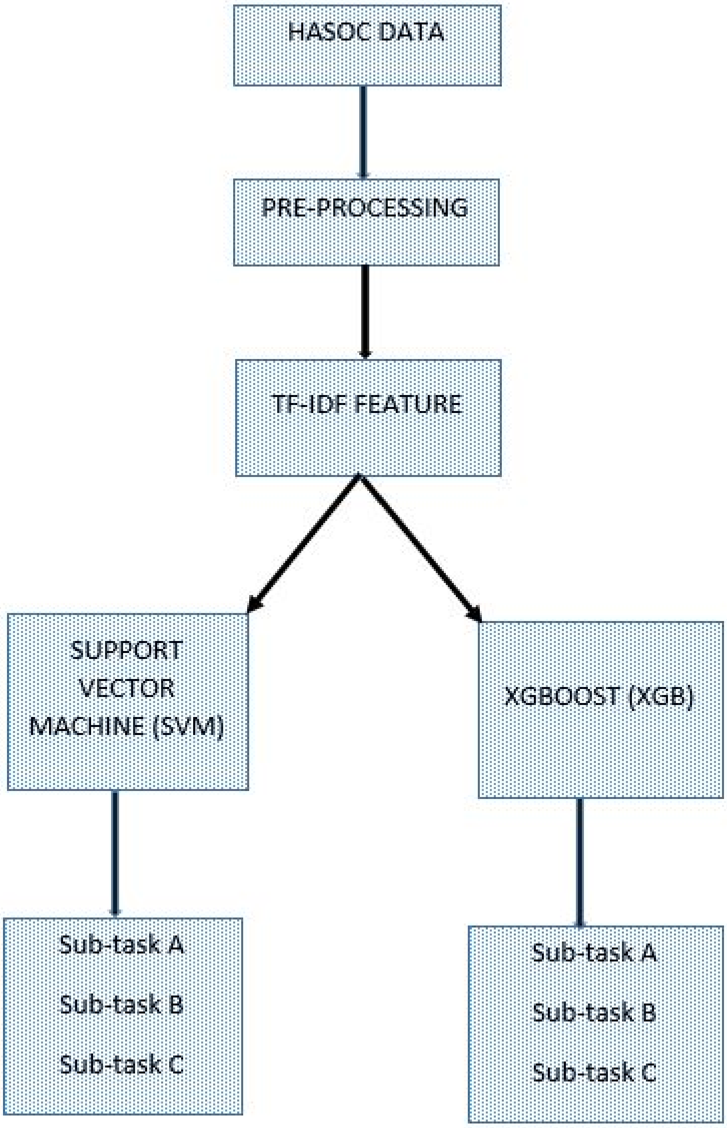

Fig. 1. Diagram of Hate speech, Offensive and Profane Classifier

4 RESULTS

We start by investigating the accuracy of our TF-IDF features based on machine learning method for this task. We first train the classifier, with each of them using a type of TF-IDF feature. The results of these experiments are listed in Table 1 and Table 2. In Sub-task A, accuracy of XGBoost is 81% better as compared to SVM 73%. The Sub-task B and Sub-task C accuracy is 80% the same for the XGBoost.

Table 1. Classifier result of HASOC dataset at Precision, Recall, F-score and Accuracy in %.

Tasks	**Model**	SVM			XGBoost		
	Labels	Precision	Recall	F_1	Precision	Recall	F_1
Sub-task A	HOF	0.47	0.65	0.54	0.69	0.41	0.51
-	NOT	0.87	0.76	0.81	0.83	0.94	0.88
Sub-task B	HATE	0.16	0.08	0.11	0.50	0.01	0.02
-	NONE	0.80	0.95	0.87	0.80	0.99	0.88
-	OFFN	0.00	0.00	0.00	0.00	0.00	0.00
-	PRFN	0.73	0.52	0.60	0.86	0.71	0.78
Sub-task C	NONE	0.85	0.80	0.83	0.83	0.96	0.89
-	TIN	0.41	0.58	0.48	0.65	0.38	0.48
-	UNT	0.00	0.00	0.00	0.00	0.00	0.00

Table 2. Classifier result on testing dataset in %.

Task	Sub-task A			Sub-task B			Sub-task C		
Model	Macro_f1	Weighted_f1	Accuracy	Macro_f1	Weighted_f1	Accuracy	Macro_f1	Weighted_f1	Accuracy
SVM	0.675	0.741	0.73	0.3949	0.7116	0.76	0.4364	0.723	0.72
XGBoost	0.6967	0.7881	0.81	0.4193	0.7283	0.80	0.4578	0.7704	080

5 CONCLUSION

In this paper we used text classification techniques to recognise among hate speech, profane and offensive posts. As a baseline we use XGBoost and a SVM classifier. The best result was achieved by XGBoost achieving 81% accuracy. The results displayed in this paper showed that identification of profanity from abusive language is a very challenging task.

References

1. Aroyehun, S.T., Gelbukh, A.: Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In: Proceedings of the

First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). pp. 90–97 (2018)
2. Association, A.P., et al.: Apa resolution on bullying among children and youth. Washington, DC: American Psychological Association pp. 1–4 (2004)
3. Britannica, E.: Britannica academic. Encyclopædia Britannica Inc (2015)
4. Dinakar, K., Jones, B., Havasi, C., Lieberman, H., Picard, R.: Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS) **2**(3), 18 (2012)
5. House, W.: Background on white house conference on bullying prevention (2011)
6. Committee on Injury, V., Prevention, P., et al.: Role of the pediatrician in youth violence prevention. Pediatrics **124**(1), 393–402 (2009)
7. Johnson, N., Leahy, R., Restrepo, N.J., Velasquez, N., Zheng, M., Manrique, P., Devkota, P., Wuchty, S.: Hidden resilience and adaptive dynamics of the global online hate ecology. Nature pp. 1–5 (2019)
8. Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M.: Benchmarking aggression identification in social media. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). pp. 1–11 (2018)
9. Malmasi, S., Zampieri, M.: Detecting hate speech in social media. arXiv preprint arXiv:1712.06427 (2017)
10. Matsuda, M.J.: Public response to racist speech: Considering the victim's story. In: Words That Wound, pp. 17–51. Routledge (2018)
11. 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)
12. Nahar, V., Li, X., Pang, C., Zhang, Y.: Cyberbullying detection based on text-stream classification. In: The 11th Australasian Data Mining Conference (AusDM 2013) (2013)
13. Orasan, C.: Aggressive language identification using word embeddings and sentiment features. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). pp. 113–119 (2018)
14. Ramakrishnan, M., Zadrozny, W., Tabari, N.: UVA wahoos at SemEval-2019 task 6: Hate speech identification using ensemble machine learning. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 806–811. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/S19-2141, https://www.aclweb.org/anthology/S19-2141
15. Reynolds, K., Kontostathis, A., Edwards, L.: Using machine learning to detect cyberbullying. In: 2011 10th International Conference on Machine learning and applications and workshops. vol. 2, pp. 241–244. IEEE (2011)
16. Risch, J., Krestel, R.: Aggression identification using deep learning and data augmentation. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). pp. 150–158 (2018)
17. Sui, J.: Understanding and fighting bullying with machine learning. Ph.D. thesis, Ph. D. dissertation, The Univ. of Wisconsin-Madison, WI, USA (2015)
18. Wiegand, M., Siegel, M., Ruppenhofer, J.: Overview of the germeval 2018 shared task on the identification of offensive language (2018)
19. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Predicting the Type and Target of Offensive Posts in Social Media. In: Proceedings of NAACL (2019)

20. 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)