A Review on Offensive Language Detection

Rahul Pradhan, Ankur Chaturvedi, Aprna Tripathi and Dilip Kumar Sharma

Abstract Offensive language, hate speech, and bullying behavior is prevalent during textual communication happening online. Users usually misuse the anonymity available online social media, use this as an advantage, and engage in behavior that is not acceptable socially in actual world. Social media platforms, analytics companies, and online communities had shown much interest and involvement in this field to cope up with this problem by stopping its propagation in social media and its usage. In this paper, we will propose the work done by researchers to form effective strategies for tackling this problem of identifying offense, aggression, and hate speech in user's textual posts, comments, microblogs, etc.

Keywords Hate speech · N-gram · Offensive language · tf-idf · Machine learning · Twitter · Offensive language detection · Antisocial behavior online

1 Introduction

Abusive and offensive language is the prime concern of technical companies nowadays due to exponential growth in number of Internet users around the world and since these people are from different walks of life and different culture. There is a fine line between hate speech and offensive language, and to detect and differentiate among them is a big challenge. In literature, researchers generally classify the text into three classes:

R. Pradhan (✉) · A. Chaturvedi · A. Tripathi · D. K. Sharma
GLA University, Mathura, India
e-mail: rahul.pradhan@gla.ac.in

A. Chaturvedi
e-mail: ankur.chaturvedi@gla.ac.in

A. Tripathi
e-mail: aprna.tripathi@gla.ac.in

D. K. Sharma
e-mail: dilip.sharma@gla.ac.in

M. L. Kolhe et al. (eds.), *Advances in Data and Information Sciences*,
Lecture Notes in Networks and Systems 94,
https://doi.org/10.1007/978-981-15-0694-9_41

- Hateful,
- Offensive, and
- Clean

In this paper, we showcase the study we perform on the research held in this area with some light on what can be done next in order to make it more efficient. Our objective behind carrying this work is to come up with a study of papers and research work done in this field so far.

2 Terminology

In this paper, we use the term hateful, offensive, and clean. We come to a conclusion in favor of the usage of these terms since they can have broader meaning and can be used in various contexts in user-generated content to define it first. Hateful text or speech is not a very common phrase to refer to such text in legal world but in general terms day-to-day speaking we use it quite often.

Following is the list of terms used in literature [1]:

- abusive messages,
- hostile messages, or
- flames.

This will help readers to go further in literature on this topic. There is a recent trend in the NLP world that author prefers to use the word cyberbullying [2–7].

Hateful speech or hate speech is commonly referred to as conversation, communication that mocks a group of people or a single person on the grounds of social status, race, color, ethnicity, nationality, gender, sexual preferences, religions, and many others [8].

3 Literature Survey

Researchers in past have proposed various machine learning approaches and their variant to deal with the problem of offensive language. Detecting sarcasm had been the point of research for many researchers around in area of NLP or text mining, with need of hour nowadays people are more focusing on detecting the wrongs prevailing in social media. This concern of government and public leads to open new research domains as fake news detection, rumor detection, offensive language detection, etc. Many of these proposed works use feature extraction from text such as bag of words (BOW) and dictionaries. Major work in this area is focused on feature extraction from text. Dictionaries [9] and bag of words [10] were among the lexical features that were used widely by researchers to detect the offensive language or phrases.

Gaydhani et al. [11] used tf-idf and N-gram as features for their classification of tweets with 95.6% accuracy.

It was found out that these features could not understand the context of sentences. Approaches that involve N-gram show better results and perform better than their counterparts [12]. Lexical features are proving to outperform other features in automatic detection of offensive language and phrases, without taking into consideration the syntactic structures as bag of word approach could not detect offensiveness if words are used in different sequences [13].

Gaydhani et al. [11] form a dataset which is the combination of three different datasets. The first dataset which they used is publicly available on Crowdflower1, which was used in [14, 15]. Dataset Crowdflower1 has tweets classified into three classes: "Hateful", "Offensive", and "Clean". All the tweets in this dataset are manually annotated. The second dataset they used is crowdflower2 having tweets manually classified into same three classes. Github3 is the third dataset they integrate with other two to build their dataset for study. This third dataset consists of two columns: tweet-ID and class. "Sexism", "Racism", and "Neither" are the three categories or classes in which each of these tweets are classified. This dataset is used by [14, 16]. They have considered logistic regression, naive Bayes, and support vector machines for text classification. They used training of dataset on each model by performing grid search for all the combinations of feature parameters and performed 10-fold cross-validation. They analyzed performance on the basis of average score of the cross-validation.

Davidson et al. [17] reduce the dimensionality of the data using a logistic regression with L1 regularization. They show a comparative study on prior work such as logistic regression, naive Bayes, decision trees, random forests, and linear SVMs. They use fivefold cross-validation, with keeping 10% of the sample for evaluation to help prevent overfitting on all the models. Their study suggests that logistic regression and linear SVM perform slightly better than other models. They further use logistic regression with L2 regularization for the final model as it has shown better result in previous work. They use tweets from Hatebase.org which contains lexicon compiled by Internet users containing words and phrases that are considered to be hate speech. Using these words from lexicon they crawled the twitter using the Twitter API which collects tweets containing these words. They collect 33,458 user's tweets as sample. They get these tweets annotated by CrowdFlower workers into three categories: hate speech, offensive but not hate speech, or neither offensive nor hate speech. Getting these manually annotated helps in clear tagging as they not just look for words but also context of tweets. They found majority of the tweets fall into category of offensive language. They use features from these tweets and used them to train a classifier.

Lee et al. [18] use the dataset titled "Hate and Abusive Speech on Twitter" [19] recently released. This dataset contains the tweets classified into four categories, namely, "normal", "spam", "hateful", and "abusive".

70 character dimensions using 26 lower character dimensions were used to convert the tweets into one hot encoded vector with 10 digits and special characters up to 34 including whitescape. This encoding is used for character-level representation.

Table 1 Distribution of categories among tweets

Categories	Normal	Spam	Hateful	Abusive
Number	42,932	9,757	3,100	15,115
(%)	(60.5)	(13.8)	(4.4)	(21.3)

Before this encoding, they have removed user ID, emojis, and URLS, and replace them by special tokens.

Table 1 shows the distribution of tweets among four categories, which are discussed below:

Aken et al. [20] consider two datasets to evaluate their proposed algorithm: one of the datasets they pick from Kaggle's second challenge on toxic comment classification which contains comments on Wikipedia talk pages presented by Google Jigsaw and other datasets they consider are based on Twitter by Davidson et al. [21]. Class distribution of both datasets is shown in Tables 2 and 3. 24,783 tweets were extracted from Twitter which constitute to dataset of Davidson et al. [21], and all these tweets were annotated by CrowdFlower workers with the labels "hate speech", "offensive but not hate speech", and "neither offensive nor hate speech".

They propose an ensemble to figure out that a single classifier is most effective on certain kind of comment. The ensemble classifier analyzes the features from comments, weights, and for a given feature combination it identifies the suitable single classifier. To attain the goal of identifying the classifier using gradient boosting decision tree, they perform validation across the average final predictions on five trained models.

The most valuable contribution by Aken et al. [20] is Error Classes of False Negatives they have defined. These classes are as such Doubtful labels, and these are the labels that cannot be clearly identified as toxic because for a particular user it is toxic but there are users or annotators that consider it as nontoxic. Second class of false-negative error is Tweets that contain toxicity without any kind of hate words or swear words that this class of error needs to overcome which will require investigating some semantic embeddings for obtaining better classification on different paradigmatic contexts. Third class of error identified by the author is Rhetorical

Table 2 Wikipedia comment dataset

Categories	Clean	Toxic	Obscene	Insult	Identity hate	Severe toxic	Threat
Number	2,01,081	21,384	12,140	11,304	2,117	1,926	689
%	80.23	8.53	4.84	4.51	0.84	0.77	0.27

Table 3 Twitter dataset

Categories	Offensive	Clean	Hate
Number	19,190	4,163	1,430
%	77	17	6

Questions and these are the kind of text sentences that does not contain any toxic words but have sarcastic questions in it, usually such text contains question marks and question words. Other classes they introduced are Metaphors and comparisons, and idiom that can be twisted in meaning by looking at context which are difficult to see in short text and such text usually requires knowledge about the implications of language or some additional contextual knowledge. Aken et al. [20] find that different approaches fail in identifying different texts and make errors, but this can be combined into an ensemble with F1-measure. They find some combination of shallow learners with deep neural networks showing remarkable results and proved it to be very effective.

Mathur et al. [22] explore the usage of mixed language in their work and identify the offensive text or hate speech. They choose Hinglish as their subject because of its ease in communication and being popular on Twitter due to its reachability to larger audience in native language. They faced difficulty as this mix of two languages has inherent variations of spellings and absence of grammar induces considerable amount of ambiguity to text and makes the problem even harder to disambiguate and understand the true meaning of text. They proposed the multi-input multichannel transfer learning (MIMCT)-based model is used to identify and detect the hate speeches and offensive language in Hinglish tweets. They use the dataset proposed by them and named it as Hinglish Offensive Tweet (HOT) dataset. Their proposed learning model uses multiple feature inputs using transfer learning. They employed word embedding with secondary extracted features as input to train their multichannel CNN LSTM which is pretrained on English tweets.

Table 4 shows the distribution of tweets among different classes in HOT dataset.

Pitsilis et al. [23] address the effectiveness of identifying the class (being offensive or not offensive) of new tweet or post, using the identity and history of user who has posted the tweet and other tweets posted by him or by other user related to him. They use LSTM for classification and classify the tweets into three classes, namely, neutral, racism, and sexism. The dataset they used is proposed by Waseem et al. [24] and contains about 16,000 short messages collected across Twitter (Table 5).

The biggest issue with this dataset is of dual labeled tweets in the dataset. The number of these tweets is not that small that they can ignore them. Being more precisely, there are 42 tweets that are annotated as both “Neutral” and “Sexism”,

Table 4 Hinglish offensive tweet (HOT) dataset

Categories	Non-offensive	Abusive	Hate inducing
Number	1121	1765	303
%	35.15	55.35	9.5

Table 5 Waseem Twitter dataset

Categories	Racism	Sexism	Neutral
Number	1943	3166	10,889
%	12.15	19.79	68.06

while 06 tweets were classified as “Racism” and “Neutral” both. According to the dataset providers, the labeling was performed manually.

Wiedemann et al. [25] explore different techniques for automatic detection of offensive text or hate speech on Tweets written in German language. They also employ deep learning for this task and use a series BiLSTM and CNN neural network in sequence. They improve the accuracy of three learning transfer task for improving the classification performance using context and historical data. They compare supervised categories such as near offensive to weakly supervised categories that contain emojis, and they also show comparison to unsupervised category using tweets of same topic by clustering them using latent Dirichlet allocation (LDA).

4 Conclusion

In this paper, we try to present the work done recently in this field of automatic detection of offensive language. We show that how research goes from using tf-idf to popular classifiers such as naïve Bayes, support vector machine (SVM), logistic regression, and then research work goes to variant of these classifiers such as linear SVM, logistic regression with L2, and from here researchers further explore ensemble classifiers using the combination of these classifiers by decomposing the task into subtasks, and then lastly the usage of deep learning and we found many researchers using approaches such as LSTM, CNN, and RNN. Each of these techniques has their own advantages and for classification accuracy LSTM models have outperformed others.

References

1. Spertus, E. (1997). Smokey: Automatic recognition of hostile messages. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence, AAAI'97/IAAI'97* (pp. 1058–1065). Providence, RI, USA: AAAI Press.
2. Xu, J.-M., Jun, K.-S., Zhu, X., & Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 656–666). Montreal, Canada: Association for Computational Linguistics.
3. Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Detection of cyberbullying incidents on the instagram social network. abs/1503.03909.
4. Zhong, H., Li, H., Squicciarini, A. C., Rajtmajer, S. M., Griffin, C., Miller, D. J., & Caragea, C. (2016). Content-driven detection of cyberbullying on the instagram social network. In *IJCAI* (pp. 3952–3958). New York City, NY, USA: IJCAI/AAAI Press.
5. Hee, C. V., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., & Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. In *Proceedings of Recent Advances in Natural Language Processing, Proceedings*, Hissar, Bulgaria (pp. 672–680).

6. Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). Improving cyberbullying detection with user context. In *Proceedings of the European Conference in Information Retrieval (ECIR)*, Moscow, Russia (pp. 693–696).
7. Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems*, *2*(3), 18:1–18:30.
8. Nockleby, J. T. (2000). Hate speech. In L. W. Levy, K. L. Karst, & D. J. Mahoney (Eds.), *Encyclopedia of the American constitution* (pp. 1277–1279, 2nd ed.). Macmillan.
9. Liu, S., & Forss, T. (2015). New classification models for detecting Hate and Violence web content. In *2015 7th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3K)*, Lisbon (pp. 487–495).
10. Burnap, P., & Williams, M. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1).
11. Gaydhani, A., Doma, V., Kendre, S., & Bhagwat, L. (2018). Detecting hate speech and offensive language on twitter using machine learning: An N-gram and TFIDF based approach. arXiv:1809.08651.
12. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web-WWW'16*.
13. Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *2012 international conference on privacy, security, risk and trust and 2012 international conference on social computing*.
14. Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access, 6,* 13825–13835.
15. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *International AAAI conference on web and social media*.
16. Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*.
17. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *ICWSM 2017*.
18. Lee, Y., Yoon, S., & Jung, K. (2018). Comparative studies of detecting abusive language on twitter. arXiv:1808.10245.
19. Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*.
20. van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. arXiv:1809.07572.
21. Davidson, T., Warmsley, D., Macy, M. W., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *ICWSM 2017*.
22. Mathur, P., Sawhney, R., Ayyar, M., & Shah, R. (2018). Did you offend me? Classification of offensive tweets in hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)* (pp. 138–148).
23. Pitsilis, G.K., Ramampiaro, H., & Langseth, H. (2018). Detecting offensive language in tweets using deep learning. arXiv:1801.04433.
24. Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop, San Diego, California, June 2016*. Association for Computational Linguistics.
25. Wiedemann, G., Ruppert, E., Jindal, R., & Biemann, C. (2018). Transfer learning from LDA to BiLSTM-CNN for offensive language detection in twitter. arXiv:1811.02906.