Online Hate Classifier

Aneri Kapadia
Computer Science Department
Nirma University
Ahmedabad,India
18bce013@nirmauni.ac.in

Jaineet Shah
Computer Science Department
Nirma University
Ahmedabad,India
18bce083@nirmauni.ac.in

***Abstract*—The rapid growth of social media has empowered people to express their opinions extensively on the internet. However this freedom is highly misused in the from of hate speech and offensive/foul language. Therefore there is an urgent need to deploy effective online hate detection and classification models on various social media platforms in order to curb unwelcoming and abusive comments on the internet. To address this issue, we have utilised a tweets based dataset, pre-processed it and applied various machine learning algorithms to produce an effective online hate classifier.**

Keywords—online hate, machine learning, natural language processing.

I. INTRODUCTION

A. Motivation

Online hate which is otherwise known as abusive language [1], aggression [2], cyberbullying [5], derogatory comments, insults [6], racism, sexism etc has been recognised as an extremely serious threat on social media platforms.Even though the platforms provide means of flagging hateful and and abusive content, the feature is significantly ineffective and under-utilised among the users. Manual methods like flagging are both ineffective and unscalable and also have a risk of partiality under subjective judgments by human annotators. Since an automated mechanism is faster and more efficient than human annotation, machine learning and deep learning models to automatically detect and classify online abuse have been gaining popularity. Therefore in this paper we have attempted to apply different machine learning models on a tweets based dataset to check the improvement and accuracy of such models in classifying online hate with respect to other conventional methods.

B. Research Contribution

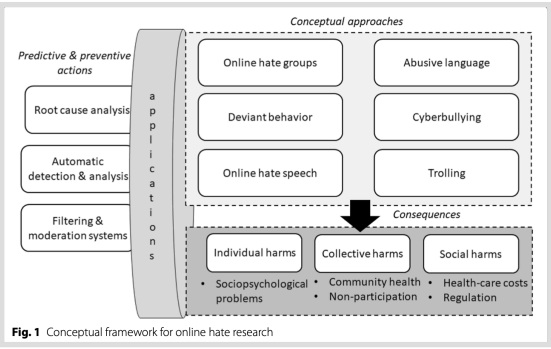


Fig. 1 Conceptual framework for online hate research

Figure - 1

1) Theoretical Foundation of Online hate -

Online hate can be considered as cross-disciplinary as it has been researched and studied using multiple theoretical frameworks and lenses, including social psychology, human–computer interaction, politics, and legislation/regulatory aspects.

Figure 1 displays a conceptual framework of the pivotal areas in the online hate research which highlights the complex dynamics of online hate; thus making its automatic classification and detection more complicated.

First, online hate is considered to be the use of abusive, profane or offensive language [4]. These studies concentrate more on the linguistic aspects of online hate such as linguistic styles, vocabularies and means of expression. Some of these studies tend to focus on counter-speech, which refers to the methods of nullifying the hateful comments with linguistic strategies.

Second, some research deals particularly with online hate as hate speech, i.e., offensive post, motivated, in whole or in part, by the writer's bias against an aspect of a group of people. The pivotal dimension is directing the hate towards a specific target such as refugees, women, caste, religion etc.

Third, another significant part of the online hate research is taking into account the group dynamics, which can be noticed in the studies dealing with online hate groups and group prejudice, persuasive storytelling as hate conditioning, radicalisation via extremist content, cultural transmission of hate, social exclusion and so on. Since these aspects of online hate possess a high degree of subjective and contextual factors, they are often researched using interpretative methods.

Fourth, some studies deal with the consequences of online hate, i.e., its effects on an individual or a group of people. This research often involves a predictive machine learning aspect for the detection and classification of toxicity in social media platforms and certain communities. The focal attribute of the toxicity studies is that they consider online hate as not only the use of language but also as an act having a solid outcome or effect. These outcomes include, a user quitting a toxic discussion, reduced participation in social media, radicalisation, group polarisation where prejudices held in the past are enforced, degraded heath of an online community and so on. The computer science research in this field tends to focus on automating the online hate classification and detection.

2) Progression of online hate detection -

The progression of online hate detection can be partitioned into three temporal stages: (1) simple lexicon or keyword-based classifiers, (2) classifiers utilising distributed

semantics, (3) deep learning classifiers with advanced linguistic features.

Keyword-based classifiers: these classifiers involve a set of hateful words which are known as keywords. These keywords are utilised in the classification of online hate by matching their occurrence in the sentences. However, this approach has well known limitations. The primary issue is that the linguistic diversity of online hate cannot be captured completely by a dictionary, i.e., keywords cannot identify sarcasm or forms of humour. Also the dictionary of hateful words and offensive content needs to be constantly updated as new terminology and slangs rapidly evolve on social media. Moreover, different communities have different standards, i.e., what is interpreted as hate by one community may be considered as humorous or normal discourse in another community. One more problem with the keyword matching approach is polysemy, i.e., same word possesses several different meanings which is also known as word-sense disambiguation in Natural Language Processing (NLP).

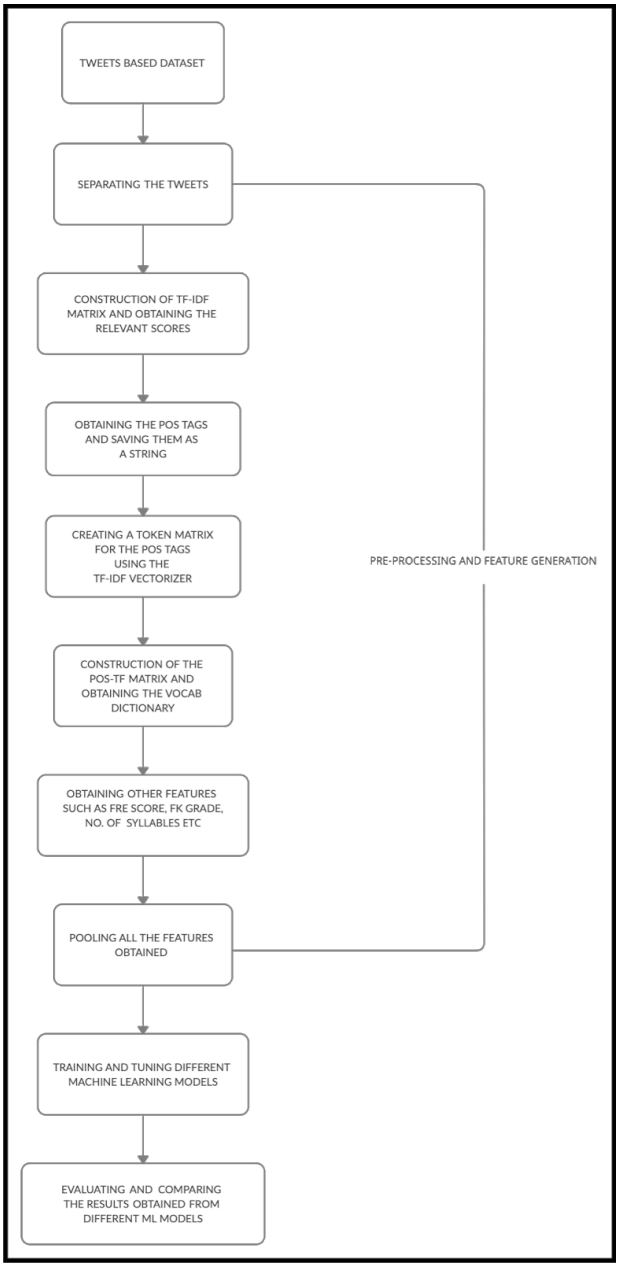
Distributional semantics: these classifiers involve the deployment of a wide range of more sophisticated feature representations such as n-grams, syntactic features and distributional semantics , i.e., word embeddings and vector space models. TF-IDF (term frequency - inverted document frequency) model, Bag of Words (BOW) model, Labeled Latent Dirichlet Allocation (LLDA) models are some examples of distributional semantics classifiers. Overall, distributional semantics do a better job of classification and detection of online hate than keyword classifiers.

Deep learning classifiers: these classifiers are the most recent and upcoming in the online hate classification and detection research. These include variants of recurrent neural network (RNN) [9] architecture, convolutional neural network (CNN) [10] architecture or their combination and generate state of the art results. Most of the research in this classification area is done with text features but there are studies also utilising other features such as user features [7] and knowledge graphs [8]. These extra features boost the performance of online hate detection and classification but at the same time, they are extremely rare and scarce.

3) Research Gaps -

Even though the domain of online hate has been thoroughly researched previously, the concept of cross-platform evaluation of online hate classifiers is scarcely discussed in articles despite the well known fact that online hate is not tied to a single platform or context. The absence of cross-platform evaluation limits the generalisation of models built on datasets from a particular online platform solely to that platform. More research efforts are required for developing cross-platform classifiers which can be applied universally.

II. SYSTEM MODEL



III. PROPOSED APPROACH

Many social networks have developed user laws forbidding hate speech; however, implementing these regulations necessitates a lot of manual labour to go through each article. Facebook, for example, recently expanded the number of content moderators. Automatic tools and methods may help speed up the evaluation process or better assign human resources to positions that need close attention.In this section, we'll look at some automated technique for identifying hate speech in text [11].

Sr.No.	Concept	Acronym	Definition	References

1	Feature Extraction.	FE	It's a process of dimensionality reduction by which the raw data is reduced to more manageable groups for processing. It is mapping from text data to real-valued vectors.	[12]
2.	Term Frequency - Inverse Document Frequency.	TF-IDF	It's a feature representation technique for a text in the document collection that reflects significance of a word." It is based on a combination of the frequency in which a word appears in a document and the number of documents that include that word.	[13]
3.	Machine Learning Classifiers.	ML Classifier	These are implemented to numeric features vectors to build the predictive model that is used to determine class labels.	[14]
4.	Logistic Regression.	LR	It uses a log sigmoid function to show the relationship between one independent variable and one or more independent variables	[15]
5.	Decision Tree.	DT	It is a supervised algorithm. It generates the classification rules in the tree-shaped form, where each internal node denotes attribute conditions, each branch denotes conditions for outcome and leaf node represents the class label.	[16]
6.	Artificial Neural Networks.	ANN	ANN is a modeling technique inspired by the human nervous system that allows learning by example from representative data that describes a physical phenomenon or a decision process	[17] [24]
7.	Support Vector Machines.	SVM	It's a supervised classification algorithm which constructs an optimal hyperplane by learning from training data which separates the categories while classifying new data.	[18]
8.	K Nearest Neighbour	KNN	Classification algorithm, which classifies the new data points based on some similarity measure(nearest neighbors) by comparing it with the existing data.	[19]

IV. PERFORMANCE EVALUATION

A. Dataset Description

The data are stored as a CSV and as a pickled pandas data-frame (Python 2.7). Each data file contains 5 columns:
count = number of CrowdFlower users who coded each tweet (min is 3, sometimes more users coded a tweet when judgments were determined to be unreliable by CF).
hate_speech = number of CF users who judged the tweet to be hate speech.
offensive_language = number of CF users who judged the tweet to be offensive.
neither = number of CF users who judged the tweet to be neither offensive nor non-offensive.
class = class label for majority of CF users. 0 - hate speech 1 - offensive language 2 - neither

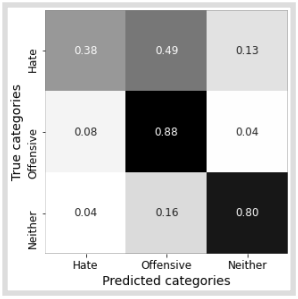
	Class	**Total Instances**	**Training Instances**	**Testing Instances**
1	Hate Speech	1430	1266	164
2.	Offensive	19190	17285	1905
3.	Neither	4163	3753	410
4.	Total	24783	22304	2479

B. Preprocessing

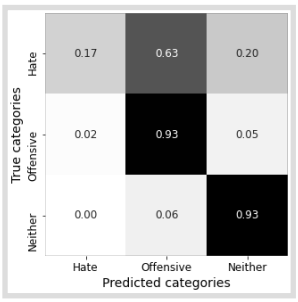
In preprocessing, tweets are converted to lowercase. All the URLs are replaced with 'URL HERE'', multiple white spaces are replaced with single instances, hashtags and mentions are eliminated. Further, the **tokenize function** removes punctuations, stop-words and excess white space. **Porter Stemmer** is used that converts words into their stem or root form [25].

C. *Evaluation*

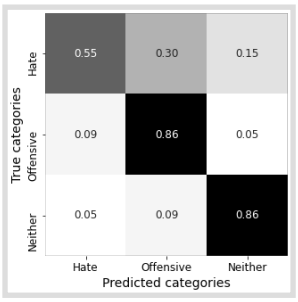
Logistic Regression:



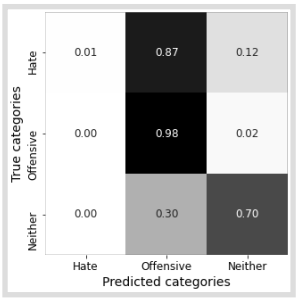
Decision Tree:



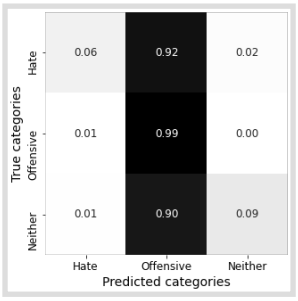
ANN (Artificial Neural Network):



SVM (Support Vector Machine):



KNN (K Nearest Neighbours):



V. RESULTS

In this section we can see the overall results of five analyses. The following table shows the Accuracy, Precision, Recall and F1 score of all the five approaches. In the above confusion matrices given for each of the five machine learning algorithms. The bold values represented are the maximum and minimum result values [automatic hate speech detection].

In all the analysis, the lowest precision (0.76), recall(0.78), accuracy(0.78) and f measure(0.71) found in KNN classifier using tf-idf features. Similarly, the highest accuracy (0.88), precision (0.87), recall(0.88), and F1 score(0.87) are obtained using the Decision Tree algorithm.

	Accuracy	**Precision**	**Recall**	**F1 Score**
DT	0.88	0.87	0.88	0.87
ANN	0.84	0.87	0.84	0.85
SVM	0.87	0.87	0.87	0.83
KNN	0.78	0.76	0.78	0.71
LR	0.84	0.85	0.84	0.84

VI. CONCLUSION

In this paper, we have explored the different text classification methods by comparing five ML algorithms - LR, SVM,ANN,KNN and Decision Tree. We have attempted to tune the hyper-parameters for each of these models in such a way that each one of them gives the best possible result. On comparing and contrasting, we can conclude that the lowest performance in terms of the various parameters (accuracy, precision, recall) is observed in the case of KNN and the best performance is given by decision tree algorithm for this dataset. The result from this study is significant for comparing the future research between various text classification techniques for automatic hate speech detection. The research can be extended for identifying the severity of the messages in place of a fixed number classes - hate, offensive and neither. Thus, it is an important consideration for future research. The model can be further improved by using lexicon based techniques, comparing with other state-of-the-art results and collecting more data instances [21].

REFERENCES

1. Joni Salminen, Maximilian Hopf, Shammur A. Chowdhury, Soon-gyo Jung, Hind Almerikhi, Bernard J. Jansen; Developing an online hate classifier for multiple social media platforms; 2020.
2. Castelle M. The linguistic ideologies of deep abusive language classification. In: Proceedings of the 2nd workshop on abusive language online (ALW2), Brussels; 2018. P. 160–70.
3. Wachs S et al (2019) Understanding the overlap between cyberbullying and cyberhate perpetration: moderating effects of toxic online disinhibition. Crim Behav Mental Health 29(3):179–188. https://doi.org/10.1002/cbm.2116.
4. Lee H-S et al (2018) An abusive text detection system based on enhanced abusive and non-abusive word lists. Decis Support Syst..
5. Kumar S, et al. Community interaction and conflict on the web. In: Proceedings of the 2018 world wide web conference on world wide web; 2018. P. 933–43
6. Salminen J, et al. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online

news media. In: Proceedings of the international AAAI conference on web and social media (ICWSM 2018), San Francisco; 2018

7. Unsvag EF, Gambäck B. The effects of user features on twitter hate speech detection.
8. Jafarpour B, Matwin S. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs.
9. Karan M, Šnajder J. Cross−domain detection of abusive language online. In: Proceedings of the 2nd workshop on abusive language online (ALW2); 2018. P. 132–137
10. Zhang Z et al (2018) Detecting hate speech on twitter using a convolution−gru based deep neural network.
11. Hate speech detection: Challenges and solutions Sean MacAvaney ,Hao-Ren Yao,Eugene Yang,Katina Russell,Nazli Goharian,Ophir Frieder Published: August 20, 2019 https://doi.org/10.1371/journal.pone.0221152
12. Mujtaba, G., et al., Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study. Journal of forensic and legal medicine, 2018. 57: p. 41-50
13. Ramos, J. Using tf-idf to determine word relevance in document queries. in Proceedings of the first instructional conference on machine learning. 2003. Piscataway, NJ.
14. Kotsiantis, S.B., I.D. Zaharakis, and P.E. Pintelas, Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 2006. 26(3): p. 159-190
15. Wenando, F.A., T.B. Adji, and I. Ardiyanto, Text classification to detect student level of understanding in prior knowledge activation process. Advanced Science Letters, 2017. 23(3): p. 2285-2287.
16. Abacha, A.B., et al., Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug–drug interaction extraction and classification. Journal of biomedical informatics, 2015. 58: p. 122- 132.
17. Received 2 September 2020, Revised 27 September 2020, Accepted 28 October 2020, Available online 1 November 2020.
18. Joachims, T. Text categorization with support vector machines: Learning with many relevant features. in European conference on machine learning. 1998. Springer
19. Zhang, M.-L. and Z.-H. Zhou, A k-nearest neighbor based algorithm for multi-label classification. GrC, 2005. 5: p. 718-721.
20. Shaikh, S. and S.M. Doudpotta, Aspects Based Opinion Mining for Teacher and Course Evaluation. Sukkur IBA Journal of Computing and Mathematical Sciences, 2019. 3(1): p. 34-43
21. Shibly F.H.A., Sharma U., Naleer H.M.M. (2021) Classifying and Measuring Hate Speech in Twitter Using Topic Classifier of Sentiment Analysis. In: Gupta D., Khanna A., Bhattacharyya S., Hassanien A.E., Anand S., Jaiswal A. (eds) International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol 1165. Springer, Singapore. https://doi.org/10.1007/978-981-15-5113-0_54
22. P Vijayaraghavan, H Larochelle, D Roy - arXiv preprint arXiv:2103.01616, 2021 - arxiv.org
23. N. Hettiarachchi, R. Weerasinghe and R. Pushpanda, "Detecting Hate Speech in Social Media Articles in Romanized Sinhala," 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), 2020, pp. 250-255, doi: 10.1109/ICTer51097.2020.9325465.
24. Y. Zhou, Y. Yang, H. Liu, X. Liu and N. Savage, "Deep Learning Based Fusion Approach for Hate Speech Detection," in IEEE Access, vol. 8, pp. 128923-128929, 2020, doi: 10.1109/ACCESS.2020.3009244.
25. IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 8, 2020