Aggression In Social Media: Detection Using Machine Learning Algorithms

Chayan Paul, Deepak Sahoo, Pronami Bora

Abstract: Social media have found a remarkable jump in the number of users and their popularity in the last decade. The users of these social media platforms are found to express their opinion, views on different diverse topics. The discussion may be on a simple opinion regarding a particular product or opinion for a social issue. It might also be someone's political view or view on some religious issue. At some point of time these discussions may enter into controversial topics and users may engage in some very provocative discussion in the social media platforms. For some considerable amount of time these issues have become common in social media. Users become aggressive at time in their opinion expressed in their posts. The aggressions in social media sometimes lead to disturbances in the social equilibrium. Many a time the situation goes so wrong that it disturbs the law and order situation may also lead to loss of life and public properties. Thus detection and control of these aggressions in social media websites is an important issue. In this paper we endeavor to make a systematic survey of various research works done in the area of detection of aggression in social media sites.

Index Terms: NLP, Sentiment Analysis, Text Processing, Text Analysis, Social Network Analysis, Aggression in Social Networks

1 INTRODUCTION

The growth in the popularity and number of users for various social networking sites has been enormous in the last half decade. These sites became a crucial source of information exchange and dissemination for different individuals, groups, political parties, celebrities and many more. People registered in these sites pass a considerable amount of time of their life for accessing the information shared in these sites. These social networking sites have provided a new dimension to the way people interact with each other. These sites made a considerable impact on various issues like business, politics, education, entertainment and many more. These sites assist different groups of people like customers and service providers to exchange information and strengthen associations (Tang, Zhang, & Philip , 2014). People from diverse back ground utilize these social networking sites for forming new associations (Kapoor, Tamilmani, Rana, Patil, Dwivedi, & Nerur, 2017). Political parties utilize social media sites for personal and targeted campaigns based on users' information to reach out maximum possible people (Kruikemeier, Sezgin, & Sophie , 2016). Information shared by students and prospective candidates in the social media websites can be a criteria to filter the candidates for a particular job role (Christofides,, Muise, & Desmarais, 2009).

These social media sites are massively populated with user generated contents, photos and videos. These contents are highly influential and have high impact in various fields such as consumer affairs, politics, and social activities (Greenwood & Gopal, 2015). The growth in the online platforms has been enormous and needless to mention that a major part of this growth is in the area of social networking sites. Since the number of people contributing to these social networks increased, incidents of different types of aggression like hate speech, cyber bullying etc. also increased (Kumar, Reganti, & Bhatia, 2018). The growth in the number of users in the social networking sites has been enormous. The sites are primarily popular among the young population. There are a large number of celebrities, athletes, politicians present in these sites. Users who are followers of these celebrities get a chance to directly get in touch with their favorite celebrities. Other possible reasons are the facilities of instant spread of messages these sites provide with. These sites also provide the users with the facility to showcase their achievements, success stories, relationships etc. Users also have the facilities to share photographs and short stories about friends, entertaining videos and gifs, coupons and discounts for marketing news stories, music videos from their favorite singers / celebrities. The contents listed here are only a few possibilities. There are specialized social networking sites for different purposes. For example, linkedin is one of the social networks where professionals get connected with the various people in and around their profession throughout the world. Another social media tiktok became very popular in recent time, which allows the users to record short videos and post it. The following graph shows the growth of different social networking sites (Ortiz-Ospina, 2019)

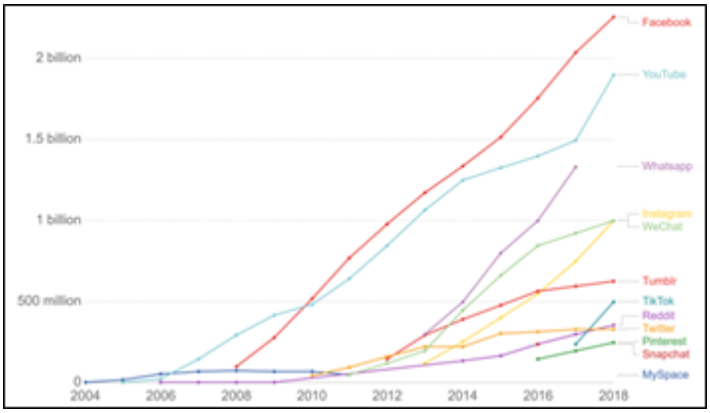

Fig 1: *Displaying growth in the number of users in social networking sites*

India has grown as country where there are large number of users in the fields of social networking sites. One of the primary reasons for that is the large amount of young population in India. As far as Facebook is concerned, India has the largest number of users registered throughout the world. The following figure shows the country wise numbers of

- *Chayan Paul, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.*
- *Deepak Sahoo, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.*
- *Pronami Bora, Department of ECE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.*

users registered in Facebook. The number of users in the graph is presented in millions (Leading countries based on number of Facebook users as of October 2019(in millions) 2020). As it is evident from the figure, India is a clear leader in this field.

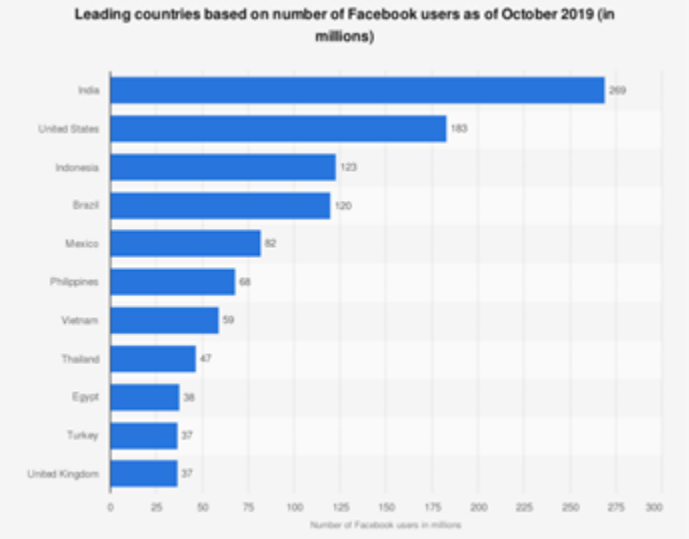

Fig 2: *Country wise number of users registered in Facebook*

The social networking sites provide the users with a platform, where they are allowed to hide their real identity. Users' have the option to choose how much information about themselves they really want to display. So there are a large chunk of the users who intentionally hide their whole personal information or a part of their personal information. This option leads to a huge numbers of fake ids in these social networking sites. As far as these fake users are concerned, the social networking sites do not have to worry a lot about them. This is primarily because, they count the total number of registered users in their networks and higher the number, higher I the gain for the networks. If we only consider Facebook, it has a large number of fake users. The following figure shows the number of fake users, Facebook reported in their annual report till the year 2018 (Nicas 2019).

Fig 3: *Bar chart showing number of fake users, Facebook has reported till the end of 2018.*

Mostly these are the users in social media networks who involve in exchange of derogatory contents. Since their identity is not disclosed, so it is difficult to track their original existence. And they take this advantage to engage themselves in cyber bullying, trolling, hate speech etc. Although most of the prominent social networks have options to report these fake accounts but in reality the time lag taken for reporting and managing the fake accounts is more than what is expected. Hence a huge chunk of the content in the social media is full of superfluous collection of texts, photographs, audio, video which propagate a lot of anti-social messages. This gives the authorities and local governments a tough time to maintain law and order situation. Since these messages are widely circulated in the social media applications, it can impact heavily in crating riot like situation in very less times. The local authorities get really less time to react which leads to a huge loss to public and personal properties. The other side of the story is, it can damage the image of any individual, group, establishment in very less amount of time. Thus we need tools and techniques in place to combat these situations. Once these aggressive contents are detected, it can be reported to the authorities, which allows them to combat these situations really fast.

3 SURVEY OF LITERATURE:

There has been a considerable amount of research in the field of detecting hate speech, cyber bullying etc. With the advent in the algorithms in the fields of Machine Learning and AI, the researchers are equipped with the right set of computational methods, which can be used to detect and report aggressive contents in Social Media Networks. Detection of aggression or hate speech is predominantly a classification problem. To solve this we need good data sets which are labeled. Kumar et. el. (Kumar, Reganti and Bhatia 2018) used Facebook and twitter data to develop an aggression tag set and annotated corpus which is Hindi and English code mixed data. In their paper the annotated the corpus using hierarchical tag set of 3 top level tags and 10 level 2 tags. They released the final dataset for research fraternity, which contains approximately 18 Thousand tweets and 21 thousand Facebook comments. Another are of this study can be calculating the amount of hate people propagate through their messages in the social media networks. In other words we can try to quantify the amount of hate. Mandal et el (Mandal, Silva and Benevenuto 2017) developed a methodology for quantifying the amount of hate in hate speeches in social media networks. They used Twitter and Whisper for collecting their data and validated their methodology using the datasets. They claimed their study can detect hate speeches and also facilitates prevention. In many of the situations the hate speech or the aggression is directed towards a specific individual or a group of individuals. Sometimes the aggression is totally generalized. ElSherief et el (ElSherief, et al. 2018) found that the directed hate speeches are more aggressive and angrier. On the other hand the generalized aggression is more related to religion and more lethal words like murder and kill are used frequently. An ideal situation would be process or a system that is able to detect the aggressive contents automatically. The options available are to develop a system based on lexical analysis or develop a system based on supervised learning. Davidson et. el (Davidson, et al. 2017) found the lexical analysis method of low precision and supervised learning process failed to classify the categories. They proposed a third method where they collected the data through crowd sourcing and trained a multiclass classifier. Social media is a platform where a lot of people propagate hate speech, aggression and discrimination. Users in many cases, use this for their political gain. Ben-David et. el (BEN-DAVID and MATAMOROS-FERNÁNDEZ 2016) questioned the policy of social networks website in controlling the propagation of aggressive contents in social media platforms. They supported their argument with a

longitudinal multimodal content and network analysis of political parties of Spain in a specific time frame. There can be several types of aggression propagated in social networking sites. It could be racial abuse, sexual abuse, gender biased or many more. Sometimes it is important to classify the types of the aggressive contents into different types to better understand the nature of the aggression. Silva et el (Silva, et al. 2016) provided a systematic classification of different hate speech and aggressive contents in the social media websites. They considered two social networking sites i.e., Twitter and Whisper. Based on their studies, the classified the aggressive contents in ten different types. In the recent past, deep learning methods have been used efficiently for in the field of data science. Natural Language processing is one of the areas where, deep learning can open a lot of possibilities. Gambäck et el ((Gambäck and Sikdar 2017) introduced deep learning in classifying the aggressive contents into four different types. These are the important works done in the area of identifying aggressive contents and hate speech. In the nest section we will see some important algorithms used in the area detecting aggressive contents and hate speech.

4 METHODS FOR DETECTING AGGRESSION:

Detection of aggressive contents can be thought of as a classification problem in the broader area of natural language processing. So basically the process of detecting the aggression can be approached in two different ways, one is rule based classification and the other one is machine learning based classification algorithms.

2.1 Lexicon Based Approaches

In rule based process the classification is done based on some handcrafted linguistic rule. This approach is also known as lexical approach. There are a considerable amount of works have been done in the classification of hate speech and aggression detection ((Martins, et al. 2918), (Gitari, et al. 2015)). In rule based systems, one needs to prepare a set of list of words. The number of lists should be equal to the number of categories. Then on receiving a new document, it can be classified as the type from which the document has the maximum number of words. The advantage of this process is, it is comprehensible and can be improved over time. But the major disadvantage with this approach is, it lacks speed and scalability is always an issue. In the era of social media networks, where the data received are characterized by the properties of big data, if an approach cannot provide speed and scalability, the approach becomes one of those which are preferred least.

2.2 Machine Learning algorithms:

Classification algorithms in machine learning have outperformed the classical lexicon based approaches. The algorithms in this section are found to be faster and more scalable. Algorithms like Naïve Bayes, Decision Trees, logistic regression, random forests, Support Vector Machines, and deep learning models like LSTM (Kwok and Wang 2013), (Magerman 1995), (Gao and Huang 2017), (Fauzi and Anny 2018), (SVM) (MacAvaney, et al. 2019) (Vigna, et al. 2017) have been extensively used in identifying aggressive contents in different social media websites. In the section below we will see some brief description of the algorithms and how they are applicable to the studies related to detection of aggressive contents: Naïve Bayes Classifier: This is a classification algorithm based on probabilities. This algorithm is mainly used for text classification or in any classification problem where the features are categorical. Like other supervised learning algorithms Naïve Bayes algorithm also needs the labels be present in the datasets. This algorithm computes the conditional probabilities of each class values and depending on the highest values of the conditional probability, it predicts the class. Decision Trees: Decision Trees are supervised learning algorithms which can be used for both classification as well as regression. Decision trees are constructed by splitting the data set based on different values like entropy and information gain. This algorithms is a good approach which can be used for predicting class values where most of the features are categorical. Logistic Regression: Logistic regression is one of the widely used classification algorithm used for binary classification problems. Unlike Naïve Bayes and Decision Tree algorithms, logistic regression works well when the features are continuous. Random Forest: Random forest is an algorithm, also belongs to the class of ensemble methods, which is widely used for classification problems. One of the main advantages of random forest is its ability to work efficiently with high dimensional noisy data. As far as the performance of the algorithm is concerned, it is one of the best algorithms in classification. But since it is an ensemble method, human interpretability is much less compared to individual decision trees. Also, training is expensive both in terms of memory as well as computational power. Support Vector Machine (SVMs): Support Vector Machine is a widely used algorithm for supervised learning. Though it can be used for regression and classification both, most of the applications are found in classification only. The main principle that works behind the idea of support vectors is, to find a hyper plane that can separate the classes in the dataset. Many a time when the data is not possible to separate into classes, the data needs to be transformed into higher or lower levels, and the mathematical functions facilitating this process is known as kernel in support vector machines. Some popular kernel functions used are, linear, polynomial, radial basis function and sigmoid. Long Short Term Memory (LSTM): LSTM are a subtype of recurrent neural networks capable of learning the sequence. These networks are capable of taking inputs from previous cycles. These features make the algorithm more suitable for text classification.

5 CONCLUSIONS:

Aggression detection or hate speech detection in social media websites is becoming a problem of increasing complexity. In this paper we surveyed different works done in this area. All the works exploited set of Machine learning and deep learning algorithms. Although different algorithms have their own advantages and disadvantages, we found that specifically when we need to work with deep learning algorithms line LSTM needs huge number of data sets to train the models. So in case one is not in possession of large amount of data sets for training, it is better to avoid the use of deep learning algorithms. Having said this, all the machine learning algorithms also needs a considerable amount of training data for obtaining a considerable accuracy.

REFERENCES

[1] Fauzi, M. Ali, and Yuniarti Anny. "Ensemble Method for Indonesian Twitter Hate Speech Detection." Indonesian Journal of Electrical Engineering and Computer Science, 2018: 294-299.

[2] Gitari, Njagi Dennis, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. "A Lexicon-based Approach for Hate Speech Detection." International Journal of Multimedia and Ubiquitous Engineering, 2015: 215-230.

[3] BEN-DAVID, ANAT, and ARIADNA MATAMOROS-FERNÁNDEZ. "Hate Speech and Covert Discrimination on Social Media:Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain." International Journal of Communication, 2016: 1167–1193.

[4] Davidson, T, D Warmsley, M. W Macy, and I Weber. "Automated Hate Speech Detection and the Problem of Offensive Language." ICWSM . 2017.

[5] ElSherief, M, V Kulkarni, Nguyen D, W. Y Wang, and E. M. Belding-Royer. "A Target-Based Linguistic Analysis of Hate Speech in Social Media." ICWSM. 2018.

[6] Gambäck, Björn , and Utpal Kumar Sikdar. "Using Convolutional Neural Networks to Classify Hate-Speech." Proceedings of the First Workshop on Abusive Language Online. Vancouber, Canada: Association for Computational Linguistics, 2017. 85-90.

[7] Gao, Lei , and Ruihong Huang. "Detecting Online Hate Speech Using Context Aware Models." RANLP 2017. 2017.

[8] Kumar, Ritesh, Aishwarya N. Reganti, and Akshit Bhatia. "Aggression-annotated Corpus of Hindi-English Code-mixed Data." Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA), 2018.

[9] Kwok, Irene , and Yuzhou Wang. "Locate the Hate:Detecting Tweets against Blacks." Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence, 2013. 1621-1622.

[10] "Leading countries based on number of Facebook users as of October 2019(in millions)." https://www.statista.com/. 1 1, 2020. https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/ (accessed February 5, 2020).

[11] MacAvaney, Sean , Hao-Ren Yao, Eugene Yang, Katina Russell, and Nazli Goharian. "Hate speech detection: Challenges and solutions." PLOS ONE, 2019.

[12] Magerman, David M. "Statistical decision-tree models for parsing." ACL '95: Proceedings of the 33rd annual meeting on Association for Computational Linguistics. ACM, 1995. 276–283.

[13] Mandal, Mainack , Leandro Araújo Silva, and Fabrício Benevenuto. "A Measurement Study of Hate Speech in Social Media." Proceedings of the 28th ACM Conference on Hypertext and Social Media. ACM, 2017. 85-94.

[14] Martins, Ricardo , Marco Gomes, Jo˜ao Jos´e Almeida, Paulo Novais, and Pedro Henriques. "Hate speech classification in social media using emotional analysis." 2018 7th Brazilian Conference on Intelligent Systems (BRACIS). Sao Paulo, Brazil: IEEE, 2918. 61-66.

[15] Nicas, Jack. "Does Facebook Really Know How Many Fake Accounts It Has?" https://www.nytimes.com/. January 31, 2019. https://www.nytimes.com/2019/01/30/technology/facebook-fake-accounts.html (accessed February 05, 2020).

[16] Silva, Leandro , Mainack Mondal, Denzil Correa, Fabr´ıcio Benevenuto, and Ingmar Weber. "Analyzing the Targets of Hate in Online Social Media." Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016). Cornel University, 2016. 687-690.

[17] Vigna, Fabio Del , Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. "Hate me, hate me not: Hate speech detection on Facebook." Proceedings of the First Italian Conference on Cybersecurity (ITASEC17),. Venice Itally, 2017. 86-95.

[18] Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y., & Nerur, S. (2017). Advances in Social Media Research: Past, Present and Future. Inf Syst Front.

[19] Christofides,, E., Muise, A., & Desmarais, S. (2009). Information Disclosure and Control on Facebook: Are They Two Sides of the Same Coin or Two Different Processes? CYBERPSYCHOLOGY & BEHAVIOR, 341-345.

[20] Greenwood, B. N., & Gopal, A. (2015). Tigerblood: Newspapers, Blogs, and the Founding of Information Technology Firms. Information Systems Research, 1-17.

[21] Kruikemeier, S., Sezgin, M., & Sophie , C. (2016). Political Microtargeting: Relationship Between Personalized Advertising on Facebook and Voters' Responses. CYBERPSYCHOLOGY, BEHAVIOR, AND SOCIAL NETWORKING, 367-372.

[22] Kumar, R., Reganti, A. N., & Bhatia,, A. (2018). Aggression-annotated Corpus of Hindi-English Code-mixed Data. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA).

[23] Tang, J., Zhang, P., & Philip , W. F. (2014). Categorizing consumer behavioral responses and artifact design features: The case of online advertising. Inf Syst Front.