International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 08 | Aug 2023 www.irjet.net p-ISSN: 2395-0072

Automatic Prediction and Countering of Communal Tweets using Machine Learning Techniques

R. Durga[1], V. Pavithra[2]

[1]Student, M.E, Dept. of Computer Science and Engineering, T.J.S Engineering College, Tamil Nadu, India
[2]Assistant Professor, Dept. of Computer Science and Engineering, T.J.S Engineering College, Tamil Nadu, India

---

Abstract - *Social networks such as Twitter and Facebook are seriously affected by abusive and offensive content. A lot of research has been carried out in recent years for automatic identification of different types of hate speech. Hate speech can come under several categories such as gender, ethnicity, nationality, racial groups and religions of the targeted groups. This work explores methods which aim to counter potentially harmful communal tweets which are pointed towards religious groups such as Hindus, Sikhs, Muslims, Jains, Christians, Buddhists. Communal tweets are largely posted during calamities and taking advantage of such situations, hatred and misinformation are propagated in the affected region which may result in serious riot and violence. Considering the potentially adverse effects of communal tweets, a classifier to identify communal tweets and non communal tweets is developed using the Support Vector Machine algorithm which performs better than existing approaches. Users who post communal tweets are identified. Anti communal tweets are identified from the non communal tweets using a classifier and by finding the appropriate anti communal tweet to a communal tweet, this system counters communal tweets by making use of such anti communal tweets posted by some users. The proposed system will be really helpful to reduce the increasing communal venom in society by blocking the users who repeat posting such communal tweets. It also changes the negative opinion of communal tweeters by promoting anti communal contents.*

Key Words: Anti communal tweets, Classifier, Communal tweets, Religious groups, Users

1. INTRODUCTION

In recent years, there is an increase in propagation of communal speech on social media and the urgent need for effective countermeasures are needed. Twitter is a social broadcast network that enables people to publicly share brief messages instantly around the world. This message brings a variety of people with different voices, ideas and perspectives. Twitter prevents and filters only abusive contents and offensive words from the tweets but not the communal contents. Using this, communal tweets are pointed towards certain religions, racial communities, politicians or certain groups. Communal tweets are posted during times of calamities or emergency situations. Hence, analyzing the communal tweets on twitter during calamities is focused. While looking through these tweets, it has been observed that a large amount of communal tweets are posted rather than anti communal tweets. A calamity generally affects the morale of the masses making them vulnerable. Often, taking advantage of such situations, hatred and misinformation are propagated in the affected region, which may result in serious deterioration of law and order.

A detailed analysis of communal tweets during disaster situations are detected and countered. This system characterizes users who initiate and promote communal tweets and also suggest a way to counter such communal content with anti communal tweets that asks users not to spread communal hatred. Considering the adverse effects of communal tweets, a Support Vector Machine Classifier is developed to automatically separate the communal tweets from the non communal ones. After identifying communal tweets, the nature of communal tweets and the users who post them are identified. Initiators, who initiate the communal tweets are identified and propagators who retweet the communal tweets posted by initiators or copy the contents of other initiators and post their own tweet with minor changes are also identified. Apart from the communal tweets, non communal tweets are analyzed to automatically identify anti communal tweets for countering. This will dissuade people from posting communal contents. Hence, a convincing way to counter the communal tweetis is to counter them with anti communal tweets to the initiators and propagators who post communal tweets. A real time system is also developed that automatically collects tweets from the user and counters with anti communal tweets if the posted tweet is communal and blocks the users who repeat posting the communal tweets through which communal venom in the society can be reduced.

2. EXISTING SYSTEM

It has been observed that offensive tweets are often posted during calamities to which the attackers are affiliated. However, it is quite surprising that in certain geographical regions such as the Indian subcontinent, communal tweets are posted even during natural disasters such as floods and earthquakes and also during man-made disasters. Several studies have attempted to identify online content that is potentially hate speech or offensive

in nature by a supervised bag-of-words (BOW) model to classify racist content in web pages. Along with words, context features are also incorporated to improve the classification accuracy. Offensive content in Youtube comments using obscenities, profanities and pejorative terms as features with appropriate weightage. Similarly, cyberbullying was identified by Dinakar et al. using features like parts-of-speech tags, profane words, words with negative connotations and so on. More recently, Burnap and Williams proposed a hate term and dependence feature-based model to identify hate speeches posted during a disaster event called the Woolwich attack. Alsaedi et al. proposed a classification and clustering based technique to predict disruptive events like riot. Burnap and Williams proposed a model to detect cyber hate on Twitter across multiple protected characteristics such as race, disability, sex, etc. The most prevailing hate speeches are targeted toward certain races, while religion-induced hate speech is very sparse. Hence, a general purpose hate speech identifier may fail to capture all the nuances of a rare category like religion-based hate speech, especially, when tweets from the rare categories are posted in huge volumes for a short period of time. We actually demonstrate in this paper that the classifier proposed in can hardly capture communal tweets. Consequently, in recent times, researchers focus on more granular levels of hate speech detection in Twitter. For example, Chaudhry tried to track racism in Twitter and Burnap and Williams detected religious hate speeches posted during the Woolwich attack. On the contrary, this paper focuses on Twitter and it has been widely demonstrated that the standard Natural Language Processing based methodologies which have been developed for formally written text, do not work well for short, informal tweets. Hence, new methodologies are necessary to deal with communal and noisy content posted on Twitter.

3. PROPOSED WORK

Detailed analysis of communal tweets posted during calamity situations such as automatic identification of such tweets, analyzing the users who post such tweets and also suggest a way to counter such content. In this paper, we try to identify communal tweets, characterize users initiating or promoting such contents and counter such communal tweets with anti communal posts that ask users not to spread communal venom. Support Vector Machine classifier is proposed to extract communal tweets and this classifier can be directly used over any future event without further training. Second, users are classified into two categories, one is originators who post a tweet and the other one is propagators who retweets the content of originators. We not only rely on retweets but also explore similarities between tweets, their timestamps in order to identify initiators and propagators more accurately. Rest of the analysis is performed on these modified groups of users. Apart from that, temporal patterns of the identified set of communal users are analyzed to understand their outraging phenomenon. Third, another Support Vector classifier is proposed to detect anti communal tweets and such tweets can be used to neutralize the effect of communal tweets. A real time system is developed to automatically identify communal and anti communal tweets posted during and counter with anti communal tweet to the communal tweet and also block the user if communal tweets are posted by them again.

4. METHODOLOGY

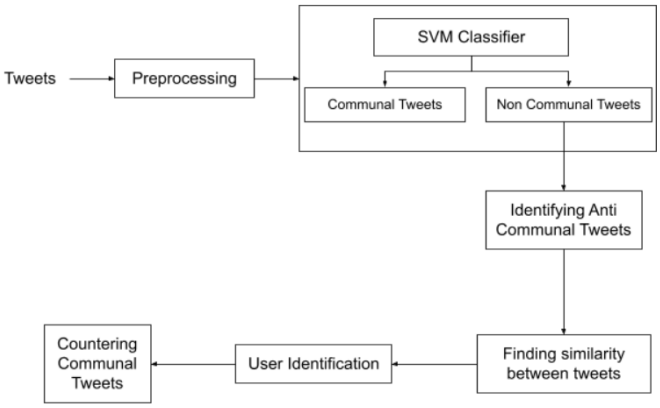

Fig -1: Proposed Methodology

4.1 Preprocessing

The dataset is collected in two ways, first the implicit dataset is given from the recent disaster event in the admin page and second, to develop the real time service which automatically collects tweets from the user in the user page. The tweets are preprocessed using stopwords removal and stemming techniques. Dictionary based approach is used to remove stopwords. Articles, prepositions, pronouns, conjunctions, etc are stop words and do not add much information to the text. Examples of a few stop words are “the”, “a”, “an”, “so”, “what”. Removal of stop words reduces the dataset size and reduces the training time due to the fewer number of tokens involved in the training. Stemming is the process of reducing inflected words to their word stem, base or root form, generally written in a word form.

4.2 Communal Tweet Classification

Support Vector Machine Classifier is used to classify the tweets into communal and non communal tweets. This is the algorithm which gives high accuracy for text classification. Support Vector Machine algorithm creates the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM

chooses the extreme points or vectors that help in creating the hyperplane. A model that can accurately identify whether it is a communal or a non communal tweet should be created by using the Support Vector Machine algorithm. We will first train our model with lots of communal and non communal tweets using communal keywords so that it can learn about which is communal and non communal tweet and then we test it with a new tweet. On the basis of the support vectors, it will classify. So as the support vector creates a decision boundary between these two data (communal and non communal) and chooses extreme cases (support vectors), it will see the extreme case of communal and non communal.

4.3 Identifying Anti Communal Tweets

After classifying communal tweets and non communal tweets, anti communal tweets from the non communal tweets are identified using the Support Vector Machine Classifier which has been trained using anti communal contents which try to dissuade people from posting communal tweets.

4.4 Finding suitable Anti Communal Tweet

Anti communal tweet which is suitable to counter communal tweet is identified through a training model. This anti communal tweet is used to counter the initiator or propagator with such anti communal tweets for those who post the communal tweet.

Table -1: Sample Communal and Anti Communal Tweets

Communal Tweets	Anti Communal Tweets
Soul Vultures, Evangelical Vultures	Tears and blood have no religion. All they know is only pain. It is not just in your country. It is everywhere.
Huh, it's Muslims behind attack, Islam attacks Paris	Muslims are not terrorists. This is not Islam. Nothing to do with Islam.

4.5 User Identification

Users who posted the communal tweets are found. We have two categories of users, one is initiators who post a tweet and the other one is propagators who retweet the posted tweet of some user or post a tweet by modifying an already posted tweet. All the users who posted communal tweets are stored in a database so if a user name repeats for posting a communal tweet then the user will be blocked.

4.6 Countering Communal Tweet

Anti communal tweet which is suitable to counter the communal tweet is identified and countered to the initiator or propagator who posted the communal tweet.

The various outputs are shown below,

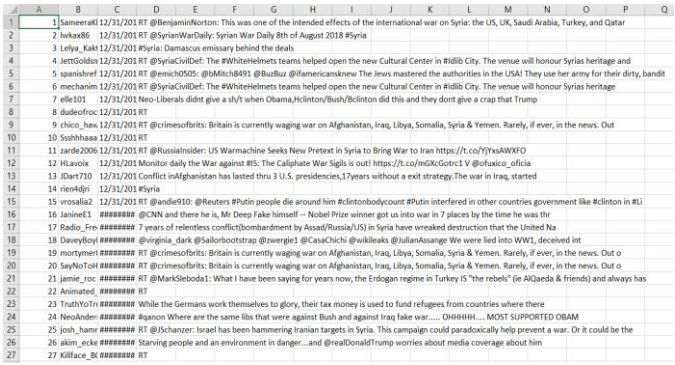

Fig -2: Tweets Dataset

Fig -3: Admin Page to upload Implicit Dataset

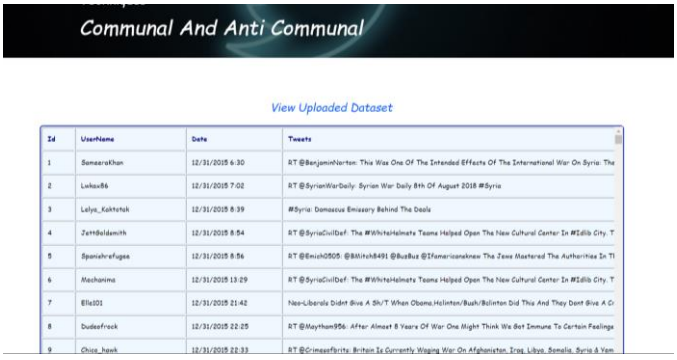

Fig -4: Uploaded Dataset

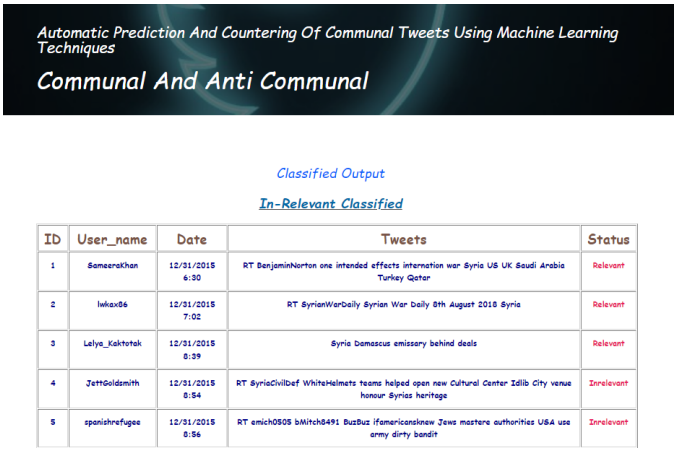

Fig -5: Classified Communal and Non Communal Tweet

Fig -6: Classified Anti Communal Tweet

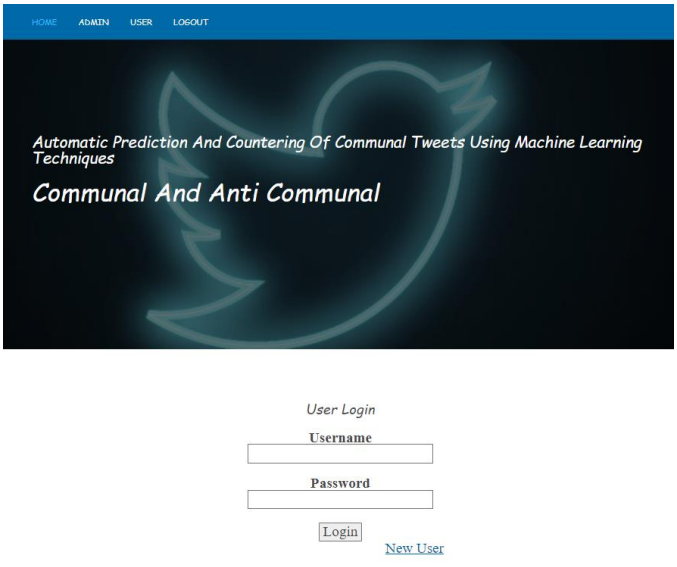

Fig -7: User Login

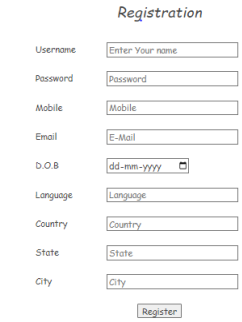

Fig -8: New User Registration

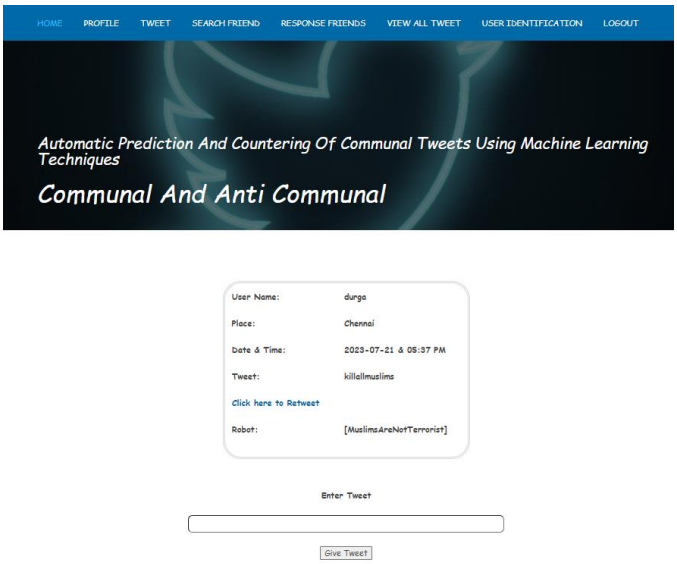

Fig -9: Real Time User's Tweet Page

5. CONCLUSION

This proposed system counters communal tweets. Support Vector Machine classifier is used to filter out communal tweets. Anti communal tweets are found from the non communal contents and that is used to counter the communal tweets. A real time service is developed with user and admin login where admin can upload the tweet dataset and see all the classified communal, non communal and anti communal tweets and the users who posted the communal tweets. Tweets are collected from dataset and also through real time service in which users post tweets in real time which can be used directly in the future to identify and counter the communal tweets with anti communal tweets and also the user will be blocked if the user posts communal tweet again. It gives 91% precision, 95% recall, 93% F1 Score and 92% of overall accuracy. The accuracy of the classifier in the proposed system is comparatively higher than the existing system.

6. FUTURE ENHANCEMENTS

- The proposed communal tweet classifier can be used as an early warning signal to identify communal tweets, and then celebrities, political personalities can be made aware of the situation and requested to post anti communal tweets so that such tweets get higher exposure. Anti Communal content needs to be promoted via mentioning popular celebrities, political persons. Our real time system can be used by the Government in taking decisions regarding filtering communal content, promoting anti communal content etc.

- The number of distinct anti communal tweets is much less in number due to the availability of a small number of anti communal tweets. In the future, the proposed set of lexicons will be enlarged.

- Tweets are known to be informally written and noisy in nature, containing misspellings, abbreviations, smiley etc. In the future, these variations will be handled to improve our classifiers.

- It would be effective if the posted communal tweets are blocked. It would be more effective if the user is intimated or given warning before posting a communal tweet or the system should not allow users posting communal tweets.

REFERENCES

[1] P. Burnap and M. L. Williams, "Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making," Policy Internet, vol. 7, no. 2, pp. 223–242, 2015.

[2] L. A. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, "Analyzing the targets of hate in online social media," in Proc. ICWSM, Mar. 2016, pp. 687–690.

[3] M. Mondal, L. A. Silva, and F. Benevenuto, "A measurement study of hate speech in social media," in Proc. ACM HT, 2017, pp. 85–94.

[4] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in Proc. WWW, 2015, pp. 29–30.

[5] W. Magdy, K. Darwish, N. Abokhodair, A. Rahimi, and T. Baldwin, "#ISISisNotIslam or #DeportAllMuslims?: Predicting unspoken views," in Proc. ACM Web Sci., 2016, pp. 95–106.

[6] K. Rudra, A. Sharma, N. Ganguly, and S. Ghosh, "Characterizing communal microblogs during disaster events," in Proc. IEEE/ACM ASONAM, Aug. 2016, pp. 96–99.

[7] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert, "The bag of communities: Identifying abusive behavior online with preexisting Internet data," in Proc. ACM CHI, 2017, pp. 3175–3187.

[8] Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata, "Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study", 10.1109/ICACSIS.2017.8355039, 2017.

[9] K. Rudra, S. Banerjee, N. Ganguly, P. Goyal, M. Imran, and P. Mitra, "Summarizing situational tweets in crisis scenario," in Proc. 27th ACM Conf. Hypertext Social Media (HT), 2016, pp. 137–147.