JES Journal of Engineering Sciences

Vol 11, Issue 3, March/2020
ISSN NO:0377-9254

AN EVENT-INDEPENDENT CLASSIFIER FOR FILTER OUT COMMUNAL TWEETS EARLY

[1]B Sandhya, [2]Dr.G. Venkata Rami Reddy , [3]Mr Katroth Balakrishna Maruthiram

[1]Master of Technology, Department of Software Engineering, School of Information Technology (JNTUH), Kukatpally, Hyderabad-500085.

[2]Professor and Additional Controller of Examinations, School of Information Technology (JNTUH), Kukatpally, Hyderabad-500085.

[3]Assistant Professor of (C), School of Information Technology (JNTUH), Kukatpally, Hyderabad-500085.

Abstract— ***The huge amount of tweets posted during a disaster event includes information about the present situation as well as the emotions/opinions of the masses. While looking through these tweets, we realized that a large amount of communal tweets, i.e., abusive posts targeting specific religious/racial groups are posted even during natural disasters—this paper focuses on such category of tweets, which is in sharp contrast to most of the prior research concentrating on extracting situational information. Considering the potentially adverse effects of communal tweets during disasters, in this paper, we develop a classifier to distinguish communal tweets from noncommunal ones, which performs significantly better than existing approaches. We also characterize the communal tweets posted during five recent disaster events, and the users who posted such tweets. Interestingly, we find that a large proportion of communal tweets are posted by popular users (having tens of thousands of followers), most of whom are related to media and politics. Further, users posting communal tweets form strong connected groups in the social network. As a result, the reach of communal tweets is much higher than noncommunal tweets. We also propose an event-independent classifier to automatically identify anticommunal tweets and also indicate a way to counter communal tweets, by utilizing such anticommunal tweets posted by some users during disaster events. Finally, we develop a real-time service to automatically collect tweets related to a disaster event and identify communal and anticommunal tweets from that set. We believe that such a system is really helpful for government and local monitoring agencies to take appropriate decisions like filtering or promoting some particular contents.***

1. INTRODUCTION

ONLINE social media (OSM) such as Twitter and Face- book are today seriously plagued by offensive and abusive content, such as trolling, cyberbullying, hate speech, and so on. A lot of research has been carried out in recent years for automatic identification of different types of offensive content [1]–[5]. Hate speech can come under several categories where people target various attributes such as religion, gender, sex, ethnicity, nationality, etc., of the target group [6]. Out of different types of hate speech, we in this paper focus on an especially harmful and potentially dangerous category—communal tweets, which are directed toward certain religious or racial communities such as "Hindu," "Muslims," "Christians," etc. Especially, we study communal tweets that are posted during times of disasters or emergency situations. A disaster situation generally affects the morale of the masses making them vulnerable. Often, taking advantage of such situation, hatred and misinformation are propagated in the affected region, which may result in serious deterioration of law and order situation. In this paper,

we provide a detailed analysis of communal tweets posted during disaster situations—such as automatic identification of such tweets, analyzing the users who post such tweets—and also suggest a way to counter such content.

2. RELATED WORK

Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making [1]

The use of "Big Data" in policy and decision making is a current topic of debate. The 2013 murder of Drummer Lee Rigby in Woolwich, London, UK led to an extensive public reaction on social media, providing the opportunity to study the spread of online hate speech (cyber hate) on Twitter. Human annotated Twitter data was collected in the immediate aftermath of Rigby's murder to train and test a supervised machine learning text classifier that distinguishes between hateful and/or antagonistic responses with a focus on race, ethnicity, or religion; and more general responses. Classification features were derived from the content of each tweet, including grammatical dependencies between words to recognize "othering" phrases, incitement to respond with antagonistic action, and claims of well-founded or justified discrimination against social groups. The results of the classifier were optimal using a combination of probabilistic, rule-based, and spatial-based classifiers with a voted ensemble meta-classifier. We demonstrate how the results of the classifier can be robustly utilized in a statistical model used to forecast the likely spread of cyber hate in a sample of Twitter data. The applications to policy and decision making are discussed.

In this article we have developed a supervised machine learning classifier for hateful and antagonistic content in Twitter. The purpose of the classifier is to assist policy and decision makers in monitoring the public reaction to large-scale emotive events, such as the murder of Drummer Lee Rigby in Woolwich in 2013. Previous research showed that 58 percent of hate crimes following 9/11 were perpetrated two weeks following the event (4 percent of the at-risk period). Data are available in near-real time from online social networks and microblogging websites such as Twitter, which can allow us to monitor the prevalence of hateful and antagonistic responses online in the period immediately following the event, when risk of hateful responses is highest. Hateful and antagonistic responses have led to imprisonment of the person posting the tweet possibly as part of a risk reduction response by the judicial system. The classification results showed very high levels of performance at reducing false positives and produced promising results with respect to false negatives. Our implementation of individual probabilistic, rule-based, and spatial classifiers performed similarly across most feature sets, but the combination of the classification output of these base classifiers using a voted meta-classifier based on maximum probability matched or improved on the recall of the base classifiers in every experiment, suggesting that an ensemble classification approach is most suitable for classifying cyber hate, given the current feature sets. This could be due to the noise and variety of types of response within the data, with some features proving more effective with different classifiers. The novel inclusion of syntactic features using typed dependencies within tweets as machine learning features reduced the false negatives by 7 percent over the baseline BoW features, providing a significant improvement when considering the volumes of data produced in response to such events. Our corpus of 450,000 tweets was collected in the first two weeks following the event, and it would be extremely difficult for human effort to manually parse these data to determine levels of public antagonism within all the responses. The improvement in machine classification using typed dependencies also suggests that cyber hate comprises content that is not instantly identifiable by words that are traditionally associated with hateful and discriminatory remarks, and requires a more nuanced approach to text classification beyond words alone. For instance, there was a prevalence of "othering" terms, such as "send them home" and "get them out," as well as incitements to undertake hateful retribution such as "burn korans" and "should be hung." The typed dependency approach was able to identify these as useful features for classification. We developed an illustrative example using cyber hate as classified by a machine as a predictive feature in a statistical

regression model. The model produced IRRs for retweet activity given a set of features for each tweet. The model showed a reduction in retweet rate ratio when a tweet contained a hateful or antagonistic response, suggesting a stemming of the flow of content on Twitter when a tweet contained cyber hate. This combination of machine classification and statistical modeling can—while accepting the limitations of machines with respect to utilizing a learned set of predictive features that are not an absolute reflection of all the possible combinations and permutations of cyber hate characteristics—produce aggregated statistics and prevalence indicators for hateful and antagonistic responses to an event on social media, including the relative spread of cyber hate on Twitter over time.

Analyzing the Targets of Hate in Online Social Media [2]

Social media systems allow Internet users a congenial platform to freely express their thoughts and opinions. Although this property represents incredible and unique communication opportunities, it also brings along important challenges. Online hate speech is an archetypal example of such challenges. Despite its magnitude and scale, there is a significant gap in understanding the nature of hate speech on social media. In this paper, we provide the first of a kind systematic large scale measurement study of the main targets of hate speech in online social media. To do that, we gather traces from two social media systems: Whisper and Twitter. We then develop and validate a methodology to identify hate speech on both these systems. Our results identify online hate speech forms and offer a broader understanding of the phenomenon, providing directions for prevention and detection approaches.

The fight against perceived online hate speech is beginning to reach a number of concerned parties, from governments to private companies, as well as to a growing number of active organizations and affected individuals. Our measurement study about online hate speech provides an overview of how this very important problem of the modern society currently manifests. Our effort even unveils new forms of online hate that are not necessarily crimes, but can be harmful to people. We hope that our dataset and methodology can help monitoring systems and detection algorithms to identify novel keywords related to hate speech as well as inspire more elaborated mechanisms to identify online hate speech. Building a hate speech detection system that leverages our findings is also part of our future research agenda.

3. FRAMEWORK

In this paper, we try to identify communal tweets, characterize users initiating or promoting such contents, and counter such communal tweets with anticommunal posts that ask users not to spread communal venom. Although there exist prior works on communal tweet identification, to our knowledge this paper is the first on characterizing communal tweets and users who post such tweets during disasters, and it tries to find out how social media platforms are used to spread communal content even during natural disasters in some regions.

Fig1: Word cloud of two events. (a) NEQuake. (b) PAttack.

Our communal tweet characterization approach was first proposed in a prior study [9]. This paper extends our prior work as follows. First, we have proposed a rule-based classifier using low-level lexical features to extract communal tweets and this classifier can be directly used over any future event without further training. Second, earlier we had classified users into two categories: 1) originators who post a tweet and 2) propagators retweeting the content of originators. In this paper, we not only rely on retweets but also explore similarity between tweets, their timestamps in order to identify initiators and propagators more accurately.

Journal of Engineering Sciences
Vol 11, Issue 3, March/2020
ISSN NO:0377-9254

4. EXPERIMENTAL RESULTS

In this paper author is describing concept to detect communal hate tweets spread in social media networks during disaster events occurred. Sometime during natural disaster event, peoples may use social media networks to spread current situation or relief activities happening at disaster area and some corrupt peoples may use this situation to spread hate messages to disturb peace. To detect such hate tweets author is using rule base concept to detect such tweets, in this rule we will find out that such hate tweets may get more re-tweet counts and may have some hate words such as Muslims, Christians, terror, attacks etc and we look such words from tweets to define or classify as communal tweets and tweets not contains such words may be consider as non communal. Author is saying we can download such hate words for religion and race from different web sites such as http://www.translationdirectory.com/glossaries, www.hatebase.org etc. To implement above concept I downloaded tweets dataset from http://www.cnergres.iitkgp.ac.in/disasterCommunal/dataset.html website First we will upload tweets dataset and then build train classifier by assign each tweets to either communal or non communal class by applying rule concepts and by using hate words.

Fig.2: Home screen

Fig.3: Dataset screen

Fig.4: View classifier screen

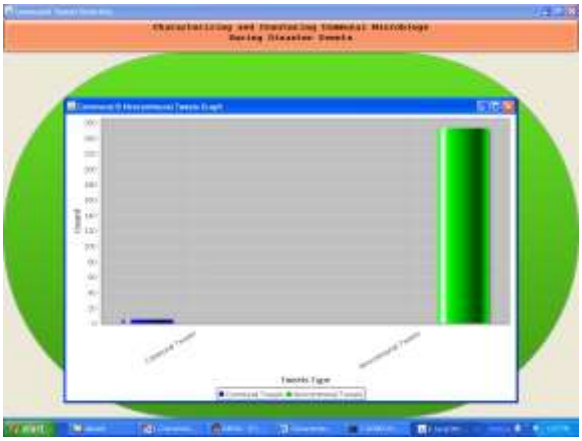

Fig.5: Graph screen

5. CONCLUSION

To our knowledge, this paper is the first attempt in the direction of characterizing communal tweets posted during the disaster scenario and analyzing the users involved in posting such tweets. We proposed an event-independent classifier that can be used to filter out communal tweets early. We also found that communal tweets are retweeted heavily and posted by many popular users; mostly belong to news media and politics domain. Users involved in initiating and promoting communal contents form a strong social bond among themselves. Additionally, most of the users get angry suddenly due to such kind of events and express their hates to specific religious communities involved in the event. We observe that, during a disaster, some users also post anticommunal content asking people to stop spreading communal posts, and it is necessary to counter the potential adverse effects of communal tweets. We have proposed an event-independent classifier to identify such anticommunal tweets. However, we have found such anticommunal tweets are retweeted much less compared to communal tweets and they are also very few in number compared to communal tweets. Finally, we proposed a realtime system DisCom which can be used directly in the future disaster events to identify communal and anticommunal tweets.

6. FUTURE WORK

We believe that our present study has many potential future applications. For instance, 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 anticommunal tweets so that such tweets get higher exposure. We need to promote anticommunal content via mentioning popular celebrities, political persons. Our real-time system DisCom can be used by the Government in taking decisions regarding filtering communal content, promoting anticommunal content etc. We plan to pursue some potential directions of countering communal tweets in the future. This paper also raises many intriguing social questions like "interaction between communal and anticommunal users," "demographic biases," etc. We will try to address these questions in the future.

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] I. Chaudhry, "#Hashtagging hate: Using Twitter to track racism online," First Monday, vol. 20, no. 2, 2015. [Online]. Available: http://firstmonday.org/ojs/index.php/fm/article/view/5450

[3] 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.

[4] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," Int. J. Multimedia Ubiquitous Eng., vol. 10, no. 4, pp. 215–230, 2015.

[5] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in Proc. 27th AAAI Conf. Artif. Intell., 2013, pp. 1621–1622.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] E. Greevy and A. F. Smeaton, "Classifying racist texts using a support vector machine," in Proc. SIGIR, 2004, pp. 468–469.