Identifying and Characterizing the Communal and Anti-Communal Tweets during Disaster Events by Event Independent Classifier

PROF.KALAM NARREN

[1]M.Tech Scholar, Department of Computer Science and Engineering, JNTUH College of Engineering, Jagityal.

[2]Professor of CSI, ISIE and Computer Science and Engineering, JNTUH College of Engineering, Jagityal.

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, cyber bullying, 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 . 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. Earlier it has been observed that such offensive tweets are often posted during man-made disasters like terrorist attacks. For instance, Burnap and Williams have shown that the U.K. masses targeted a certain religious community during Woolwich attack to which the attackers are affiliated. However, it is quite surprising that in certain geographical regions such as Indian subcontinent, communal tweets are posted even during natural disasters such as floods and earthquakes. Some examples of communal tweets are shown in Table I. Such kind of communal tweets help in developing hatred and agnosticism among common masses, which subsequently deteriorates communal harmony, law and order situation. In the midst of disaster, this kind of situation is really difficult for government to handle. 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.

2. RELATED WORK

[2] Locate the Hate: Detecting Tweets against Blacks

Although the social medium Twitter grants users freedom of speech, its instantaneous nature and retweeting features also amplify hate speech. Because Twitter has a sizeable black constituency, racist tweets against blacks are especially detrimental in the Twitter community, though this effect may not be obvious against a backdrop of half a billion tweets a day.1 We apply a supervised machine learning approach, employing inexpensively acquired labeled data from diverse Twitter accounts to learn a binary classifier for the labels "racist" and "nonracist." The classifier has a 76% average accuracy on individual tweets, suggesting that with further improvements, our work can contribute data on the sources of anti-black hate speech.

We have shown that our bag-of-words model is insufficient to accurately classify anti-black tweets. Although the discrepancy found in our survey reflects the difficulty in achieving this accuracy, this challenge should serve as motivation for searching for ways to further refine our classification. Our algorithms need to include bigrams, as well as sentiment analysis and classification, word sense disambiguation, etc. Future explorations may include how often we need to incorporate new vocabulary, how we may utilize popular hash tags to collect more training data, how we may predict and classify deliberate misspellings, how we may involve the racial identity of Twitter users, whether anti-black tweets are targeted to individuals or to groups, how often anti-black tweets are woven into various conversations, etc. As more and more people participate in social media networks, platforms like Twitter become an intersection for diverse groups and individuals, which in turn makes our research increasingly relevant.

[3] A Measurement Study of Hate Speech in Social Media

Social media platforms provide an inexpensive communication medium that allows anyone to quickly reach millions of

users. Consequently, in these platforms anyone can publish content and anyone interested in the content can obtain it, representing a transformative revolution in our society. However, this same potential of social media systems brings together an important challenge these systems provide space for discourses that are harmful to certain groups of people. This challenge manifests itself with a number of variations, including bullying, offensive content, and hate speech. Specifically, authorities of many countries today are rapidly recognizing hate speech as a serious problem, specially because it is hard to create barriers on the Internet to prevent the dissemination of hate across countries or minorities. In this paper, we provide the first of a kind systematic large scale measurement and analysis study of hate speech in online social media. We aim to understand the abundance of hate speech in online social media, the most common hate expressions, the effect of anonymity on hate speech and the most hated groups across regions. In order to achieve our objectives, we gather traces from two social media systems: Whisper and Twitter. We then develop and validate a methodology to identify hate speech on both of these systems. Our results identify hate speech forms and unveil a set of important patterns, providing not only a broader understanding of online hate speech, but also offering directions for detection and prevention approaches.

3. FRAMEWORK

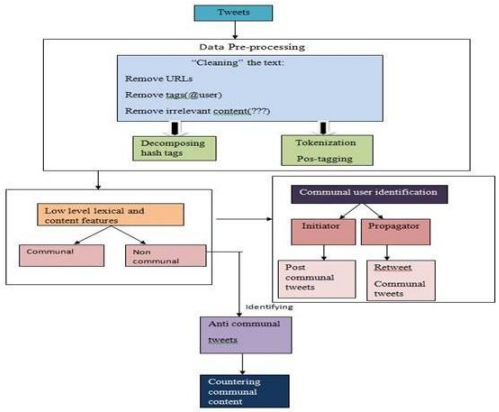

Fig.1. System Design

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.

4. EXPERIMENTAL RESULTS

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.

Fig.3. In above screen we can see test tweet and classified train tweet with predicted value as communal or non communal.

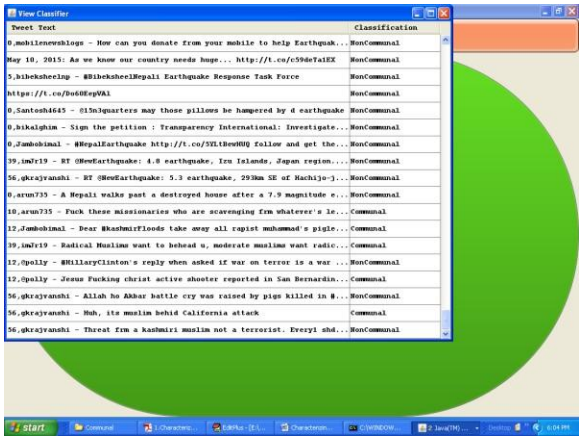

Fig.2. we can see each tweet assign with class. Now upload test dataset and predict the class name from train dataset.

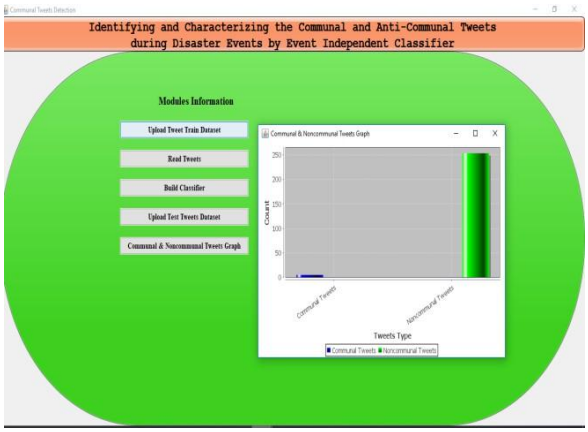

Fig.4. In above graph x-axis represents tweet type as communal or non communal and y-axis represents count.

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.

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.

[11] N. Pendar, "Toward spotting the pedophile telling victim from predator in text chats," in Proc. ICSC, Sep. 2007, pp. 235–241.

[12] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in Proc. Int. Conf. Social Comput. Privacy, Secur., Risk Trust (PASSAT), (SocialCom), Sep. 2012, pp. 71–80.

[13] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," ACM Trans. Interact. Intell. Syst., vol. 2, no. 3, p. 18, 2012.

[14] P. Burnap et al., "Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack," Social Netw. Anal. Mining, vol. 4, no. 1, p. 206, 2014.