Characterizing and Countering Communal and Anti-Communal Tweets During Disasters

Archana K. S, Latha M, Sheela Gowr P

Abstract: Various tweets shared during a disaster situation encompasses data related to current scenario and about emotions/opinions. By analyzing these communal tweets, abusive posts which targets various religious and racial groups during natural calamities has been found. By reviewing its effects, a classifier has been developed to distinguish between communal and non-communal messages, which shows better performance. People posting such communal tweets has been analyzed which says that most of them are posted by popular users from media, political and form strong correlated groups in the social network which makes it to reach higher. An event-independent classifier has been proposed which identifies anti-communal tweets automatically and propose a way to counter back. A real-time service has been developed to find tweets automatically related to an emergency segregating communal and anti-communal tweets. Government and local monitoring agencies can use this system for making decisions like filtering or to promote some news.

Keywords: Tweets, Communal, Non-Communal, Event-Independent classifier.

I. INTRODUCTION:
Communal tweets were analyzed during disaster situations which identifies the messages along with the users and it helps to counter that information. Offensive messages are getting shared at the time of terrorist attacks. For example, Burnap et al, have focused about the religious community which got noticed during Woolwich [1]. In few geographical regions including Indian subcontinent, they are being shared when floods and earthquakes occur. Such tweets develop hatred among common people, resulting in loss of communal harmony, causing problem in law and order situation. Even the government feels difficult to handle such situation during disasters [1,5].

II. LITERATURE SURVEY:
In 2013 Drummer Lee Rigby in Woolwich got murdered which gave wide retroaction on social media. In Rigby's murder twitter data was collected which used a text classifier based on machine learning that helped to find messages related to race, ethnicity, or religion. Form the content available, the features for classification were extracted including grammatical dependencies between words. Statistical model is used to show the Optimal results obtained by probabilistic classifiers to predict the spread of cyber hate in Twitter [2,8].

III. EXISTING SYSTEM:
Various studies identified online contents like speeches that are hate or offensive in nature. A supervised BOW model helped to classify those contents which includes words, features with respect to context are also incorporated for high classification accuracy. Obscenities, profanities were used to find offensive content in You tube forums. Dinakar et al., identified cyberbullying via tags, discussions with negative meanings. In recent times, Williams et al., have discussed about model based on feature which helped to find danger speeches that are shared at the time of disaster happening [4,7].

IV. PROPOSED WORK:
Communal tweets analysis during disaster situations was done. Communal tweets were identified along with users initializing or encouraging such messages, and counter them with anti-communal posts instructing people to avoid communal venom. Low-level lexical features were used by a classifier for extracting communal messages [4]. Two categories were specified: 1) originators 2) propagators. Here in addition to retweets the similarity between tweets, timestamps identifying initiators and propagators were identified more accurately performing analysis on these groups of users. Identified users with their patterns was extracted to learn the events happening with their phenomena. Rule-based classifier has been proposed which helps to neutralize the effect. A real time system Dis Com has been developed which automatically identifies the communal and anti-communal tweets during emergency situations.

V. IMPLEMENTATION:
The proposed model has the following modules and module description has been explained below.
1. Authentication
2. Data Collection
3. Pre-processing
4. Communal Tweets Identification
5. Anti-Communal Tweets Identification
6. Communal User Identification
7. Performance Analysis

**Authentication:**
This module belongs to both admin and User. Admin can login with his available details and the User needs to register by providing the details and can login with the details. If both matches, then the valid user can proceed further.

**Data Collection:**
Twitter dataset has been collected based on the Disaster Events like Syria war, PA attack(Coordinated terrorist attack in Paris). This dataset contains tweets of the different people belonging to communal and non-communal tweets. By uploading the dataset, the communal and non-communal tweets can be classified from the dataset.

**Preprocessing:**
In this step, admin can view the uploaded dataset and it should be preprocessed. Here stop-words and stemming techniques have been used. Tweets may contain url, tags(@user), irrelevant content(???) and these unwanted contents should be removed. which includes removal of URLs, Tags and irrelevant expressions. Now the stop-words technique is used to remove the unwanted data like is, this, and, a, at, the etc. from the tweet.

1) **Stopwords Removal:**
A stopword list which holds 75 stopwords has been used. The target text stores individual words as an array while reading a single word from the list. It is matched with output text as an array via sequential search technique. If a match is found, the word gets removed from the array while continuing the comparison till the end. After completion of one stopword, next one gets read from the list and algorithm runs till the end. Output displays the text without stopwords, number of stopwords removed, total words in target text.

2) **Stemming Technique:**
After removing the unwanted words from the tweet, stemming technique is processed. Reducing derived words to its stem is done in this process. Its root form will be in a written word form. There is no need that the stem and the morphological root of the word to be equal; it is enough that correlated words map to the same stem, even if its not in a valid root.

**Communal Tweets Identification:**
After preprocessing, classification of communal and non-communal tweets has been done. To classify the tweets, Rule based algorithm is implemented along with the SVM Algorithm.[9]

**Anti-Communal Identification:**
After the classification of tweets, analysis of non-communal tweets reveals that few users post anti-communal messageshoping to stop people from sharing communal contents. Hence communal tweets can be countered back by promoting such anti communal content[10].

**Communal User Identification:**
Once communal tweets are found, finding its nature and user is essential. Two categories of users have been identified.
1. Initiators
2. Propagators

**Initiators:** People Initiating communal tweets.

**Propagators:** People who retweet communal tweets / copy them and post like their own tweet with few modifications.

**Performance Analysis:**
Finally, Performance has been analyzed and found out following Parameters
i) Precision Value
ii) Recall Value
iii) F1 Measure
iv) Accuracy.

The various outputs have been shown below.
Table 5.1 shows the performance measures for identifying tweets with corresponding values. Event dependent and event independent classifier outputs has been shown with values. The proposed classifier works for event independent showing better accuracy rate and the comparison chart has been shown below.

### Table 4.1 Performance analysis of LDA

| Classifier          | Precision | Recall | F1 Measure | Accuracy |
|---------------------|-----------|--------|------------|----------|
| Baseline models     | 0.846     | 0.385  | 0.463      | 0.55     |
| (Event dependent)   |           |        |            |          |
| LDA                 | 1.0       | 0.416  | 0.587      | 1.228    |
| (Event independent) |           |        |            |          |

### VI. CONCLUSION

Characterizing communal tweets during disaster and analysing its users involved has been a big issue. The event-independent classifier that filters out communal tweets and retweeted messages heavily which are posted by popular users has been developed. Various users initiate and promote communal contents and create a group. Also, there are users who get angry because of such happenings and start to show their hate towards specific religious communities involved.

During disaster, few users also post anti-communal content saying to stop sharing communal posts, providing the necessity to reduce the effects of such tweets. The proposed classifier identifies such anti-communal tweets and the proposed system helps to differentiate tweets during disaster events.

### REFERENCES:

1. L. A. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, “Analyzing the Targets of Hate in Online Social Media”, Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016).
2. 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.
3. I. Chaudhry, "#Hashtagging hate: Using Twitter to track racism online". First Monday, vol. 20, no. 2, 2015.
4. Njagi, Dennis & Zuping, Z. & Hanyurwimfura, Damien & Long, Jun. (2015). “A Lexicon-based Approach for Hate Speech Detection”, International Journal of Multimedia and Ubiquitous Engineering, 10. 215-230. 10.14257/ijmue.2015.10.4.21. [5] I. Kwok and Y. Wang, “Locate the Hate: Detecting Tweets against Blacks,” Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013, pp. 1621–1622.
5. 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.
6. 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.
7. W. Magdy, K. Darwish, N. Abokhodair, A. Rahimi, and T. Baldwin, “#ISISisNotIslam or #DeportAllMuslims?: predicting unspoken views”, Proceedings of the 8th ACM Conference on Web Science, Pages 95-106.
8. E. Greevy and A. F. Smeaton, “Classifying racist texts using a support vector machine”, Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pages 468-469.
9. K. Rudra, A. Sharma, N. Ganguly, and S. Ghosh, “Characterizing communal microblogs during disaster events”, in Proceedings of the 2016 IEEE/ACM international Conference on Advances in Social Networks Analysis and MiningPages 96-99.

AUTHORS PROFILE:

M.Latha, Assistant Professor, Department of Computer Science and Engineering, Vels Institute of Science, Technology and advanced Studies, Chennai, India. Area of Research : Medical Imaging

P.SheelaGowr, Assistant Professor, Department of Computer Science and Engineering, Vels Institute of Science, Technology and advanced Studies, Chennai, India. Area of Research :Cloud Computing and Networking.