# Effective Countering of Communal Hatred during Disaster Events in Social Media

**N. Antony Sophia[1], J. Angelin Jenifer[2], D. Hinduja[3]**

*[1]Assistant Professor, Department of Computer Science and Engineering, Jeppiaar SRR Engineering College, Tamil Nadu, India*

*[1,2]Student, Department of Computer Science and Engineering, Jeppiaar SRR Engineering College, Tamil Nadu, India*

---***---

**Abstract -** *As we are living in an era of digitization and information technology is progressively growing, we are using websites like Twitter, Facebook, Instagram, etc. The use of social media has been in a hike in our day to day life. Especially the teenagers are highly affected by the use of it. Our daily life, social involvement are affected by social media. Social media has changed the way people communicate and socialize on the web. There is a positive effect on business, politics, socialization as well as some negative effects such as cyberbullying, privacy, fake news and communal hate speech. Communal hatred and offensive words in twitter are mainly addressed in this paper. People are forwarding information without checking the credibility of the message. In this scenario a tweet with hatred message also propagates swiftly leading to unrest in the society. To check these kinds of messages propagating and to find a way not to promote these messages, Machine learning technology is used to address these issues. This paper not only helps in identifying hatred speech against communities and religions but also classifies vulgar and offensive or hatred words. The main objective of this project is to plot a graph which displays the percentage of offensive or hatred words used in the tweets. By doing so, the propagation of communal hate speech can be reduced.*

***Key Words***: Cyberbullying, communal hatred, Machine learning, vulgar, offensive, hatred words, hate speech.

## 1. INTRODUCTION

Social media reach is widespread such that information travels very fast without any geographical border or constraints. The age of Internet has changed the way people express their views and opinions. It is now mainly done through social media. Nowadays, millions of people are using social network sites like Twitter, Facebook, Google Plus, etc. to express one's emotions, opinion and share views about their daily lives. But they may not be aware of the issues and clashes arousing among different class of people in the society.

In 2013, an article in the Hindustan Times cited Professor Badri Narayan from the GB Pant Social Science Institute in Allahabad as saying, "From word of mouth, communal polarization is now moving online. This is a dangerous trend as the internet is very potent." Twitter users are likely to stick around the content that has already gotten a lot of retweets and mentions, compared with content that has fewer. The flow of this misinformation on Twitter is a function of both human and technical factors. Human's role is the major factor: Since we are more likely to react to content that taps into our existing grievances and beliefs, inflammatory tweets will generate quick engagement. It is only after that engagement happens that the technical side kicks in: If a tweet is retweeted favorited or replied to by enough of its first viewers, the newsfeed algorithm will show it to large number of users, at which point it will tap into the biases of those users too – prompting even more engagement, and so on. And because of this reason, these buzz tweets are bubbling virally.

This paper mainly focuses on a machine learning technique to analyze and classify the tweets on the basis of parameters like offensive, hatred or neither. The output is displayed in the form of a graph which is shown to the user who posted such tweet.

## 2. RELATED WORKS

Koustav Rudra, Ashish Sharma, Niloy Ganguly, and Saptarshi Ghosh[1]

Have proposed a rule- based classifier to automatically separate communal tweets from non-communal tweets. The tweets are mainly collected from initiators, who initiate a communal tweet and propagators, who retweet the communal tweets. Those users are identified in this paper. After the first-level classification an analysis is made on the non-communal tweets to separate the anti-communal tweets from it. The anti-communal tweets are used to encounter the communal tweets.

Ying Chen, Sencun Zhu, Yilu Zhou and Heng Xu[2]

Have proposed a Lexical Syntactic Feature architecture to detect offensive content and to identify potential offensive users in social media. A hand-authoring syntactic rule is being introduced to identify the name-calling harassments. The user's potentiality to send offensive content is predicted using certain features like user's writing style, structure and specific cyberbullying content.

Pete Burnap and Matthew L. Williams[3]

Have proposed a study on online hate speech based on the massive public reaction that arouse during the murder of Drummer Lee Rigby, Woolwich, London. Human annotated Twitter data's regarding the Woolwich attack was collected to train and test a supervised machine learning text classifier that distinguishes between hateful responses that focusses religion and race. Classification features were derived from the twitter contents including grammatical dependencies between words to recognize "othering" phrases. The results of the classifier was optimal using a combination of probabilistic, rule-based, spatial- based classifiers with a voted ensemble meta-classifier.

Edel Greevy and Alan F. Smeaton[4]

Have proposed a text categorization system for PRINCIP project to automatically identify the racist text. Support Vector Machine (SVM) learning technique is used here to automatically categorize web pages and identifies whether it is racist or not.

## 3. DATA COLLECTION

### 3.1 Accessing Twitter API:

OAuth is an open standard framework for accessing the delegation, which is used by users to grant permission or applications access to their information on other websites but without giving the passwords. This mechanism is used by Twitter to permit the users to share information about their accounts with third party applications or websites without revealing their credentials. OAuth defines four major roles:

- Resource Owner: The resource owner is the one who owns an application and allows users to access their account.
- Client: The client is the application that wants to access the user's accounts.
- Resource Server: The resource server conducts the protected user accounts.
- Authorization Server: The authorization server verifies the identity of the user and then issues access tokens to the application.

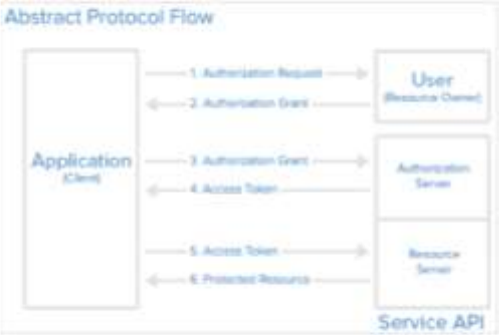
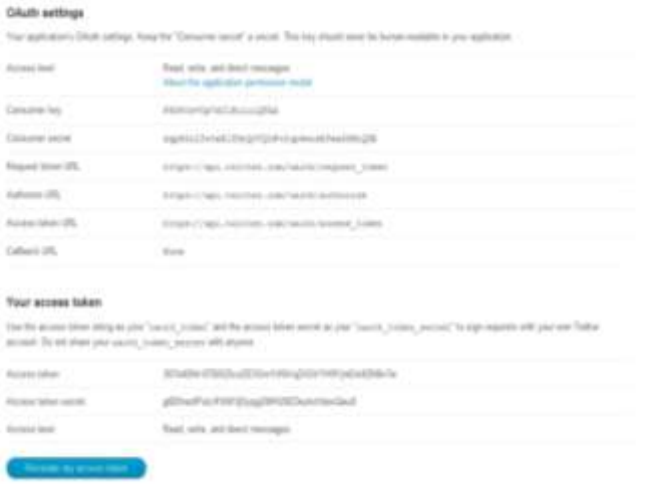


**Fig – 1** A protocol flow of OAuth

Tweepy supports accessing Twitter via OAuth. Tweepy is a library in python that enables Python to communicate with Twitter platform and use its API.



OAuth settings

[illegible]

| | |
|---|---|
| Access level | Read, write, and direct messages<br>[illegible] |
| Consumer key | [illegible] |
| Consumer secret | [illegible] |
| Request token URL | [illegible] |
| Authorize URL | [illegible] |
| Access token URL | [illegible] |
| Callback URL | None |

Your access token

[illegible]

| | |
|---|---|
| Access token | [illegible] |
| Access token secret | [illegible] |
| Access level | Read, write, and direct messages |

**Fig – 2** Access tokens

The above screenshot has the data needed to talk with Twitter Network. The main classes that are used in the Twitter API are Tweets, Users, Entities and Places. With Tweepy it is possible to get any object and use any method offered by Twitter API. Access to each returns a JSON-formatted response and traversing through information is made much easier in Python. The important tasks of Tweepy are monitoring the tweets and doing actions on it. The key component is Stream Listener this object monitors tweets in real-time and obtains them. Fig. 1. Illustrates the raw data collected from the Twitter server.

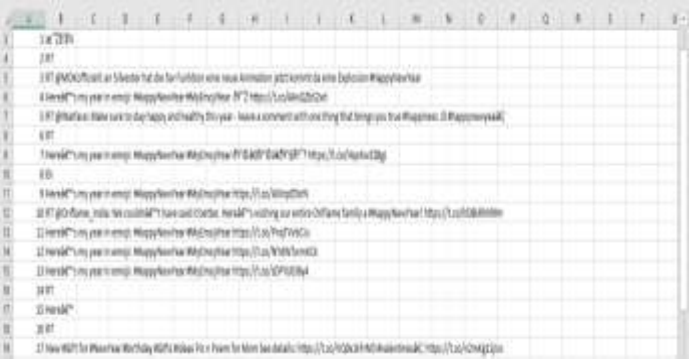


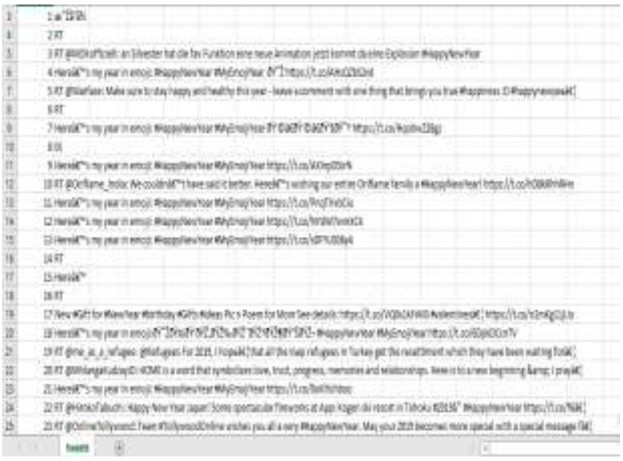
**Fig – 3** View of the raw text data collected

## 4. DATA PREPROCESSING

After collecting the raw Tweets from the Twitter data pre-processing take place. Following steps illustrates the process involved

- The fetched Tweets that are obtained in the JSON format is formatted using Pandas

- Measures are also taken to handle and display exceptions.
- A minimum of 40 tweets under this hash tag will be obtained as a csv file. Later a cleaning process is done to remove spaces, emoji's, URL and links.
- This would be obtained in another csv file.
- The missing values are also cleaned and the data is encoded. The pandas will recognize both empty cells and 'NA' types as missing values.



```
1   1 â€[illegible]
4   2 RT
5   3 RT @ARDdrofficial: an Silvester hat die [illegible] Funktion eine neue Animation jetzt kommt die eine Explosion #HappyNewYear
6   4 Hereâ€™s my year in emoji: #HappyNewYear #MyEmojiYear [illegible] https://t.co/[illegible]
7   5 RT @[illegible]: Make sure to stay happy and healthy this year - leave a comment with one thing that brings you true #happiness :D #happynewyearâ€¦
8   6 RT
9   7 Hereâ€™s my year in emoji: #HappyNewYear #MyEmojiYear [illegible] https://t.co/[illegible]
10  8 [illegible]
11  9 Hereâ€™s my year in emoji: #HappyNewYear #MyEmojiYear https://t.co/[illegible]
12  10 RT @Oriflame_India: We couldnâ€™t have said it better. Hereâ€™s wishing our entire Oriflame family a #HappyNewYear! https://t.co/[illegible]
13  11 Hereâ€™s my year in emoji: #HappyNewYear #MyEmojiYear https://t.co/[illegible]
14  12 Hereâ€™s my year in emoji: #HappyNewYear #MyEmojiYear https://t.co/[illegible]
15  13 Hereâ€™s my year in emoji: #HappyNewYear #MyEmojiYear https://t.co/[illegible]
16  14 RT
17  15 Hereâ€™
18  16 RT
19  17 New #Gift for #NewYear #Birthday #Gifts #Ideas Pic s Poem for Mom See details: https://t.co/[illegible] #valentinesâ€¦ https://t.co/[illegible]
20  18 Hereâ€™s my year in emoji: [illegible] #HappyNewYear #MyEmojiYear https://t.co/[illegible]
21  19 RT @me_as_a_refugee: @Refugees For 2019, I hopeâ€¦that all the Iraqi refugees in Turkey get the resettlement which they have been waiting forâ€¦
22  20 RT @[illegible]: HOME is a word that symbolises love, trust, progress, memories and relationships. Here is to a new beginning &amp; I prayâ€¦
23  21 Hereâ€™s my year in emoji: #HappyNewYear #MyEmojiYear https://t.co/[illegible]
24  22 RT @[illegible]: Happy New Year Japan! Some spectacular fireworks at Appi Kogen ski resort in Tohoku #2019 [illegible] #HappyNewYear https://t.co/[illegible]
25  23 RT @OnlineTollywood: Team #TollywoodOnline wishes you all a very #HappyNewYear. May your 2019 becomes more special with a special message frâ€¦
    tweets
```

**Fig – 4** The pre-processed data

## 5. IDENTIFICATION OF COMMUNAL TWEET

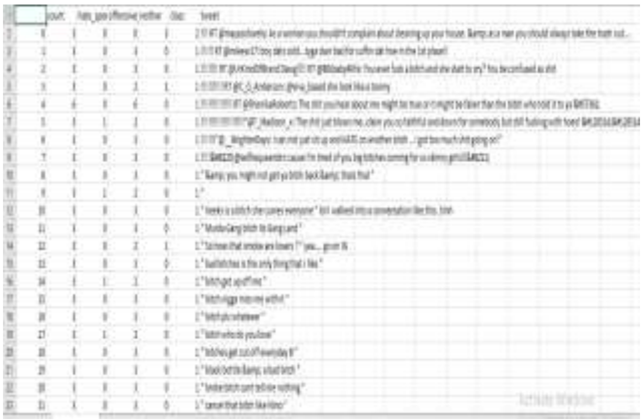
A manually prepared training dataset is used to train a supervised algorithm called Support Vector Machine learning algorithm. Supervised Learning is a machine learning method used to map an input to the output based on examples input-output pairs. It infers a function from labelled training data consisting of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (supervisory signal). A supervised learning algorithm analyses the dataset prepared for training and produces an inferred function, which can be used for mapping new examples. Supervised learning problems are grouped into Regression and Classification problems.

- **Regression –**
  When the output variable is a real value then it is said to be a regression problem Eg: such as 'dollars' or 'weight'.

- **Classification –**
  When the output is a category, then it is said to be a classification problem Eg: such as 'red' or 'blue' or 'disease' or 'no disease'.

In this paper, a supervised classification algorithm called the Support Vector Machine learning algorithm is used.

Support Vector Machine (SVM) is one among the supervised learning techniques which can be used for either classification or regression challenges. However, it is mostly used in classification problems. This algorithm is used to plot each data item as a point in n-dimensional space where n is number of features. Here in this paper three features are used. They are, hatred, offensive and neutral words. Based on these features, the algorithm classifies the tweets collected from the Twitter. The algorithm takes two inputs. It takes specific keywords list which is the dataset prepared for training the algorithm which is shown in Fig – 5 as one input and the tweet to be identified as another input. The results were plotted in a graph.



| | count | hate_spe | offensive | neither | class | tweet |
|---|---|---|---|---|---|---|
| 0 | 3 | 0 | 0 | 3 | 2 | !!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. &amp; as a man you should always take the trash out... |
| 1 | 3 | 0 | 3 | 0 | 1 | !!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!! |
| 2 | 3 | 0 | 3 | 0 | 1 | !!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit |
| 3 | 3 | 0 | 2 | 1 | 1 | !!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny |
| 4 | 6 | 0 | 6 | 0 | 1 | !!!!!!!!!!!!! RT @ShenikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who told it to ya &#57361; |
| 5 | 3 | 1 | 2 | 0 | 1 | !!!!!!!!!!!!!!!!!!"@T_Madison_x: The shit just blows me..claim you so faithful and down for somebody but still fucking with hoes! &#128514;&#128514; |
| 6 | 3 | 0 | 3 | 0 | 1 | !!!!!!"@__BrighterDays: I can not just sit up and HATE on another bitch .. I got too much shit going on!" |
| 7 | 3 | 0 | 3 | 0 | 1 | !!!!&#8220;@selfiequeenbri: cause I'm tired of you big bitches coming for us skinny girls!!&#8221; |
| 8 | 3 | 0 | 3 | 0 | 1 | " &amp; you might not get ya bitch back &amp; thats that " |
| 9 | 3 | 1 | 2 | 0 | 1 | " |
| 10 | 3 | 0 | 3 | 0 | 1 | " Keeks is a bitch she curves everyone " lol I walked into a conversation like this. Smh |
| 11 | 3 | 0 | 3 | 0 | 1 | " Murda Gang bitch its Gang Land " |
| 12 | 3 | 0 | 2 | 1 | 1 | " So hoes that smoke are losers ? " yea ... go on IG |
| 13 | 3 | 0 | 3 | 0 | 1 | " bad bitches is the only thing that i like " |
| 14 | 3 | 1 | 2 | 0 | 1 | " bitch get up off me " |
| 15 | 3 | 0 | 3 | 0 | 1 | " bitch nigga miss me with it " |
| 16 | 3 | 0 | 3 | 0 | 1 | " bitch plz whatever " |
| 17 | 3 | 1 | 2 | 0 | 1 | " bitch who do you love " |
| 18 | 3 | 0 | 3 | 0 | 1 | " bitches get cut off everyday B " |
| 19 | 3 | 0 | 3 | 0 | 1 | " black bottle &amp; a bad bitch " |
| 20 | 3 | 0 | 3 | 0 | 1 | " broke bitch cant tell me nothing " |
| 21 | 3 | 0 | 3 | 0 | 1 | " cancel that bitch like Nino " |

**Fig – 5** The training dataset

## 6. DATA VISUALIZATION

The trained model can be stored as a Python object so that one need not train each and every time we need to predict. The accuracy of the trained model is calculated before prediction. A convolution matrix is calculated. We use data visualization to present the result in a visualized manner. Fig – 6 Shows the output of the pie chart.

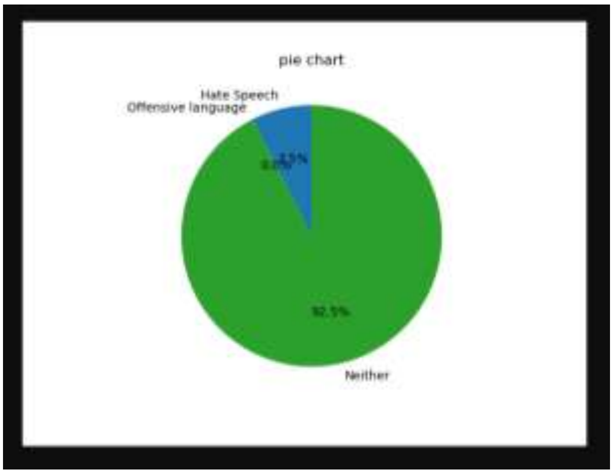


**Fig – 6** Plotted graph

## 7. CONCLUSION

This paper is the first attempt which is involved in identifying not only communal tweets but also the tweets which contains offensive or vulgar contents that has to be concealed from children below 18 years of age. It also helps in preventing conflicts that may arouse between people who belongs to different communities. This paper is a prototype built on the idea given in the paper "Characterizing and Countering Communal Microblogs during Disaster Events"[1]. Here live Tweets are obtained and pre-processed. Those cleaned Twitter data's are fetched to the machine which is trained with manual dataset using Support Vector Machine Learning Technique. A classifier is used to classify the pre-processed data. Now, a graph is displayed which gives information regarding the percentage of hatred, offensive and neutral contents available in the collected Tweets. The Tweets collected not only concentrates on particular incident or a community or a personality, but considers the Tweets randomly that does not come under any particular category and identifies communal tweets[1], offensive contents[2], online hate speech[3] and vulgar Tweets. Finally, a real-time system that automatically classifies the Tweets is proposed in this paper.

## 8. LIMITATIONS OF THE PROPOSED SYSTEM

The proposed system has some limitations as follows:

- Only the Tweets in English are taken into account. This system cannot be applied to words not present in the English dictionaries. Those words are ignored in this process, which is one of the major limitation faced in this paper.
- Some Tweets may contain words with improper spellings, abbreviations, emoji's are just ignored while pre-processing the tweets, which can be treated.
- A graph is being displayed to the user on the basis of his Tweet and it can be removed only by the user, even if it is found to be offensive.

## 9. FUTURE ENHANCEMENTS

- It can be used by the Government in taking decisions like regarding eliminating the troublesome tweet, and find a solution to stop the problems that arise in the society.
- The communal, offensive, hatred or vulgar Tweets can be replaced with neutral Tweets, so that it does not create any problems or clashes in the society.
- It would be effective if such trouble causing Tweets can be blocked immediately.
- It would be more effective, when the user is intimated while typing an offensive or a vulgar content and not allowing to proceed further.

## 10. REFERENCES

[1] K. Rudra, A. Sharma, N. Ganguly, and S. Ghosh, "Characterizing and Countering communal microblogs during disaster events," IEEE Transactions on Computational Social Systems 5 (2), 403-417

[2] 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.

[3] 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.

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

[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] 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] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh, "Extracting situational information from microblogs during disaster events: A classification-summarization approach," in Proc. ACM CIKM, 2015, pp. 583–592.

[8] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, Oct. 2011.