Abusive Text Mining on Twitter

Deven Shah[1], Kuntal Surwade[2], Rohan Shah[3], Pranav Thakkar[4]

[1,2,3,4]*Student, Information Technology Department, Shah & Anchor Kutchhi Engineering College, Mumbai, India*

Abstract—**In social media, there has been tremendous increment in profaning the victims, tweeting abusive contents to make victims profile a notorious one. This have immensely affected the sentiments of the users on the social media like twitter. To counterpart the derogatory done by the perpetrators, we have proposed to detect offensive content and identify potential offensive users in social media like twitter. This preventive approach can provide information about users who are being targeted and can be used to build monitoring tools to aid finding and stopping potential bullies. We have used python as our base language & applied various filtering algorithms to remove unwanted tweets and other anomalies from the real-time data. The content which are filtered are then lemmatize for preprocessing the data being gathered at real-time. This in return will display the offensive contents posted by the user onto the victim's post or tweets in real-time. Therefore, we have attempted to resolve the profanity carried out on twitter in live environment.**

Index Terms— Bernoulli, Classification, lemmatization, Naïve Bayes, NLTK, SVC, Tokenization

I. Introduction

We proposed a new approach to deal with the offensive content on Twitter using text classification algorithms. We have made an attempt to contemplate the user’s tweets and report the same to him/her. For this purpose, we have developed interface for the Twitter admin to monitor the tweets at real time environment.

The model is developed using Python as base language, applying numerous libraries, modules, classes and functions to classify the tweets as for abusive text mining. The system operates with the successful login of the administrator on our interface developed. The user can have sign-in from his mobile or desktop device to the twitter account. Nevertheless, we have no interference in the login of any user.

The user, after successful login into his/her twitter account can then tweets on any subject, matter, or any other subject. On back-end, the Admin can monitor the tweets posted by any user. After the tweet is posted, the admin can check with the training data sets. If any abusive content is found, the user account is flagged and the same is reported to him/her.

II. Proposed System

The main motive of the system is to detect the abusive content tweeted by a user to bully the victim. To extract the run-time data from twitter, we have taken the Twitter API access token secret key, consumer key, private key and public key to gain the real-time tweet access. This API allows the administrator to extract the user’s tweets. After this connection, the python script is executed; the admin will enter any keyword for extraction of the real-time tweets.

After the extraction procedure, the administrator make preprocessing step to further remove the anomalies of the tweets. Once anomalies are removed, the sentimental analysis is performed on those tweets. In this analysis, our focus is on the negative tweets posted by the users. After extracting negative tweets, preprocessing is done. In preprocessing, word tokenization and lemmatization is used to filter out tweets into arrays of words. Lemmatization method is used on those words which help to get root word. These words are then compared with the abusive dataset to detect any abusive words.

To check the presence of profanity, these words are stored in the database and then the same are evaluated along with the dataset present in the back-end i.e. administration side. If the words are matched with the dataset, the users' twitter account is reported. In result, a summary is generated at run-time environment disclosing the tweet time, user's name, tweet that he/she posted, abusive words and location (optional).

So, to classify whether tweets are positive or negative we have used Naive-Bayesian Bernoulli classifier. The algorithm generates two kinds of classification based on our criteria - negative & positive. It uses SVC to plot each data item in n-dimensional space (where n is variety of options you have) with the worth of every feature being the worth of a specific coordinate. This in result distributes the tweets as containing abusive content or a regular tweet. Then, to determine the keywords of the tweet posted, tokenization is done. Tokenization is used to split the words individually from the whole tweets (sentences). To make the word generic, lemmatization is preferred. Lemmatization makes the word to its root form. Thus, it helps in determining the profanity presence in the tweet which can be compared with the dataset.

Our flow of the proposed system is given below illustrating the Activity diagram and Flow diagram to make the concept well understood. Actors used in the modelling are also described. The easiest way to prepare your document is to use this document as a template and simply type your text into it.

III. Design and implementation

The figure-1 activity diagram shows when the admin search corresponding to the keyword, the tweets are filtered out & stored the matched ones into the DB. After that, evaluation is done & results are presented on the screen.

The actors used in the system developed shows their respective functionalities in the Use case diagram in figure-2. Actors like Users, Admin and the component Dataset are the key factors of it.

The detailed flow diagram in figure-3, illustrates the phases and the roles by various actors present. The major chunk is the pre-processing activity wherein tweets are filtered out, tokenized and then compared with the dataset.

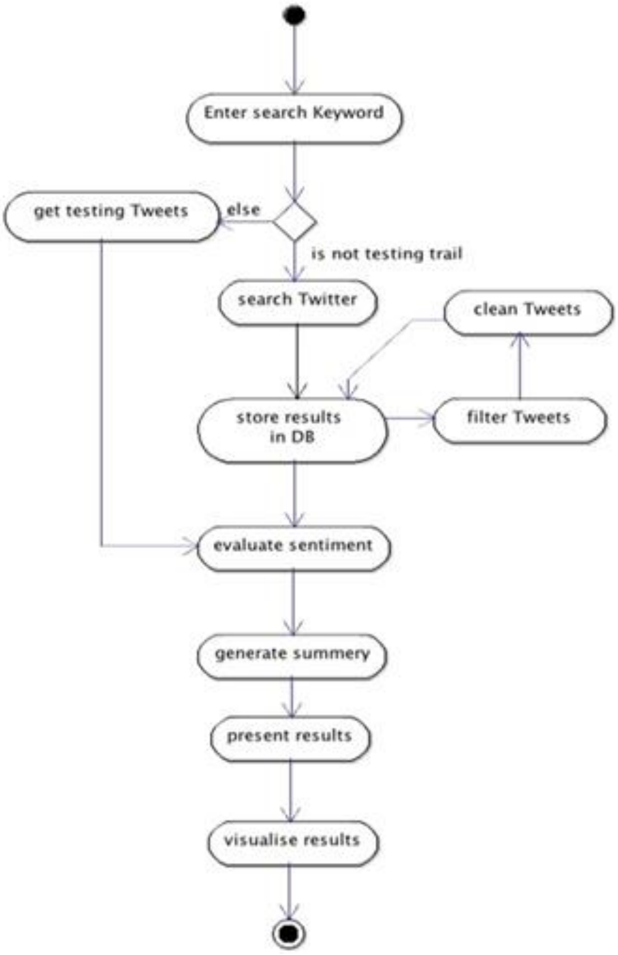

Fig. 1. Activity diagram

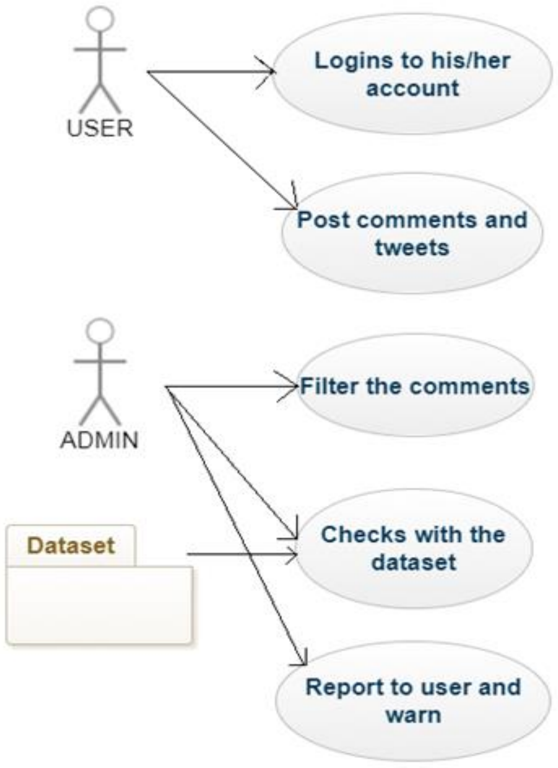

Fig. 2. Use Case diagram

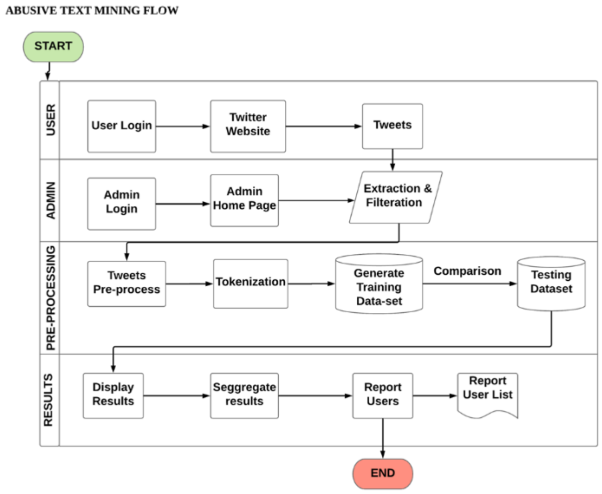

Fig. 3. Flow Diagram

IV. Conclusion

The output of this project solves the issues of online bullying and harassment which stress on the text-mining based on the data from the social media. The project enhances user experience which could be a huge boost commercially expanding horizons in this field. Its implementation from user perspective is relatively simple and easy to use.

Acknowledgement

We wish to express our profound gratitude to our project guide Mr. Lukesh Kadu for allowing us to go ahead with this project and giving us the opportunity to explore in this domain. We would also like to thank I/C Head of Department Ms. Swati Nadkarni for constant encouragement and support towards achieving this goal. We take this opportunity to express our profound gratitude and deep regards to our co-guide Ms. Nida Jawre for her exemplary guidance, monitoring and constant encouragement throughout the course of this project. We would also like to thank Ms. Manya Gidwani for her invaluable suggestions and feedback without whom our work would have been very difficult.

No project is ever complete without the guidelines of these experts who have already established a mark on this path before and have become masters of it. So we would like to take this opportunity to thank all those who have helped us in implementing this project.

Acronyms

SVC	Support Vector Classifier
NB	Naïve Bayesian
NLP	Natural Language Processing
DB	Database

References

[1] N. Pendar, "Toward spotting the pedophile telling victim from predator in text chats," in Proceedings of the First IEEE International Conference on Semantic Computing, 2007, pp. 235-241.

[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. April 2015.

[3] Y. Chen, Y. Zhou, S. Zhu and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, Amsterdam, 2012, pp. 71-80.

[4] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," *CEAS 2010 Seventh annual Collaboration, Electronic messaging, Anti Abuse and Spam Conference*, July 13-14, 2010, Redmond, Washington, US.

[5] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive Language Detection in Online User Content," *WWW '16 Proceedings of the 25th International Conference on World Wide Web*, pp. 145-153.

[6] W. Warner, and J. Hirschberg, "Detecting hate speech on the world wide web," *LSM '12 Proceedings of the Second Workshop on Language in Social Media*, pp. 19-26.

[7] R. M. Kowalski, S. P. Limber, and P. W Agatston, "Cyberbullying: Bullying in the digital age," John Wiley & Sons, 2nd Edition, 2012.

[8] S. Bauman, R. B. Toomey, and J. L. Walker, "Associations among bullying, cyberbullying, and suicide in high school students," *Journal of adolescence*, vol. 36, no. 2, pp. 341-350, April 2013.

[9] B. Fitzgerald, "Bullying on Twitter: Researchers find 15,000 bully-related tweets sent daily (study).
Source: http://www.huffingtonpost.com/2012/08/02/bullying-on twitter_n_1732952.html

[10] R. Slonje, P. K. Smith, and A. Frisén, "The nature of cyberbullying, and strategies for prevention," *Computers in Human Behavior*, vol. 29, no. 1, pp. 26-32, 2013.

[11] Cyberbullying Statistics: Bullying facts, bullying statistics, 2014.
Source: http://nobullying.com/cyber-bullying-statistics-2014/