International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Analysis of Women Safety using Machine Learning on Tweets

VIKRAM CHANDRA[1], RAMPUR SRINATH[2]

[1]*Student, MTECH-Computer Network and Engineering, Department of IS&E, The National Institute of Engineering, Mysore, Karnataka, India*
[2]*Associate Professor, Department of IS&E, The National Institute of Engineering, Mysore, Karnataka, India*

Abstract - *Nowadays women are experiencing lots of violence such as harassment in places in several cities. This starts from stalking which then leads to abusive harassment or also called abuse assault. In this paper we mainly focus on the role of social media which can be used to promote the safety of women in India, given the special reference to the participation of many social media websites or applications such as Twitter, Facebook and Instagram platforms. This paper also focuses on developing the responsibilities among the common people on the various parts of Indian cities so that the safety of women around them is ensured. Tweet on the Twitter application contains the text messages, audio data, video data, images, smiley expressions and hash-tags. This tweet content can be used to read among the people and thus can educate them in order to take strict actions if tweets are abusive to women and hence can punish such people if the harassment is made. Applications which include hash-tags, such as Twitter and Instagram, can be used to spread the messages across the entire globe and make the women feel free to express their views and feelings. By this we can know the state of their mind when they go out for work or travel in a public transportation or surrounded by anonymous men and whether it feels they are secure or not.*

***Key Words*: Women, Safety, Sexual Harassment, Hash tag, Sentimental Analysis.**

1. INTRODUCTION

Twitter in this modern era has emerged as a ultimate micro-blogging social network consisting over hundred million users and generate over five hundred million messages known as 'Tweets' every day. Twitter with such a massive audience has magnetized users to emit their perspective and judgemental about every existing issue and topic of internet, therefore twitter is an informative source for all the zones like institutions, companies and organizations.

On the twitter, users will share their opinions and perspective in the tweets section. This tweet can only contain 140 characters, thus making the users to compact their messages with the help of abbreviations, slang, shot forms, emoticons, etc. In addition to this, many people express their opinions by using polysemy and sarcasm also. Hence twitter language can be termed as the unstructured. From the tweet, the sentiment behind the message is extracted. This extraction is done by using the sentimental analysis procedure. Results of the sentimental analysis can be used in many areas like sentiments regarding a particular brand or release of a product, analyzing public opinions on the government policies, people thoughts on women, etc.
In order to perform classification of tweets and analyze the outcome, a lot of study has been done on the data obtained by the twitter. We also review some studies on machine learning in this paper and research on how to perform sentimental analysis using that domain on twitter data. The paper scope is restricted to machine learning algorithm and models.

Staring at women and passing comments can be certain types of violence and harassments and these practices, which are unacceptable, are usually normal especially on the part of urban life. Many researches that have been conducted in India shows that women have reported sexual harassment and other practices as stated above. Such studies have also shown that in popular metropolitan cities like Delhi, Pune, Chennai and Mumbai, most women feel they are unsafe when surrounded by unknown people. On social media, people can freely express what they feel about the Indian politics, society and many other thoughts. Similarly, women can also share their experiences if they have faced any violence or sexual harassment and this brings innocent people together in order to stand up against such incidents. From the analysis of tweets text collection obtained by the twitter, it includes names of people who has harassed the women and also names of women or innocent people who have stood against such violent acts or unethical behaviour of men and thus making them uncomfortable to walk freely in public.

The data set of the tweet will be used to process the machine learning algorithms and models. This algorithm will perform smoothening the tweet data by eliminating zero values. Using Laplace and porter's theory, a method is developed in order to analyze the tweet data and remove redundant information from the data set. Huge numbers of people have been attracted to social media platform such as Twitter, Facebook, Instagram. People express their sentiments about society, politics, women, etc via the text messages, emoticons and hash-tags through such platforms. There are some methods of sentiment that can be classified like machine leaning based and lexicon based learning.

2. SENTIMENTAL ANALYSIS

Sentiment analysis is the process of extracting the sentiment behind any sentence or statement. It can be called as a classification technique which is used to obtain the opinion from tweet.

This opinion is useful in formulating a sentiment which can further be used to achieve sentiment classification. Sentiments are personal to the topic and thus we need to decide what kind of specifications is formulated out of it.
Person performing the sentimental analysis wants to find the class of entities of the tweets using the programming model. The dimension of the sentimental class is an important factor in order to decide the efficiency of the algorithm. For instance, there can be two class sentimental classification of tweets – Positive and Negative or there can be three class classification – Positive, Negative and Neutral. Approaches of sentimental analysis can be broadly differentiated into two types – machine learning based and lexicon learning based. Machine learning approach includes the process of extraction of features, programming model training using dataset of features. Whereas lexicon learning based approach uses the vocabulary and scoring method to detect opinions. In this paper, we use machine learning approach. Collection of data, pre-processing the data, extraction of features, choosing base features, detection of sentiments and classification of sentiments using machine learning approaches or simple computations are the basic steps to perform sentimental analysis.

2.1 Analysis of Sentimental Data

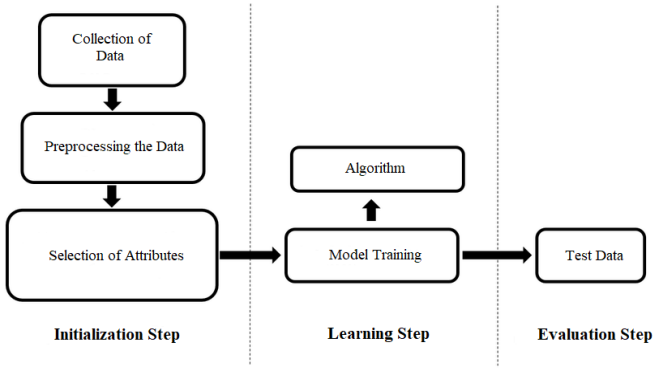

Fig -1: Process of Analysis

The process of obtaining the sentiments of tweet includes five steps:

1) Data extraction: First step involved in analysis of sentiment is the collection of information from the social network website like twitter. This helps in extracting the tweet message but this message also includes extra data like tweets likes, dislikes and comments.
2) Text Cleaning: Once the data is extracted from the twitter source as the datasets, this information has to be passed to the classifier. The classifier cleans the dataset by removing redundant data like stop words, emoticons in order to make sure that non textual content is identified and removed before the analysis.
3) Sentiment Analysis: After the classifier cleans the dataset, the data is ready for the sentimental analysis process. Machine learning and Lexicon based learning and Hybrid learning are some of the approaches of sentimental analysis. There are also some other approaches such as Nero Linguistic Programming and Natural Language Processing. Training the dataset and then testing that trained dataset involves in machine learning approach. Training data and Testing data are useful for the classifier to perform the algorithm. Maximum Entropy, Naives Bayes classification, Bayesian Networks and Network Support Vector Machine are some of the algorithm which can be used to train the classifier. Testing data is used to identify the efficiency of the sentiment classifier.
In case of Lexicon based leaning, training dataset is not used. This approach uses a built-in dictionary in which words associated with sentiments of human are present. The third approach, which is the Hybrid learning, combines both machine leaning approach and lexicon learning approach in order to improve the performance of classifier.
4) Sentiment Classification: At this step, the dataset is ready for the classification. Each and every sentence of the tweet will be examined and opinion will be formed accordingly for subjectivity. Subjective expression sentences are retained and those of objective expression sentences are rejected. Techniques like Unigrams, Negation, Lemmas and so on are used at different levels of sentimental analysis. Sentiments can be distinguished broadly into two groups – Positive and Negative. At this point of sentimental analysis, each of the subjective sentences which will be retained are classified into good, bad or like, dislike or positive and negative.
5) Output Presentation: To generate useful and meaningful information out of the raw data, sentimental analysis plays vital role. Once the algorithm is completed, the outcome of the analysis can be visualized by creating different types of graphs. Bar graphs, Time series and Pie charts are some of the examples which can be used to display the output. To measure the sentiment of the tweets in terms of Positive and Negative, Bar graphs can be used. Similarly, to measure in terms of likes, dislikes, average length of tweet for a certain period, Time series can be used. To obtain the initial source of the tweet, pie charts can be used.

2.2 System Architecture

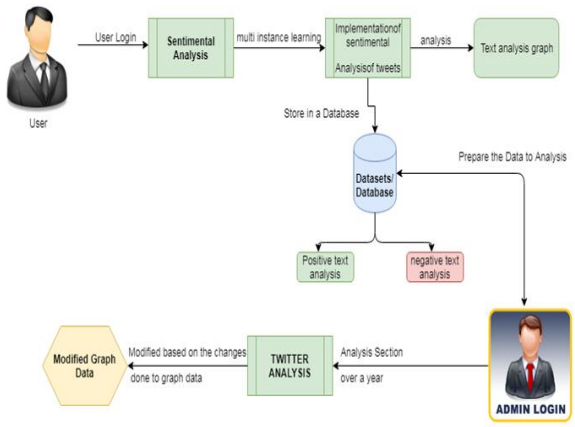

Fig -2: Architecture

Every user data such as credentials, new tweets, re-tweets and tweet score will be stored in the database for the admin to monitor and perform the analysis. The sentiment analysis is applied on the user data in order to monitor and confirm whether any tweets are abusive to women or not. Admin performs this analysis on each and every user tweets to provide safety for the women. Sentimental analysis will be implemented on the tweets of user that are stored in the database. Admin can now prepare the data to perform the analysis. The tweets made by every user of the application will be called as the initial input for the sentiment analysis and hence they will be the dataset. Along with this, text analysis graph can also be shown. Admin will store the filters in the database. Filters are the keywords for which the tweet context will be searched for in order to declare as abusive or not. There can be two types of filters – positive keyword and negative keyword. Positive keywords are those words which are abusive or disrespect the women by any means. Negative keywords are the words which are normal and will not abuse the women.

There can be 'n' number of positive and negative keywords stored in the database. When the admin implements the sentimental analysis, every keyword in the database will be compared with each and every word in the tweet of the user. If any one of the positive keyword is found in the tweet, that tweet will be classified as positive sentimental analysis and these are abusive to women. If negative keyword is found in the tweet, it will be classified as the negative sentimental analysis which is not abusive to women. Hence, by this stage there will be two types of sentimental analysis made based on the filter in the database. Under positive sentimental analysis, there will be a list of all the tweets in the application that are abusive to women. Similarly, under negative sentimental analysis there will be a list that is clean and are not abusive tweets. Along with the tweet context, user details will also be provided at each of the analysis list.

3. CONCLUSION AND FUTURE WORK

Machine learning algorithm has been discussed throughout the project. For the twitter data that includes millions of tweet and messages every day, machine learning algorithm helps to organize and perform analysis. SPC algorithm, linear algebraic are some of the algorithms which are effective in analyzing the large data that provide categorization and convert into meaningful datasets. Hence we can perform machine learning algorithms to achieve sentimental analysis and bring more safety to women by spreading the awareness.

For the future enhancement, we can extend to apply these machine learning algorithms on different social media platforms like facebook and instagram also since in our project only twitter is considered. Present ideology which is proposed can be integrated with the twitter application interface to reach larger extent and apply sentimental analysis on millions of tweet to provide more safety.

REFERENCES

[1] Apoorv Agarwal, Fadi Biadsy, and Kathleen R. Mckeown. "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009.

[2] Luciano Barbosa and Junlan Feng. "Robust sentiment detection on twitter from biased and noisy data." Proceedings of the 23rd international conference on computational linguistics: posters. Association for Computational Linguistics, 2010.

[3] Adam Bermingham and Alan F. Smeaton. "Classifying sentiment in microblogs: is brevity an advantage?." Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[4] Michael Gamon. "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.

[5] Soo-Min Kim and Eduard Hovy. "Determining the sentiment of opinions." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.

[6] Dan Klein and Christopher D. Manning. "Accurate unlexicalized parsing." Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003.

[7] Eugene Charniak and Mark Johnson. "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking." Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2005.

[8] Gupta B, Negi M, Vishwakarma K, Rawat G & Badhani P (2017). "Study of Twitter sentiment analysis using machine learning algorithms on Python." International Journal of Computer Applications, 165(9) 0975-8887.

[9] Sahayak V, Shete V & Pathan A (2015). "Sentiment analysis on twitter data." International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2(1), 178-183.

[10] Mamgain N, Mehta E, Mittal A & Bhatt G (2016, March). "Sentiment analysis of top colleges in India using Twitter data." In Computational Techniques, in Information and Communication Technologies (ICCTICT), 2016 International Conference on (pp. 525-530). IEEE.