Journal of Engineering Sciences

Vol 13, Issue 05, MAY/2022

ISSN NO:0377-9254

ANALYSIS OF WOMEN SAFETY IN INDIAN CITIES USING MACHINE LEARNING ON TWEET

Syeda surrayya jabeen, Assistant professor, MTech, Department of CSE, syedasurayyaj@gmail.com

Sheema Nargis,BE, Department of CSE, 160518733085@islec.edu.in

Saadiya,BE, Department of CSE, 160518733090@islec.edu.in

Asma Namera,BE, Department of CSE, 160518733096@islec.edu.in

ABSTRACT: Women and girls have been experiencing a lot of violence and harassment in public places in various cities starting from stalking and leading to abuse harassment or abuse assault. This research paper basically focuses on the role of social media in promoting the safety of women in Indian cities with special reference to the role of social media websites and applications including Twitter platform Facebook and Instagram. This paper also focuses on how a sense of responsibility on part of Indian society can be developed the common Indian people so that we should focus on the safety of women surrounding them. Tweets on Twitter which usually contains images and text and also written messages and quotes which focus on the safety of women in Indian cities can be used to read a message amongst the Indian Youth Culture and educate people to take strict action and punish those who harass the women. Twitter and other Twitter handles which include hash tag messages that are widely spread across the whole globe sir as a platform for women to express their views about how they feel while we go out for work or travel in a public transport and what is the state of their mind when they are surrounded by unknown men and whether these women feel safe or not?.

1. INTRODUCTION

There are certain types of harassment and Violence that are very aggressive including staring and passing comments and these unacceptable practices are usually seen as a normal part of the urban life. There have been several studies that have been conducted in cities across India and women report similar type of sexual harassment and passing off comments by other unknown people. The study that was conducted across most popular Metropolitan cities of India including Delhi, Mumbai and Pune, it was shown that 60 % of the women feel unsafe while going out to work or while travelling in public transport. Women have the right to the city which means that they can go freely whenever they want whether it be too an Educational Institute, or any other place women want to go. But women feel that they are unsafe in places like malls, shopping malls on their way to their job location because of the several unknown Eyes body shaming and harassing these women point Safety or lack of concrete consequences in the life of women is the main reason of harassment of girls. There are instances when the harassment of girls was done by their neighbours while they were on the way to school or there was a lack of safety that created a sense of fear in the minds of small girls who throughout their lifetime suffer due to that one instance that happened in their lives where they were forced to do something unacceptable or was sexually harassed by one of their own neighbor or any other unknown person. Safest cities approach women safety from a perspective of women rights to the affect the city without fear of violence or sexual harassment. Rather than imposing restrictions on women that society usually imposes it is the duty of society to imprecise the need of protection of women and also recognizes that women and girls also have a right same as men have to be safe in the City.

Analysis of twitter texts collection also includes the name of people and name of women who stand up against sexual harassment and unethical behaviour of men in Indian cities which make them uncomfortable to walk freely. The data set that was obtained through Twitter about the status of women safety in Indian society was for the processed through machine learning algorithms for the purpose of smoothening the data by removing zero values and using Laplace and porter's theory is to developer method of analyzation of data and remove retweet and redundant data from the data set that is obtained so that a clear and original view of safety status of women in Indian society is obtained.

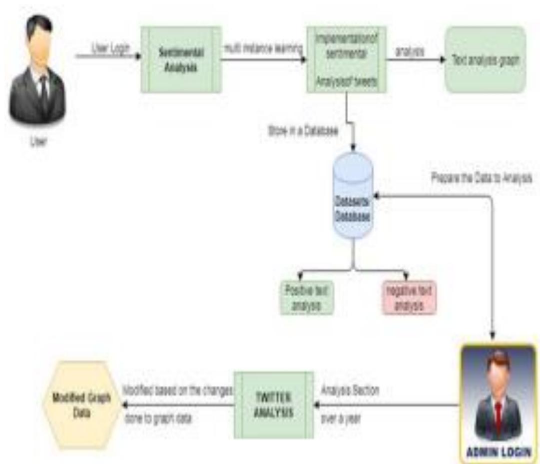

Fig.1: Example figure

Twitter in this modern era has emerged as a ultimate microblogging 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.

2. LITERATURE REVIEW

2.1 Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams :

We present a classifier to predict contextual polarity of subjective phrases in a sentence. Our approach features lexical scoring derived from the Dictionary of Affect in Language (DAL) and extended through WordNet, allowing us to automatically score the vast majority of words in our input avoiding the need for manual labeling. We augment lexical scoring with n-gram analysis to capture the effect of context. We combine DAL scores with syntactic constituents and then extract n-grams of constituents from all sentences. We also use the polarity of all syntactic constituents within the sentence as features. Our results show significant improvement over a majority class baseline as well as a more difficult baseline consisting of lexical n-grams..

2.2 Robust sentiment detection on twitter from biased and noisy data:

In this paper, we propose an approach to automatically detect sentiments on Twitter messages (tweets) that explores some characteristics of how tweets are written and meta-information of the words that compose these messages. Moreover, we leverage sources of noisy labels as our training data. These noisy labels were provided by a few sentiment detection websites over twitter data. In our experiments, we show that since our features are able to capture a more abstract representation of tweets, our solution is more effective than previous ones and also more robust regarding biased and noisy data, which is the kind of data provided by these sources..

2.3 Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis:

We demonstrate that it is possible to perform automatic sentiment classification in the very noisy domain of customer feedback data. We show that by using large feature vectors in combination with feature reduction, we can train linear support vector machines that achieve high classification accuracy on data that present classification challenges even for a human annotator. We also show that, surprisingly, the addition of deep linguistic analysis features to a set of surface level word n-gram features contributes consistently to classification accuracy in this domain.

2.4 Study of Twitter sentiment analysis using machine learning algorithms on Python:

Twitter is a platform widely used by people to express their opinions and display sentiments on different occasions. Sentiment analysis is an approach to analyze data and retrieve sentiment that it embodies. Twitter sentiment analysis is an application of sentiment analysis on data from Twitter (tweets), in order to extract sentiments conveyed by the user. In the past decades, the research in this field has consistently grown. The reason behind this is the challenging format of the tweets which makes the processing difficult. The tweet format is very small which generates a whole new dimension of problems like use of slang, abbreviations etc. In this paper, we aim to review some papers regarding research in sentiment analysis on Twitter, describing the methodologies adopted and models applied, along with describing a generalized Python based approach.

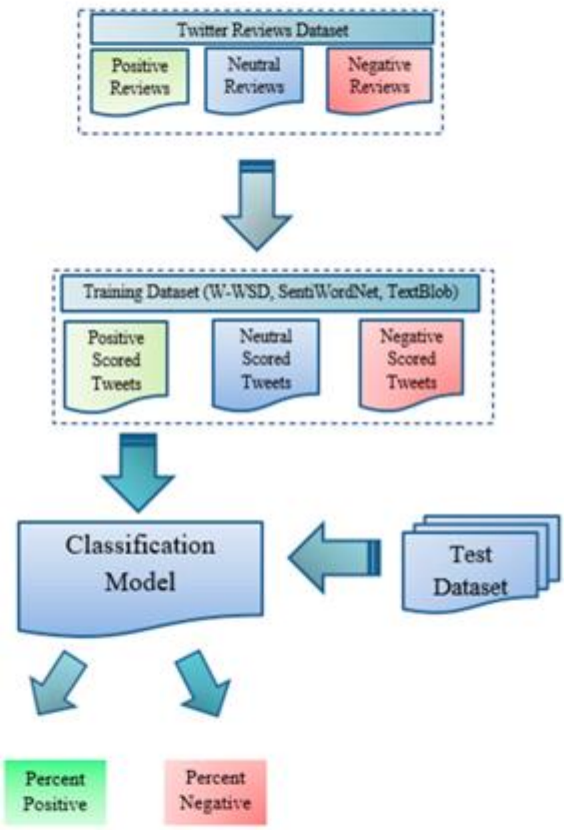

Fig.2: Twitter sentiment analysis.

2.5 Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams:

We present a classifier to predict contextual polarity of subjective phrases in a sentence. Our approach features lexical scoring derived from the Dictionary of Affect in Language (DAL) and extended through WordNet, allowing us to automatically score the vast majority of words in our input avoiding the need for manual labeling. We augment lexical scoring with n-gram analysis to capture the effect of context. We combine DAL scores with syntactic constituents and then extract ngrams of constituents from all sentences. We also use the polarity of all syntactic constituents within the sentence as features. Our results show significant improvement over a majority class baseline as well as a more difficult baseline consisting of lexical n-grams.

Vol 13, Issue 05, MAY/2022

ISSN NO:0377-9254

2.6 Robust sentiment detection on twitter from biased and noisy data:

In this paper, we propose an approach to automatically detect sentiments on Twitter messages (tweets) that explores some characteristics of how tweets are written and meta-information of the words that compose these messages. Moreover, we leverage sources of noisy labels as our training data. These noisy labels were provided by a few sentiment detection websites over twitter data. In our experiments, we show that since our features are able to capture a more abstract representation of tweets, our solution is more effective than previous ones and also more robust regarding biased and noisy data, which is the kind of data provided by these sources.

2.7 Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis:

We demonstrate that it is possible to perform automatic sentiment classification in the very noisy domain of customer feedback data. We show that by using large feature vectors in combination with feature reduction, we can train linear support vector machines that achieve high classification accuracy on data that present classification challenges even for a human annotator. We also show that, surprisingly, the addition of deep linguistic analysis features to a set of surface level word n-gram features contributes consistently to classification accuracy in this domain.

3. IMPLEMENTATION

Concept to analyze women safety using social networking messages and by applying machine learning algorithms on it. Now-a-days almost all peoples are using social networking sites to express their feelings and if any women feel unsafe in any area then she will express negative words in her post/tweets/messages and by analysing those messages we can detect which area is more unsafe for women's. But women feel that they are unsafe in places like malls, shopping malls on their way to their job location because of the several unknown Eyes body shaming and harassing these women point.

In propose work author using TWEEPY package from python to download tweets from twitter but every time INTERNET will not available to download tweets online so we downloaded MEETOO tweets on women safety and safe inside dataset folder. Application will read this tweets to detect women's sentiments. Author using NLTK (natural language tool kit) to remove special symbols and stop words from tweets and to make them clean. Author using TEXTBLOB corpora package and dictionary to count positive, negative and neutral polarity and the tweets which has polarity value less than 0 will consider as negative as and greater than 0 and less than 0.5 will consider as neutral and polarity greater than 0.5 will consider as positive.

Advantages:

1. Analysis of twitter texts collection also includes the name of people and name of women who stand up against abuse harassment and unethical behaviour of men in Indian cities which make them uncomfortable to walk freely.

2. The data set that was obtained through Twitter about the status of women safety in Indian society.

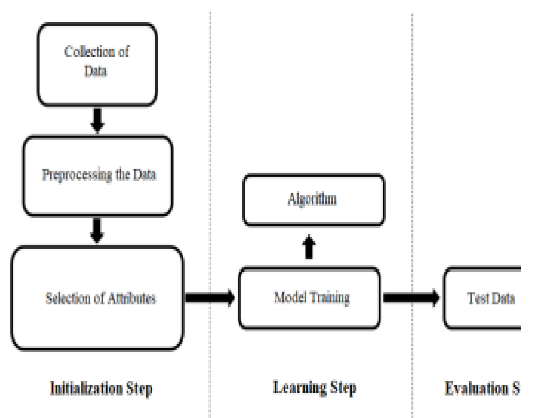

Fig.3: System architecture

MODULES:

•upload dataset: using this module we will upload dataset

•Dataset cleaning: using this module we will find out empty values in the dataset and replace with mean or 0 values.

•Train & Test Split: Using this module we will split dataset into two parts called and training and testing. All machine learning algorithms take 80% dataset to train classifier and 20% dataset is used to test classifier prediction accuracy.

4. MACHINE LEARNING

A machine learning algorithm is the method by which the AI system conducts its task, generally predicting output values from given input data. The two main processes of machine learning algorithms are classification and regression.

Machine learning (ML) algorithms are broadly categorized as either supervised or unsupervised. Supervised learning algorithms have both input data and desired output data provided for them through labeling, while unsupervised algorithms work with data that is neither classified nor labeled. An unsupervised algorithm might, for example, group unsorted data according to similarities and differences.

However, many ML approaches, including transfer learning and active learning, involve what are more accurately described as semi-supervised algorithms. Transfer learning uses knowledge gained from completing one task to help solve a different but related problem, while active learning allows an algorithm to query the user or some other source for more information. Both systems are commonly used in situations where labeled data is scant.

Reinforcement learning, sometimes considered a fourth category, is based on rewarding desired behaviors and/or punishing undesired ones to direct unsupervised machine learning through rewards and penalties.

Supervised learning vs. unsupervised learning:

Independent of these divisions, there are another two kinds of machine learning algorithms: supervised and unsupervised. In supervised learning, you provide a training data set with answers, such as a set of pictures of animals along with the names of the animals. The goal of that training would be a model that could correctly identify a picture (of a kind of animal that was included in the training set) that it had not previously seen.

In unsupervised learning, the algorithm goes through the data itself and tries to come up with meaningful results. The result might be, for example, a set of clusters of data points that could be related within each cluster. That works better when the clusters don't overlap.

Training and evaluation turn supervised learning algorithms into models by optimizing their parameters to find the set of values that best matches the ground truth of your data. The algorithms often rely on variants of steepest descent for their optimizers, for example stochastic gradient descent (SGD), which is essentially steepest descent performed multiple times from randomized starting points. Common refinements on SGD add factors that correct the direction of the gradient based on momentum or adjust the learning rate based on progress from one pass through the data (called an epoch) to the next.

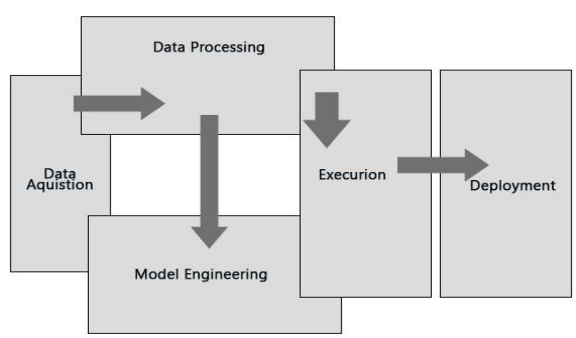

Fig.4: Machine learning algorithms process steps

There are dozens of machine learning algorithms, ranging in complexity from linear regression and logistic regression to deep neural

Journal of Engineering Sciences
Vol 13, Issue 05, MAY/2022
ISSN NO:0377-9254

networks and ensembles (combinations of other models). However, some of the most common algorithms include:

- Linear regression, aka least squares regression (for numeric data)
- Logistic regression (for binary classification)
- Linear discriminant analysis (for multi-category classification)
- Decision trees (for both classification and regression)
- Naïve Bayes (for both classification and regression)
- K-Nearest Neighbors, aka KNN (for both classification and regression)
- Learning Vector Quantization, aka LVQ (for both classification and regression)
- Support Vector Machines, aka SVM (for binary classification)
- Random Forests, a type of "bagging" ensemble algorithm (for both classification and regression)
- Boosting methods, including AdaBoost and XGBoost, are ensemble algorithms that create a series of models where each new model tries to correct errors from the previous model (for both classification and regression)

Where are the neural networks and deep neural networks that we hear so much about? They tend to be compute-intensive to the point of needing GPUs or other specialized hardware, so you should use them only for specialized problems, such as image classification and speech recognition, that aren't well-suited to simpler algorithms. Note that "deep" means that there are many hidden layers in the neural network.

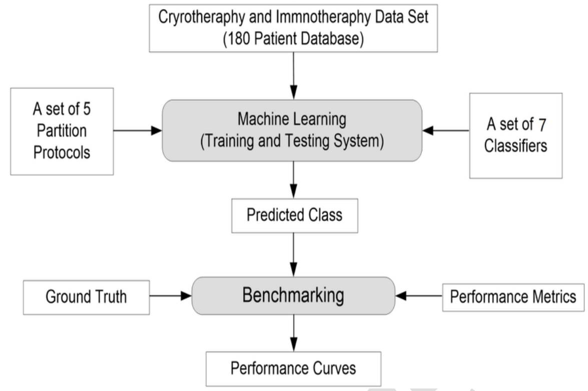

Fig.5: Machine learning algorithms architecture

5. EXPERIMENTAL RESULTS

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.

Journal of Engineering Sciences

Vol 13, Issue 05, MAY/2022

ISSN NO:0377-9254

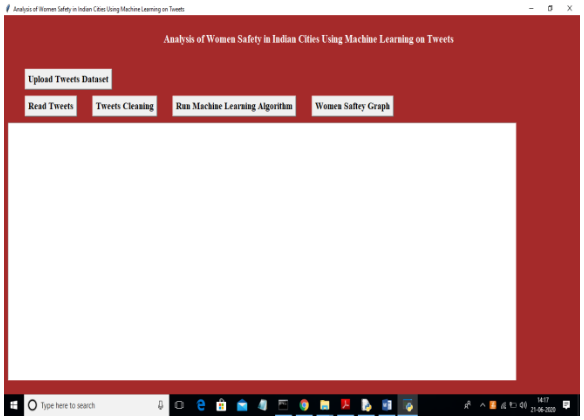

Fig.6: Home screen

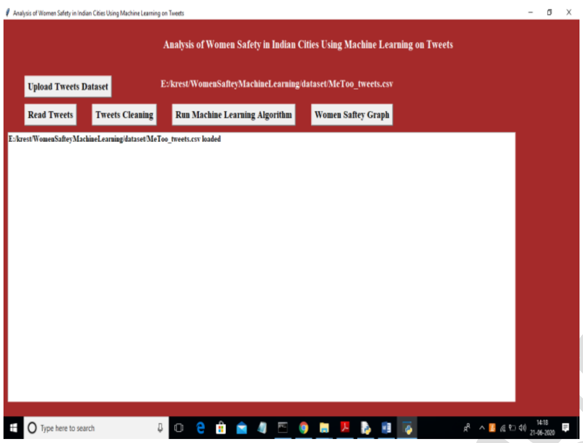

Fig.7: Data loading screen

In above screen uploading MeeToo_tweets.csv file and then click on 'Open' button to load dataset

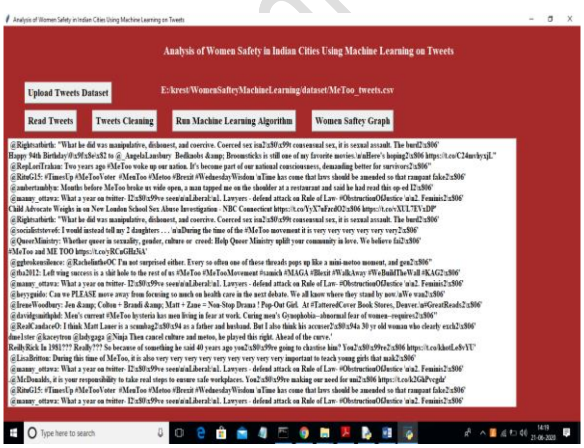

Fig.8: Read tweets

In above screen each line represents one tweet and you can scroll down above screen text area to view all tweets. In above screen we can see all tweets contains special symbols and stop words and to clean those tweets click on 'Tweets Cleaning' button

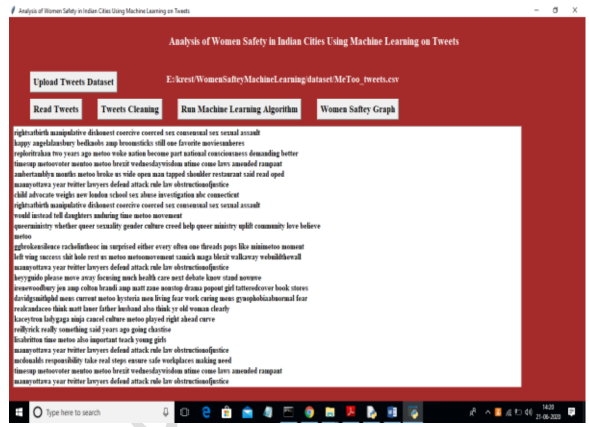

Fig.9: Tweet cleaning

In above screen we can see all special symbols and stop words remove from tweets and only clean words are there and now click on 'Run Machine Learning Algorithm' button to predict sentiments from tweets

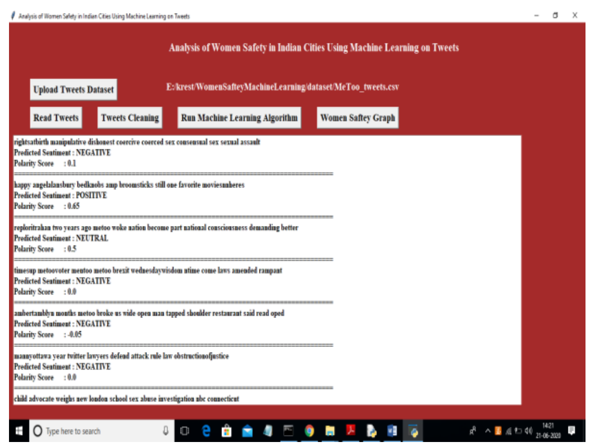

Fig.10: Machine learning algorithm

In above screen each tweet having tweet text and then displaying tweets sentiments with polarity score. Scroll down above text area to see all tweets. Now click on 'Women Saftey Graph' button to get below results and by seeing that result user can easily understand whether area is safe or not. If area is safe then more peoples will

Journal of Engineering Sciences
Vol 13, Issue 05, MAY/2022

ISSN NO:0377-9254

express either positive or neutral tweets and if not safe then more peoples will discuss negative tweets.

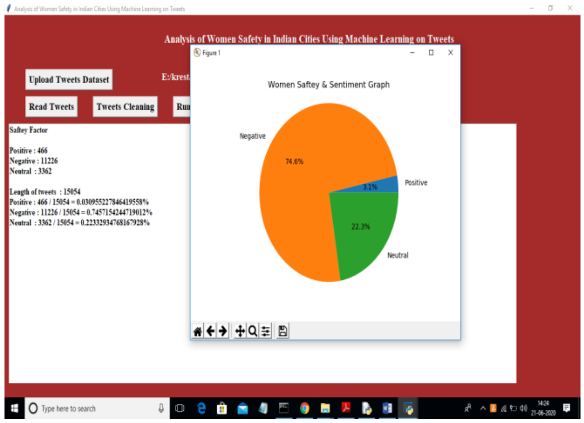

Fig.11: Women safety graph

In above screen 0.74 multiply by 100 will give 74% which means 74% peoples are talking negative and area is not safe and only 22 and 3% peoples are talking positive

6. CONCLUSION

Throughout the research paper we have discussed about various machine learning algorithms that can help us to organize and analyze the huge amount of Twitter data obtained including millions of tweets and text messages shared every day. These machine learning algorithms are very effective and useful when it comes to analyzing of large amount of dataincluding the SPC algorithm and linear algebraic Factor Model approaches which help to further categorize the data into meaningful groups. Support vector machines is yet another form of machine learning algorithm that is very popular in extracting Useful information from the Twitter and get an idea about the status of women safety in Indian cities.

7. FUTURE SCOPE

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] Agarwal, Apoorv, 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] Barbosa, Luciano, 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] Bermingham, Adam, 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] Gamon, Michael. "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] Kim, Soo-Min, and Eduard Hovy. "Determining the sentiment of opinions." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.

[6] Klein, Dan, 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] Charniak, Eugene, 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.