International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 5, May 2020 ||

Analysis of Women Safety in Indian Cities Using Python on Tweets

M.Anantha Lakshmi [1], P.Suchitra [2], N.Manoja [3], N.Kavya[4], S.Shagufta Maheen[5], S.Sadiya Samreen[6]

Assistant Professor, Department of Computer Science & Engineering, Ravindra College of Engineering for Women, Kurnool, India[1]

Students, Department of Computer Science & Engineering, Ravindra College of Engineering for Women, Kurnool, India [23456]

ABSTRACT: Social networking is a biggest resource to express people's opinions and sentiments towards different topics. Sentiment is used to determine the opinion of the client/writer. Sentiment analysis is a term that can predict by taking a piece of information and indicate whether it is a positive, negative and neutral sentiment. In this technical paper we show how to connect to the Twitter API and how we can get tweets regarding women. The algorithms we used are SVC algorithm and linear regression algorithm to classify and analyse the tweets. And we need to check the Geo location tax to know from which location the more percentage of postive and negative tweets are coming. In this we get the accurate and automatic sentiment analysis of collected tweets and we can take some measures wherever the highest negative tweets are observed.

I. INTRODUCTION

1.1 Overview of the Project

The goal of this project is to perform sentiment analysis on the data available in the twitter. Public opinions are mined from Twitter and then classified into sentiments, whether positive or negative. These results will let us know about the reviews and opinions of people. To achieve this goal, a module is created which can perform live sentimental analysis. In live sentimental analysis user can obtain the trend of any live trending topic depicted by two sentiment category (positive and negative) in live graphs.

1.2 Objective of the Project

Twitter is a small-scale blogging stage where clients generate 'tweets' that are communicated to their devotees or to another client. At 2016, Twitter has more than 313 million dynamic clients inside a given month, including 100 million clients daily . Client origins are widespread, with 77% situated outside of the US, producing more than 500 million tweets every day. The Twitter site positioned twelfth universally for activity in 2017 and reacted to more than 15 billion API calls every day. Twitter content likewise shows up in more than one million outsider sites. In accordance with this enormous development, Twitter has of late been the subject of much scrutiny, as Tweets frequently express client's sentiment on controversial issues. In the social media context, sentiment analysis and mining opinions are highly challenging tasks, and this is due to the enormous information generated by humans and machines.

1.3 Motivation of the Project

There is a feeling of insecurity among the working women. In Today's World the safety of women is in danger especially in India. The rate of crimes against women is not decreasing but in fact increasing at an alarming rate especially harassment, molestation, eve-teasing, rape, kidnapping and domestic violence. Many preventive measures have been taken by the government to stop these misbehaving activities but still has not affected the growing rate of these crimes and has remained unaffected. Women is getting kidnapped at every 44 minutes, raped at every 47 minutes, 17 dowry deaths every day[1].The fear of harassment against women is not only the condition at outside but it may also happen at homes, Women are not so physically fit as compared to men so in case of a need a helping hand would be a boon for them[2]. In this paper. We will show some queries on women safety and show the polarity of tweets. Our approach used classifiers to categorize sentiment into positive or negative and we applied a good feature extractor for enhancing accuracy. The classifiers which are built into a model to effectively classify are Naive Bayes (NB) and Support Vector Machine (SVM). We are implementing the project in google collab.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 5, May 2020 ||

II. RELATED WORK

The literature on the subjected area shows a variety of approaches, the investigator high lights briefly the significance of research in secondary education and summarizes the relevant studies that have been conducted in this area.

1. **L. A. Adamic and N. Glance[1]** has proposed "The political blogosphere and the 2004 U.S. election: Divided they blog" They classified a user into three categories: "Gender", "Age" and "Political Affiliation" with an application to Indian Twitter users. There approach automatically predicts these attributes by leveraging observable information such as the tweet behavior, linguistic content of the user's Twitter feed and the celebrities followed by the user.
2. **F. Al Zamal, W. Liu, and D. Ruths [2]** has proposed "Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors"which presents a general statistical methodology for the analysis of multivariate categorical data arising from observer reliability studies.
3. **J. An, M. Cha, K. P. Gummadi** [3] has proposed "Visualizing media bias through Twitter" to measure which candidates newspapers endorse in state and federal elections from 1940 to 2002. One sample focuses on the largest circulation newspapers in the United States from 1940 to 2002.
4. **S. Ansolabehere, R. Lessem, and J. M. Snyder** [4] has proposed "The orientation of newspaper endorsements in U.S. elections" to allow politicians and citizens increasingly engage in political conversations on social media outlets such as Twitter.

III. DESIGN, ISSUES AND IMPLEMENTATION OF THE MODEL

3.1. EXISTING MODEL:

To measure the computational perceptions of the customers we need to develop a program for sentiment analysis. In this existing system the tweets are analyzed with isolation forest algorithm. Isolation forest algorithm is an unsupervised learning algorithm for anomaly detection that works on the principle of isolating anomalies, instead of the most common techniques of profiling normal points. But it is done only to some extent.

3.2. PROPOSED SYSTEM:

In the proposed system we are using twitter API for analyzing the tweets using the SVM and linear regression algorithm.

3.3. MACHINE LEARNING ALGORITHMS:

3.3.1. Support vector machine algorithm:

SVM is a supervised machine learning algorithm which is used for classification or regression problems. It uses a the kernel trick technique to transform the data.Based on these transformations it finds an optimal boundary between the possible outputs.

3.3.2. Linear regression algorithm:

Linear regression is a statistical approach for modelling the relationship between a dependent variable with a given set of independent variables.

3.4. CLASSIFIERS :

Here we use one classifiers it is naive bayes classifier

3.4.1. Naive bayes classifier:

Naive Bayes classifiers can be extremely fast compared to more sophisticated methods. The seperation of the class conditional feature distributions should be done . That means that each distribution can be estimated independently as a one dimensional distribution.

3.5. MODULES :

3.5.1. Tweepy module:

Tweepy is open-sourced and hosted on GitHub and enables Python to communicate with the Twitter platform and use its API

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 5, May 2020 ||

3.5.2. Csv module:

CSV module is used to handle CSV files. That is to read/write data,we need to loop through rows of the CSV.we need to use the split method to get data from specified columns.

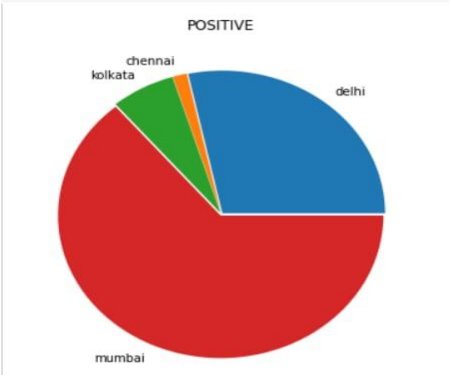
3.5.3. Pandas module:

Pandas is an open source library in Python. It runs on top of NumPy and it is popularly used for data science and data analytics.It is a low-level data structure that supports multi-dimensional arrays and a wide range of mathematical array operations.DataFrame is the key data structure in Pandas module.

IV. EXPERIMENTAL RESULTS

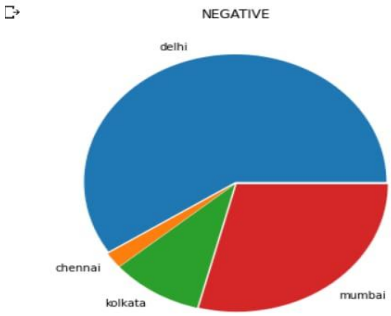
```
                                                  Delhi
0    Dear. CM kejriwal sr apke Delhi me meri mom, b...
1    RT @ameer_maawiya: @MzafarQasmi @alarafatengr ...
2    @MzafarQasmi @alarafatengr Delhi police chutiy...
3    Mother's Day hi chuan North Delhi awm lai a U ...
4    @SitamarhiDm @IPRD_Bihar @abhilashasharma @air...
..                                                 ...
189  @AHindinews @ANI Sir meri sister fsi hui h del...
190  @dm_shamli Mam namaskar mai. Mam lockdown hone...
191  @DM_Samastipur  sir namaskar  Delhi s hu tajpu...
192  @HemantSorenJMM Delhi me Bhai or Bihar me meri...
193  RT @nitin33K: @ZeeNewsCrime @DelhiPolice @DCPO...

[194 rows x 1 columns]
                                                 Mumbai
0    @GoswamiArnaw Kuch mind se jyada dimagdaro ki ...
1    @singhsahab1282 @majorgauravarya Brooo I am In...
2    RT @__MunnaJi: @untvin @Rita_Blogs Guru ji mum...
3    @untvin @Rita_Blogs Guru ji mumbai mein high p...
4    @sidhearts00009 @Truthsp25271613 @sidhearrts @...
..                                                 ...
163  @virupandit6 @PiyushGoyal @narendramodi @AmitS...
164  #SITANAVAMI  #Mother #Blessings ❤️ #Love #JaiSr...
165              @a_busy_woman hyderabadi and bombay 😳
166  Sir Mai Gorakhpur k gorakhnat Ka rahne wla hu ...
167      @BeYouDoYouSeeU Mumbai redddddddd zoneeeee 💥

[168 rows x 1 columns]
                                                Chennai
0    RT @ursyathi: Chennai poyi vachi malli start c...
1    Chennai poyi vachi malli start chesaru maa jil...
2    @Savitab84127227 @lipakshi_kapoor @kaamuk_salo...
3    Aditi Rao Hydari in Bennch at Chennai Event ht...
4    @babbarishaam Chennai MTC la oruthan semaya ir...
5    @chennaicorp Sir..... Nan apply panirukan... E...
6              @gowriakk Adai aunty Indian summa irra
7    Maa pakka village lo okadiki chennai koyambedu...
```

In these pie charts we can separately observe the positive and negative tweets so that we can easily aware of the locations from where the highest percentage of positive or negative tweets are coming.

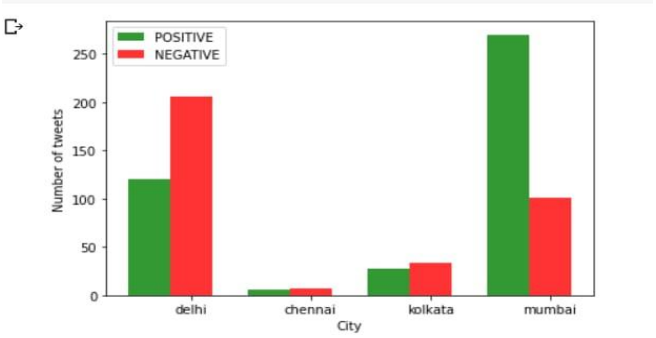


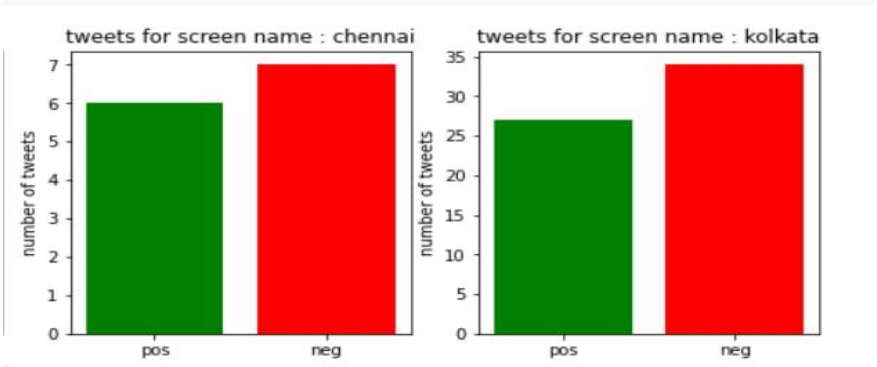
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 5, May 2020 ||



In this bar chart we can observe how tweets percentage is taking place in different locations.





V. CONCLUSION

We have discussed about various machine learning algorithms that can help us to organize and analyse 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 analysing of large amount of data including the NAÏVE BAYES CLASSIFIER and linear SVC model. 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.

REFERENCES

[1] Albert Biffet and Eibe Frank. Sentiment Knowledge Discovery in Twitter Streaming Data. Discovery Science, Lecture Notes in Computer Science, 2010, Volume 6332/2010, 1-15, DOI: 10.1007/978-3-642-16184-1_1

[2] Alec Go, Richa Bhayani and Lei Huang. Twitter Sentiment Classification using Distant Supervision. Project Technical Report, Stanford University, 2009.

[3] Alexander Pak and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proceedings of international conference on Language Resources and Evaluation (LREC), 2010.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 5, May 2020 ||

[4] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner and Isabell M. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Proceedings of AAAI Conference on Weblogs and Social Media (ICWSM), 2010. Project Thesis Report 51

[5] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002.

[6] Chenhao Tan, Lilian Lee, Jie Tang, Long Jiang, Ming Zhou and Ping Li. User Level Sentiment Analysis Incorporating Social Networks. In Proceedings of ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD), 2011.

[7] Efthymios Kouloumpis, Theresa Wilson and Johanna Moore. Twitter Sentiment Analysis: The Good the Bad and the OMG! In Proceedings of AAAI Conference on Weblogs and Social Media (ICWSM), 2011.