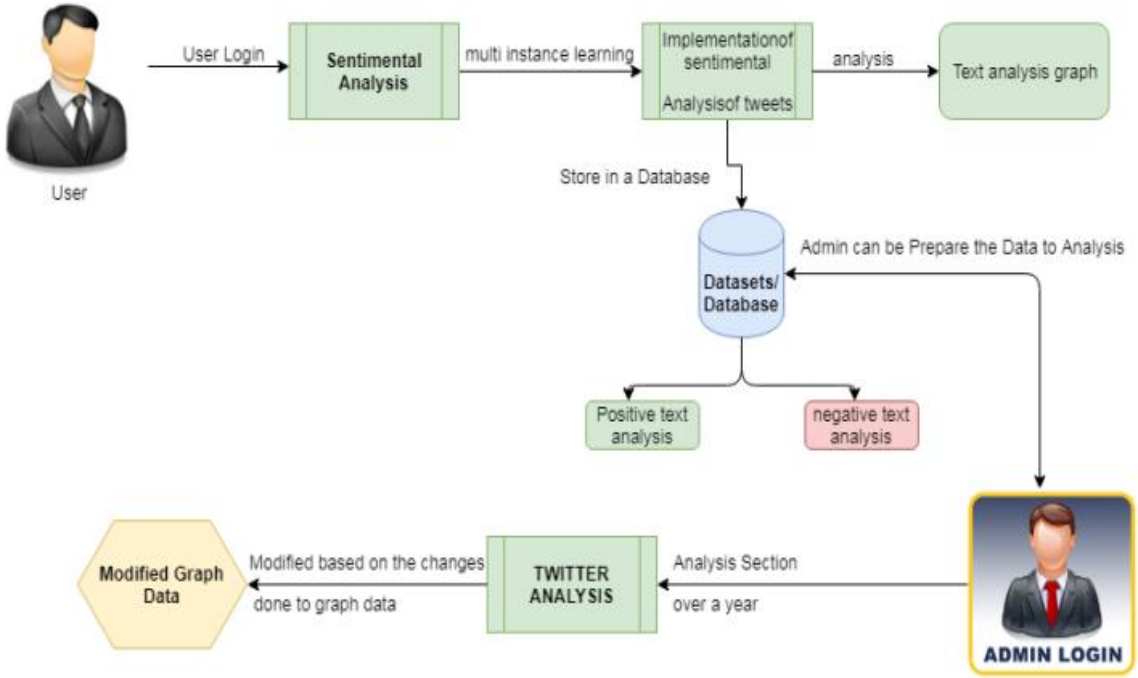
A MACHINE LEARNING BASED APPROACH FOR THE ANALYSIS OF TWEETS

[1]*B V Pranay Kumar,*[2]*J Purna Prakash,*[3]*CH Saritha,*[4]*A Navya Sri*
[1,2,3]*Assistant Professor,*[4]*Student*
Department of CSE
Christu Jyothi Institute of Technology &Science,Colombo Nagar, Telangana

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?

I. ARCHITECTURE



II. EXISTING SYSTEM
People often express their views freely on social media about what they feel about the Indian society and the politicians that claim that Indian cities are safe for women. On social media websites people can freely Express their view point and women can share their experiences where they have faced abuse harassment or where we would have fight back against the abuse

harassment that was imposed on them . The tweets about safety of women and stories of standing up against abuse harassment further motivates other women data on the same social media website or application like Twitter. Other women share these messages and tweets which further motivates other 5 men or 10 women to stand up and raise a voice against people who have made Indian cities and unsafe place for the women. In the recent years a large number of people have been attracted towards social media platforms like Facebook, . It is a common practice to extract the information from the data that is available on social networking through procedures of data extraction, data analysis and data interpretation methods. The accuracy of the Twitter analysis and prediction can be obtained by the use of behavioral analysis on the basis of social networks.

DISADVANTAGES:

1. Twitter and Instagram point and most of the people are using it to express their emotions and also their opinions about what they think about the Indian cities and Indian society.
2. There are several method of sentiment that can be categorized like machine learning hybrid and lexicon-based learning.
3. Also there are another categorization Janta presented with categories of statistical, knowledge-based and age wise differentiation approaches

III. PROPOSED SYSTEM

Women have the right to the city which means that they can gofreely 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 isthe main reason of harassment of girls. There are instances when the harassment of girls was done by their neighbourswhile 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 abusely harassed by one of their own neighbor or any other unknown person. Safest cities approach women safety from a perspective ofwomen rights to the affect the city without fear of violence or abuse 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.

ADVANTAGES:

1. Analysis of twitter texts collection also includes the name ofpeople and name of women who stand up against abuseharassment and unethical behaviour of men in Indian citieswhich make them uncomfortable to walk freely.
2. The datasetthat was obtained through Twitter about the status of womensafety in Indian society

IV. LITERATURE SURVEY

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.

Learning word vectors for slant examination

Authors:Maas, Andrew L., etal.Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Relationship for Computational Linguistics, 2011

Unaided vector-based ways to deal with semantics can show rich lexical implications, however they generally neglect to catch notion data that is vital to many word implications and significant for a wide scope of NLP assignments.

A model that uses a blend of solo and managed methods to learn word vectors catching semantic term–record data just as rich feeling content. The proposed model can use both consistent and multi-dimensional opinion data just as non-assumption explanations.

Here the model is instantiated so as to use the record level estimation extremity comments present in numerous online archives (for example star evaluations). A model utilizing little, generally utilized conclusion and subjectivity corpora and discover it out-plays out a few recently presented strategies for feeling characterization.

A huge dataset of film audits to fill in as an increasingly hearty benchmark for work here is used.[1]

Opinion mining and supposition investigation

Authors: Pang, Bo, and Lillian Lee. Establishments and patterns in data recovery 2, no. 1-2 (2008): 1-135.

In this paper data gathering conduct has consistently been to discover what other individuals think. With the developing accessibility and fame of assessment rich assets, for example, online audit destinations and individual web journals, new chances and difficulties emerge as individuals presently can, and do, effectively use data innovations to search out and comprehend the assessments of others.

The unexpected emission of action in the region of supposition mining and conclusion examination, which manages the computational treatment of assessment, feeling, and subjectivity in content, has in this manner happened in any event to a limited extent as an immediate reaction to the flood of enthusiasm for new frameworks that manage sentiments as a five star object.

The overview covers strategies and methodologies that guarantee to legitimately empower sentiment situated data looking for frameworks. For the most partspotlight must be kept up on strategies that try to address the new difficulties raised byassessment mindful applications, when contrasted with those that are as of now present in increasingly conventional reality based examination.

Here a material on outline is incorporated into request to evaluative content and on more extensive issues with respect to protection, control, and monetary effect that the advancement of sentiment arranged data access administrations offers ascend to.

To encourage future work, an exchange of accessible assets, benchmark datasets, and assessment battles is likewise provided.[2]

wistful instruction: Sentiment examination utilizing subjectivity rundown dependent on least cuts

Authors: Pang, Bo, and Lillian Lee. In Proceedings of the 42nd yearly gathering on Association for Computational Linguistics, p. 271.Relationship for Computational Linguistics, 2004.

Assessment investigation tries to recognize the viewpoint(s) fundamental a content range; a model application is arranging a motion picture audit as "approval" or "disapproval".

To decide this feeling extremity, we propose a novel AI strategy that applies content order methods to simply the emotional parts of the record.

Separating these bits can be executed utilizing productive procedures for discovering least cuts in diagrams; this extraordinarily encourages fuse of cross-sentence relevant constraints.[3]

Recursive profound models for semantic compositionally over an assumption Treebank

Authors: Socher, Richard, et al. Procedures of the meeting on exact strategies in regular language preparing (EMNLP). Vol. 1631. 2013

The paper clarifies about Semantic word spaces have been extremely helpful yet can't express the significance of longer states in a principled manner. Further progress towards understanding compositionality in undertakings, for example, opinionlocation requires more extravagant directed preparing and assessment assets and all the more dominant models of organization. To cure this, a Sentiment Treebank is presented.

It incorporates fine grained assessment names for 215,154 expressions in the parse trees of 11,855 sentences and introduces new difficulties for feeling compositionality.

To address them, a Recursive Neural Tensor Network. At the point when prepared on the new Treebank, this model outflanks every single past technique on a few measurements. It pushes the cutting edge in single sentence positive/negative arrangement from 80% up to 85.4%.

The precision of anticipating fine-grained notion marks for all expressions arrives at 80.7%, an improvement of 9.7% over sack of highlights baselines.

Ultimately, it is the main model that can precisely catch the impacts of nullification and its extension at different tree levels for both positive and negative phrases.[4]

Distributed portrayals of words and states and their compositionality

Authors: Mikolov, Tomas, et al. Advances in Neural Information Processing Systems. 2013.

The paper clarifies that the Skip-gram model is an effective strategy for adapting high caliber appropriated vector portrayals that catch an enormous number of exact syntactic and semantic word connections.

A few augmentations that improve both the nature of the vectors and the preparation speed are introduced.

By sub-sampling of the successive words we get critical speedup and furthermore adapt increasingly standard word portrayals. Here a straightforward option in contrast to the various leveled soft-max called negative inspecting is portrayed**.**

An inborn restriction of word portrayals is their impassion to word request and their powerlessness to speak to colloquial expressions.

For instance, the implications of "Canada" and "Air" can't be effectively consolidated to get "Air Canada". Spurred by this model, we present a basic strategy for discovering phrases in content, and demonstrate that adapting great vector portrayals for many expressions is possible.[5]

Regularization and variable choice by means of the flexible net

Authors: Zou, Hui, and Trevor Hastie. Diary of the Royal Statistical Society: Series B (Statistical Methodology) 67, no. 2 (2005): 301-320.

The flexible net, another regularization and variable choice technique has been proposed. True information and a reproduction study demonstrate that the versatile net frequently outflanks the rope, while getting a charge out of a comparative sparsity of portrayal.

What's more, the flexible net supports a gathering impact, where firmly corresponded indicators will in general be in (out) the model together.

The versatile net is especially valuable when the quantity of indicators (p) is a lot greater than the quantity of perceptions (n). Paradoxically, the rope is certifiably not an exceptionally acceptable variable choice technique in the p n case.

A productive calculation called LARS-EN is proposed for figuring versatile net regularization ways proficiently, much like the LARS calculation accomplishes for the lasso.[6]

Parsing regular scenes and normal language with recursive neural systems

Authors: Socher, Richard, et al. Procedures of the 28th global meeting on AI (ICML-11). 2011.

Recursive structure is usually found in the contributions of various modalities, for example, common scene pictures or normal language sentences. Finding this recursive structure encourages us to not just distinguish the units that a picture or sentence contains yet in addition how they communicate to frame an entirety.

A maximum edge structure forecast engineering dependent on recursive neural systems that can effectively recoup such structure both in complex scene pictures just as sentences was presented.

A similar calculation can be utilized both to give an aggressive syntactic parser to characteristic language sentences from the Penn Treebank and to beat elective methodologies for semantic scene division, comment and order.

For division and comment calculation gets another degree of condition of-the art execution on the Stanford foundation dataset (78.1%). The highlights from the picture parse tree beat Gist descriptors for scene order by 4%.[8]

Multivariate relapse and AI with entireties of distinguishable capacities

Authors: Beylkin, Gregory, JochenGarcke, and Martin J. Mohlenkamp. SIAM Journal on Scientific Computing 31, no. 3 (2009): 1840-1857.

Here a calculation for learning (or assessing) a component of numerous variables from dissipated information is introduced. The capacity is approximated by a total of distinguishable capacities, following the paradigm of isolated portrayals.

The focal fitting calculation is straight in both the number of data focuses and the quantity of factors, and hence is reasonable for huge informational indexes in high measurements. A numerical proof for the utility of these portrayals is introduced.

Specifically, we demonstrate that this strategy beats different strategies on a few benchmark information sets.[9]

Semantic compositionality through recursive network vector spaces

Authors: Socher, Richard, etal. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.Relationship for Computational Linguistics, 2012. APA

It states about Single-word vector space models have been fruitful at learning lexical data. Be that as it may, they can't catch the compositional significance of longer expresses, keeping them from a more profound comprehension of language.

A recursive neural system (RNN) model that learns compositional vector portrayals for expressions and sentences of self-assertive syntactic sort and length.

The model doles out a vector and a framework to each hub in a parse tree: the vector catches the intrinsic significance of the constituent, while the lattice catches how it changes the importance of neighboring words or expressions.

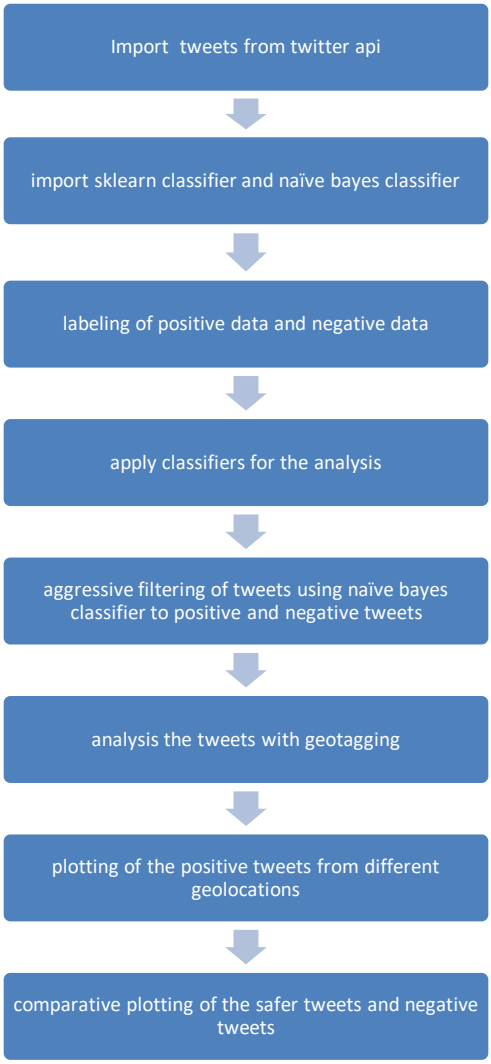
The network vector RNN can become familiar with the significance of administrators in propositional rationale and common language.

The model gets cutting edge execution on three distinct trials: anticipating fine-grained assumption circulations of modifier descriptive word sets; arranging slant names of motion picture surveys and grouping semantic connections, for example, cause-impact or point message between things utilizing the syntactic way between them.[10]

V. WORKING

In this project we are using tweepy module pandas and csv module. The data set is loaded into googlecollab

We are using nltk module to import sklearn classifier and naïve bayes classifier the procedure of processing is done as follows:



VI. CONCLUSION

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 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.

VII. RESULTS

Figure 1 LOGIN SCREEN

Figure 2 ADMIN LOGIN ANALYSIS

Figure 3 PIE GRAPH

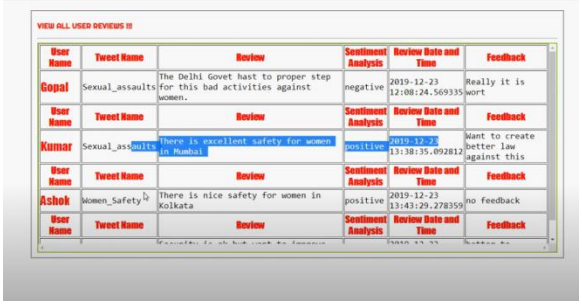


Figure 4 USER REVIEWS

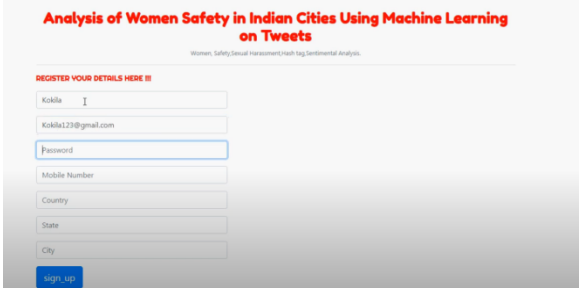


Figure 5 USER REGISTTRATION

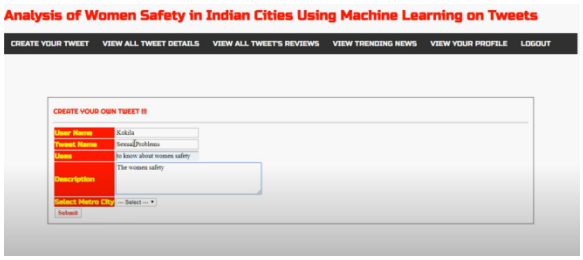


Figure 6 USER TWEET

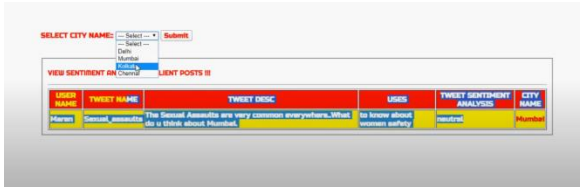


Figure 7 SENTIMENT ANALYSIS ON TWEETS ACCORDING TO GEOLOCATION

VIII. REFERENCES

[1]Agarwal, Apoorv, FadiBiadsy, 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 JunlanFeng. "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.