Sentiment Classification on Women Safety across Indian Cities Based on Twitter Data using NLP and Machine Learning

Ranjitha[1], Pradeep Nayak[2], Vedanth M[3], Mahantesh G[4], Namitha D[5]
[1,2,3,4,5]*Alvas Institute of Engineering and Technology, Moodbidri, Karnataka, India*

Abstract - **These days women are experiencing lots of violence such as persecution 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 Primarily focus on the role of social media which can be utilized to advance the safety of women in India, given the unique reference to the cooperation of numerous virtual entertainment sites or applications like Twitter, Facebook and Instagram platforms. The proposed work focuses on developing a ML model that can determine the least safe states in India based on the sentiment classification of twitter data. Tweet from the Twitter application contains the text messages, audio data, video data, images, smiley expressions and hashtags. This tweet content can be used for sentiment classification in NLP. The proposed work also implies the comparative study of four ML algorithms used in classification and the best accurate algorithm is used for analysis.**

Applications which include hashtags such as metoo has been considered for the study. Machine learning algorithms including SVM, Random Forest, NB and Logistic Regression are analyzed.

Index Terms - **Twitter, Sentiment analysis (SA), Machine Learning (ML), Random Forest, Naive Bayes (NB), Support Vector Machine (SVM), Women, Safety, Hash tag.**

I.INTRODUCTION

On the twitter, clients will share their contemplations and viewpoint in the tweets area. Twitter with such a monstrous crowd has polarized clients to radiate their viewpoints and critical about each current issue and subject of web, subsequently twitter is an enlightening source. People express their sentiments about society, politics, women, etc. via the text messages, emoticons and hash-tags through such platforms.

The sentiment analysis procedure is employed for the extraction of sentiment behind every tweet. Hence the results that are obtained from the sentiment analysis is used to categorize the tweets. Social platforms like Twitter, Facebook, Instagram are used by different individuals for their opinions called sentiment. Social media platforms have a lot of information. Women are also more likely to use social media to share incidents or opinions. Gazing at women and passing comments can be particular sorts of viciousness and disturbances and these practices, which are inadmissible. Twitter sentiment analysis frameworks permit us to sort large arrangements of tweets and determines every assertion.

Safe cities approach implicit women's protection in perspective of women's freedom, making them to impact the city unafraid of violence or sexual harassment. Rather than imposing the limitation that society normally imposes on women, it is the obligation of society to recognize the need for women safety and to understand that women and girls, similar to men, reserve a privilege to be protected in the city.

Analysis of twitter texts assortment likewise incorporates names of individuals and names of women who take a stand in opposition to lewd behavior and exploitative way of behaving of men in Indian cities which makes it difficult for them to walk uninhibitedly. The data set about the status of women's safety in Indian society that was obtained through Twitter was processed using machine learning algorithms in order to smooth the data by removing zero values and foster an analysis strategy by utilizing Porter's and Laplace theory to remove retweet and repetitive information from the data set that was obtained so a clear and unique perspective on women safety in Indian cities could be acquired.

II. SENTIMENT ANALYSIS

Sentiment analysis can be implemented on extracted data from social media and analyzed using Natural Language Processing (NLP), the TextBlob library of NLP helps to calculate polarity and sentiment of extracted text unit. This analysis will return a sentiment of positive, negative or neutral. The Sentiment analysis is used in brand monitoring, analyzing customer satisfaction to improve the branding, market research analysis, etc. Sentiment analysis can be many types such as text summarization, recommendations, social sentiment analysis etc. However, analysis of social media streams is always restricted and basic analysis can only be done using the extracted texts.

III. EXISTING SYSTEM

Existing work in view of a bunch of procedures used Naive Bayes and SVM achieved accuracy of 80% which is trained with emoticon. They analyzed sentiment classification using distant supervision model. For analysing the enormous measure of information, the SPC algorithm is used along with linear algebraic factor model.

Lexicon-based methodologies depend with respect to a predefined list of words, every one of which is connected to a specific inclination. The lexical methodologies vary contingent upon the climate in which they were grown, yet they all involve assessing report direction in view of the semantic direction of letters or expressions in the records.

IV. PROPOSED SYSTEM

Proposed work presents the women safety analysis on twitter dataset and verification of outcomes acquired from every analyzer with machine learning algorithm. The examination depends on the correlation of various sentiment analyzers and the outcomes are then approved utilizing various classifiers.

Here we are using the different ML algorithms including SVM, Random Forest, NB and Logistic Regression and use polarity values to identify the sentiment. Predict the sentiment using machine leaning techniques.

V. SYSTEM ARCHITECTURE

The below figure represents the system architecture of the overall proposed system. In this architecture, we represented all modules of implementation. The information collection method has the choice of value information for sentiment analysis. In this we have taken twitter dataset from extricated from twitter.com. In machine learning, there is highly needed for pre-processing dataset, the aim of preprocessing is to modify the unprocessed information into a structure that fits machine learning. Following data analyst has preprocessed the gathered information and parted it into train and test can continue with model training.
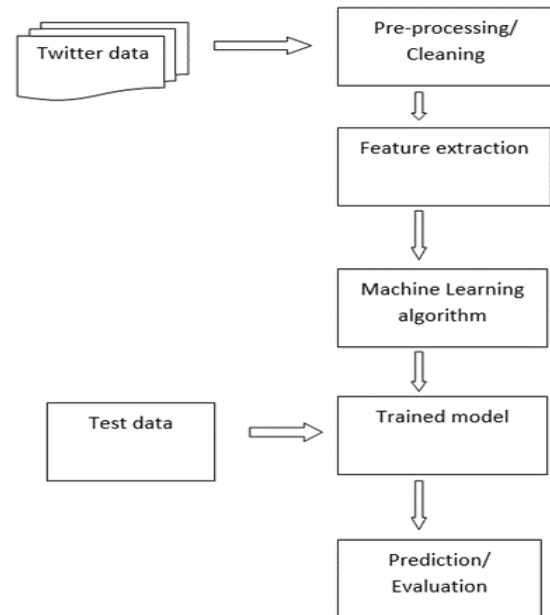
Fig-1: System Architecture

VI. MODULES

The accompanying modules are implemented in our implementation.

A. Data set preparation

The training dataset of around of twitter data based on Indian cities location for the hashtag #metoo is considered for study. The dataset containing id, label and tweet as the attributes.

B. Dataset pre-process

Data pre-processing of different stages are handled before using for machine learning. When the information has been extricated as datasets from the Twitter source, it should be conveyed to the classifier. Prior to analysis, the classifier cleans up the dataset by eliminating repetitive information, for example, stop

words and emojis to guarantee that non-literary content is recognized and eliminated before the analysis.

C. Feature Extraction

Feature extraction is the process handled before training the dataset, which facilitates subsequent learning and generalization steps. Here we used count vectorization for feature extraction. The tokenized tweets from above module given as input for feature extraction methods.

D. Machine learning and Comparison

Sentiment analysis may extract insightful information from social media platforms by identifying emotions or views from a large amount of unstructured data. Sentiment analysis makes use of negative and positive polarity. Every tweet is given a score ranging from 0 to 1, depending on the terms used, to determine its polarity. We apply SVM, Logistic Regression, Naive Bayes, and Random Forest. Values for Accuracy and Error are generated.

VII. IMPLEMENTATION OF SENTIMENTAL ANALYSIS OF TWEETS

Scikit-Learn, Pandas, Matplotlib, and other necessary libraries are used to implement the proposed work in Python 3.6.4. From twitter.com, we downloaded a dataset. Separate train set and test set data are included in the downloaded data. Test.csv is regarded as the test set, and train.csv as the train set. On the train set, ML techniques like SVM, Logistic Regression, Naive Bayes, and Random Forest are applied.

Implementation Steps
The proposed application ought to recognize the sentiment of the tweet text. Feature extraction models are used which includes N-count, N-gram and TF/IDF. We used classification models such as SVM, Random Forest, NB and Logistic Regression to classify the tweets based on their sentiments.

1 Extricate the feature by applying count vectorization
2 Apply machine leaning methods
3 Sets for training and testing are separated.
4 Use the ML algorithms which are listed in the above section on test set to classify the problem.
5 Predict the different types of sentiments.

6 Compare the accuracy of the machine learning algorithms on every feature extraction model.

VIII.RESULT

In our implementation twitter dataset is considered for study. In the proposed system, we applied three feature extraction models such as N-count, N-gram and TF/IDF, then applied machine learning algorithms including SVM, Random Forest, NB and Logistic Regression.
The training dataset from twitter is collected with the hash tag
metoo of various cities in India. Experimental is conducted using Feature extraction and ML techniques in order to predict the least safe state in India. The result shows that Random Forest is achieving high accuracy. The outcome of the analysis can be visualized by creating different types of graphs. To measure the least safe state the Bar graphs are used. The below figure depicts the Accuracy using n_count Vectorizer model. The accuracy of the algorithms is measured by N-count vectorizer. Here Naïve bayes shows the less accuracy among other four algorithms and the remaining three gives the accuracy value of 89% for tweet analysis. The below diagram depicts the Accuracy using n_gram Vectorizer model
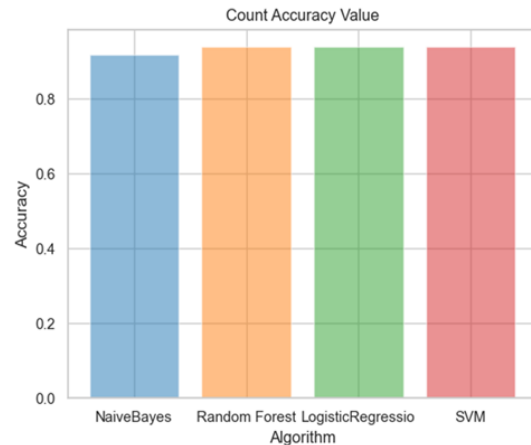

Fig-2: Evaluation metrics of N-count vectorizer
The below diagram depicts the Accuracy using n_gram Vectorizer model. The N-gram vectorizer shows highest accuracy value for Logistic Regression of 88%. This algorithm classifies the data point correctly over others. The below figure shows the Accuracy using TF-IDF Vectorizer model

Fig-3: Evaluation metrics of N-gram vectorizer

The below figure shows the Accuracy using TF-IDF Vectorizer model. The TF-IDF vectorizer shows highest accuracy value for Random Forest algorithm of 89%. Which is the highest accuracy achieved by then we choose Random Forest for the model development which can classify the problem correctly. The best accurate algorithm gives the best model problem prediction results.

Fig-4: Evaluation metrics of TF-IDF vectorizer

Assuming we run the program in different times we might get unique results at each occurrence with a little difference, in light of the tweets we bring. We ran the program for multiple times and these results are the normal of the continuous results.

This means a lot to make reference to that relies upon the information of the analysis we might get various outcomes as individuals' viewpoint might change relying upon conditions.

Fig-5: Total sentiment across different states for women safety analysis

By the above study that we have done, it tends to be obviously expressed that Mizoram is the safest city while Uttar Pradesh is the perilous city.

IX.CONCLUSION

Machine learning algorithm has been discussed throughout the project. For the twitter data that includes a huge number of tweets and messages consistently, Machine Learning algorithms helps to organize and perform analysis. SVM, Random Forest, NB and Logistic Regression 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.

X.FUTURE ENHANCEMENT

For the future enhancement, these machine learning algorithms can be reached out to be utilized on various virtual entertainment platforms, for example, Facebook and Instagram additionally 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 tweets to provide more safety.

REFERENCE

[1] Deepak Kumar1, Shivani Aggarwal2 "Analysis of Women Safety in Indian Cities Using Machine Learning on Tweets", (IEEE), 2019.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.