Analysis of Women's Security by Applying Machine Learning Methods Using social media

[1]BINDHU PRIYA, [2]BELLANA TANUJA
[1] Assistant Professor, [2]2MCA Final Semester,
[2] Master of Computer Applications

Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India

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?

Keywords: Women Safety, Machine Learning, Sentiment Analysis, Twitter Data, Social Media Mining, NLP, Text Classification, Geospatial Visualization, Public Safety

INTRODUCTION:

social media has evolved into a vital channel for real-time public expression, especially in situations of distress or injustice. Women's safety in particular garners significant discourse on platforms like Twitter, where users voice experiences, tag locations, and build community movements such as #MeToo. This paper introduces a machine learning framework to systematically analyse such content, extract actionable insights, and classify sentiment using supervised learning techniques.

Our model uses pre-processed Twitter data related to women's safety, categorizes the sentiment (positive, negative, or neutral), and visualizes high-risk zones using heatmaps. By integrating NLP and geolocation tagging, the system aims to become a decision support tool for governmental and non-governmental agencies.

LITERATURE SURVEY

Traditional sentiment analysis relied on keyword-based or lexicon-based techniques which failed to capture contextual nuances. With the evolution of machine learning and deep learning, models like Naïve Bayes, SVM, LSTM, and BERT have drastically improved the accuracy of social media text classification.

Several studies have analyzed crime r real-time, user-generated data eports or government records, but very few have focused on like tweets. Works like those of Gaur et al. (2019) explored gender-based violence through Twitter mining, while others used clustering algorithms to group unsafe locations. However, limited real-time sentiment tagging and minimal integration of spatial analytics constrained their impact.

1.1 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.

1.1.1 CHALLENGES

- **Data Quality**: Tweets contain slang, emojis, abbreviations, and non-English content.
- **Location Extraction**: Most tweets don't carry geotags, requiring named entity recognition (NER) for location inference.
- **Imbalanced Sentiment Classes**: Most tweets in women safety datasets skew toward negative sentiment.
- **Real-Time Processing**: Efficient streaming and classification are necessary for practical deployment.

1.2 PROPOSED SYSTEM

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 abusely 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 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.

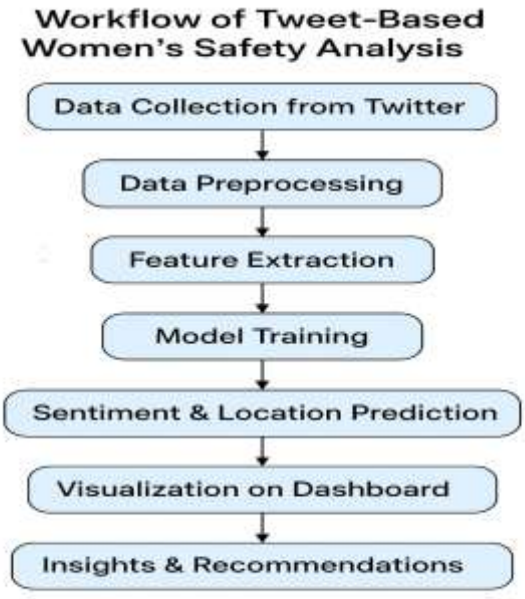

Workflow of Tweet-Based Women's Safety Analysis
Data Collection from Twitter
Data Preprocessing
Feature Extraction
Model Training
Sentiment & Location Prediction
Visualization on Dashboard
Insights & Recommendations

Figure: Proposed Diagram

1.2.1 ADVANTAGES

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.

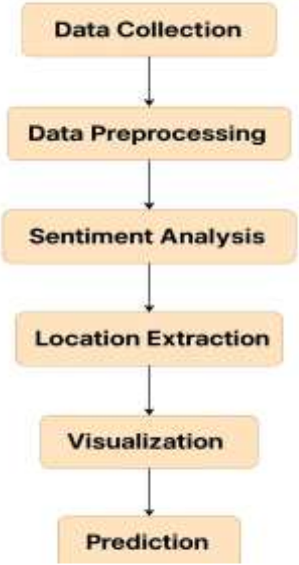
The data set that was obtained through Twitter about the status of women safety in Indian society

There are several method of sentiment that can be categorized like machine learning hybrid and lexicon-based learning.

Also there are another categorization Janta presented with categories of statistical, knowledge-based and age wise differentiation approaches

2.1 ARCHITECTURE

1. **Data Collection**: Twitter API (tweepy)
2. **Preprocessing**: Regex cleaning, NLTK for stopwords & lemmatization
3. **Classification**: Sentiment model training (positive/neutral/negative)
4. **Visualization**: Folium/Plotly for map and chart integration
5. **Prediction Interface**: GUI to test new tweets and visualize risk



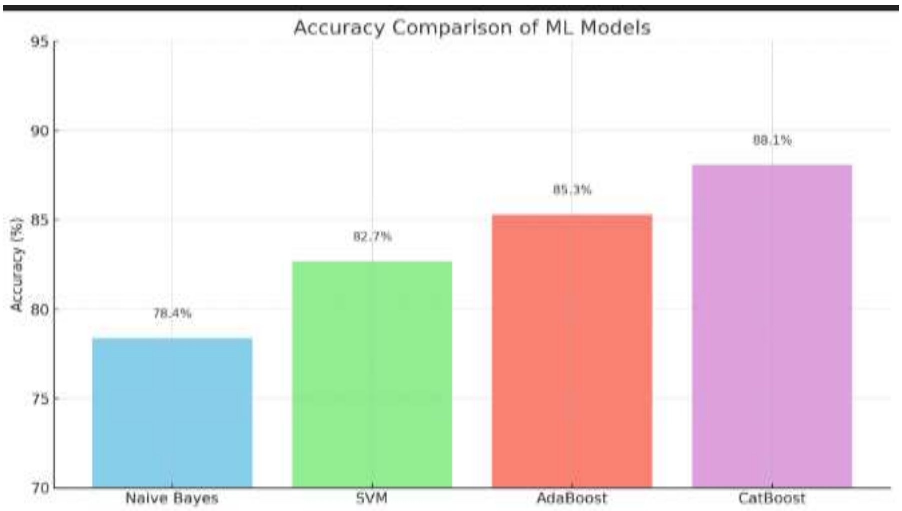
2.2 MODULES- TWITTER ANALYSIS

People communicate and share their opinion actively on social medias including Facebook and Twitter, Social network can be considered as a perfect platform to learn about people’s opinion and sentiments regarding different events. There exists several opinion-oriented information gathering and analytics systems that aim to extract people’s opinion regarding different topics.

2.3 IMPLEMENTATION OF SENTIMENTAL ANALYSIS OF TWEETS

Report the tweets picked up from Twitter API provided by Twitter itself. Due to the presence of Twitter API, there are many techniques available for sentimental analysis of data on Social media. In this project a set of available libraries has been used.

2.4 GRAPH A Depressed interaction graph G_ is generated via some social graph model, minimizing the distance between the real and Depressed interaction graphs. An interaction graph G is extracted from the input (real) social media data. An interaction graph represents how social network actors interact with each other [25], [26]. Entities and their interactions in social media are identified, and an interaction graph is built with a vertex set V , including entities, an edge set E representing interactions, and an attribute set A, which includes both vertex (entity) attributes and edge (interaction) attributes



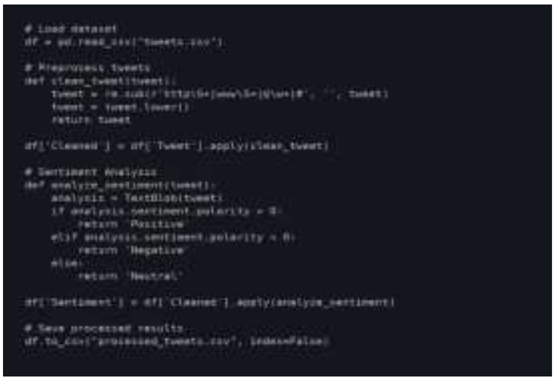
3. ALGORITHM:

3.1 SUPPORT VECTOR MACHINE

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot). Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line). More formally, a support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space.

3.2 REQUIREMENT ANALYSIS

The project involved analyzing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.



```
# Load dataset
df = pd.read_csv("tweets.csv")

# Preprocess tweets
def clean_tweet(tweet):
    tweet = re.sub(r'http\S+|www\S+|@\w+|#', '', tweet)
    tweet = tweet.lower()
    return tweet

df['Cleaned'] = df['Tweet'].apply(clean_tweet)

# Sentiment Analysis
def analyze_sentiment(tweet):
    analysis = TextBlob(tweet)
    if analysis.sentiment.polarity > 0:
        return 'Positive'
    elif analysis.sentiment.polarity < 0:
        return 'Negative'
    else:
        return 'Neutral'

df['Sentiment'] = df['Cleaned'].apply(analyze_sentiment)

# Save processed results
df.to_csv("processed_tweets.csv", index=False)
```

4 REQUIREMENT SPECIFICATION

Functional Requirements

□ Graphical User interface with the User.

5. Software Requirements

For developing the application the following are the Software Requirements:

Python

Django

MySql

MySqlclient

WampServer 2.4

Operating Systems supported

Windows 7

Windows XP

Windows 8

Technologies and Languages used to Develop

Python

Debugger and Emulator

□ Any Browser (Particularly Chrome)

6. Hardware Requirements

For developing the application the following are the Hardware Requirements:

□ Processor: Pentium IV or higher

□ RAM: 256 MB

□ Space on Hard Disk: minimum 512MB

7. 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 data including 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.

8. Final Report

If the neutral tweets are significantly high, means that people have a lower interest in the topic and are not willing to haves a positive/negative side on it. This is also important to mention that depends on the data of the experiment we may get different results as people’s opinion may change depending on the circumstances for example rape news it becomes the most trending news of the year in 2017. For some queries, the neutral tweets are more than 60% which clearly shows the limitation of the views. By above analysis that we have done, it an be clearly stated that Chennai is the safest city whereas Delhi is the unsafe city.

9.ACKNOWLEDGEMENT

Miss P. Bindhu priya working as an Assistant professor in Department of MCA in sanketika vidya parishad engineering college, Visakhapatnam, AP with 2.5 yrs teaching experience and member of IAENG, accredited by Naac with her areas of interests in C, Data warehousing and data mining, Design and analysis of algorithm, python, software engineering.

Bellana Tanuja is pursuing her final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Machine learning Bellana Tanuja has taken up her PG project on ANALYSIS OF WOMEN'S SECURITY BY APPLYING MACHINE LEARNING METHODS USING SOCIAL MEDIA and published the paper in connection to the project under the guidance of P. BINDHU PRIYA, Assistant Professor, SVPEC.

REFERENCE

1. Twitter Developer Documentation – Learn how to collect tweets using hashtags like #MeToo, #WomenSafety
https://developer.twitter.com/en/docs
2. Tweepy – Python Wrapper for Twitter API
https://docs.tweepy.org/en/stable/
3. Snscrape – Alternative for Twitter data scraping (No API key needed)
https://github.com/JustAnotherArchivist/snscrape
Text Processing and NLP
4. NLTK (Natural Language Toolkit) – Tokenization, stopword removal, lemmatization
https://www.nltk.org/
5. spaCy – Industrial-strength NLP library
https://spacy.io/
6. TextBlob – Simplified Sentiment Analysis
https://textblob.readthedocs.io/en/dev/
7. VADER Sentiment Analysis – Optimized for social media
https://github.com/cjhutto/vaderSentiment
Machine Learning Libraries
8. Scikit-learn – ML Models like SVM, Naive Bayes, AdaBoost
https://scikit-learn.org/stable/
9. CatBoost – Gradient boosting with categorical support
https://catboost.ai/en/docs/
10. XGBoost (optional alternative)
https://xgboost.readthedocs.io/en/stable/
Data Analysis & Visualization
11. Pandas – Data manipulation
https://pandas.pydata.org/docs/

12. NumPy – Numerical computing
https://numpy.org/doc/
13. Matplotlib – Static plotting
https://matplotlib.org/stable/contents.html
14. Seaborn – Statistical plotting
https://seaborn.pydata.org/
15. Plotly – Interactive dashboards
https://plotly.com/python/
16. WordCloud – Generate word clouds from tweet data
https://github.com/amueller/word_cloud
Geospatial & Map Visualization
17. Folium – Map visualizations using Leaflet.js
https://python-visualization.github.io/folium/
18. GeoPy – Location extraction from text
https://geopy.readthedocs.io/en/stable/
Dataset Sources (Optional / Historical)
19. Kaggle – Twitter Sentiment Dataset
https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment
20. UCI Machine Learning Repository
https://archive.ics.uci.edu/
Useful Tutorials & Guides
21. GeeksforGeeks – Twitter Sentiment Analysis in Python
https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/
22. Towards Data Science – Building NLP pipeline with Twitter data
https://towardsdatascience.com/sentiment-analysis-on-twitter-using-python-1e420e7d2f43
23. Analytics Vidhya – End-to-End Sentiment Analysis
https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-using-python/
Tools and Platforms
24. Google Colab – Free cloud coding platform
https://colab.research.google.com/
25. Python Official Website
https://www.python.org/