ICEST-2K24 [International Conference on Engineering, Science and Technology]
In Association with
International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN - 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)

Machine Learning Techniques for Sentiment Analysis of Users in Urban Cities for Transportation System Using Social Media Data

Prof. G.G.Sayyad[1], Barbole Prachi Prakash[2], Dhainje Sandhya Balu[2], Rokade Rohan Shivaji[2], Rayate Gaurav Rajendra[2]

[1]Assistant Professor, S.B.Patil College of Engineering, Maharashtra, India

[2]Department of Computer Engineering, Savitribai Phule Pune University, Maharashtra, India

ABSTRACT

Twitter is a popular social media site where users can share their thoughts and feelings on a variety of topics. Sentiment analysis is a method of analysing data and retrieving the sentiment it contains. sentiment data is the application of sentiment analysis to social media data in order to extract the user's expressed sentiments. The study in this sector has steadily increased during the last few decades. The reason for this is the data complex format, which makes processing tough. Because the social data format is so short, it introduces a whole new set of issues, such as the use of slang and abbreviations. In this implementation we will cover somemethodology (SVM&NLP) used and models used, as well as describing a generalised Python-based approach

Keywords: Machine Learning, Sentiment Analysis, Tweet Classify etc.

I. INTRODUCTION

In the recent decade the usage of social media has increased drastically, this drastic increase in usage of social media has led to tremendous data collection and storage to respective social media servers. Social media companies have a tremendous amount of data stored in their database servers, these servers consist of data like user's contact details, photos, videos, audio files, personal information, feeds uploaded on the social media etc. These social media companies use these data for analysis purposes and for making appropriate decisions regarding their respective businesses. The data of users collected through user social media accounts can also be used for transportation purposes in cities, villages, states, or nations worldwide. These data can help authorities handle city or security personnel or governing bodies of that particular city or area. These data can be gathered from different social media companies or through their APIs (Application Programming Interface) which is accessible through some process of the same. Considering this project we can observe that people using social media account does activities such as interacting with each other, having a discussion about some events or activities to be performed, and posting feed like photos, videos, articles, audio files, etc. Through the above observation, we come to know that above mentioned activities can tell us what is going to happen or what has happened in the city or area. These activities can be positive or negative. These positive or negative activities may sometimes lead to traffic congestion, chaos, strike, traffic blockage (Jam), etc. that leads to disturbance in the traffic flow of the city, and such result may cause GDP change of the city and peace loss. To achieve such project objectives we will be using different algorithms, techniques, and methods that will help to collect data, extract them, pre–process them, analyze them, identify sentiments from them, classify them, and then make appropriate decisions related to traffic or city good wellness.

II. LITERATURE SURVEY

1. In the proposed methodology Sentiment Analysis, Naive Bayes classification and AdaBoost algorithms are used to detect sarcasm on twitter. By using Naive Bayes classification, the tweets are categorized into sarcastic and nonsarcastic. Sarcasm is a subtle type of irony, which can be widely used in social networks. It is usually used to transmit hidden information to criticize and ridiculea person and to recognize. The sarcastic reorganization system is very helpful for the improvement of automatic sentiment analysis collected from different social networks and microblogging sites. Sentiment analysis refers to internet users of a particular community, expressed attitudes and opinions of identification and aggregation.In this paper, to detect sarcasm, a pattern-based approach is proposed using Twitter data. Four sets of features that include a lot of specific sarcasm is proposed and classify tweets as sarcastic and non-sarcastic. The proposed feature sets are studied and evaluate its additional cost classifications.
2. Twitter can be identified as one of the largest social networking sites.A large number of users have accepted Twitter as a universal platform for spreading news, sharing articles and socializing with other people globally. Subsequently, such a high-volume, high-velocity surge of Twitter data generated at each second have the potential of being utilized for significant analytical and interpretation purposes. The objective of this paper is to demonstrate an easy and simple solution, called Tweet-Analyzer. We propose a system to extract real-time Twitter data and to represent the trending Twitter hash tags and active users on a bar graph. Tweet Analyzer also makes use of the user's current location coordinates to represent the tweets on a world map. The proposed system can be easily deployed and used for various real-world applications such as job search, news updates, and business intelligence.
3. Twitter can be identified as one of the largest social networking sites. A large number of users have accepted Twitter as a universal platform for spreading news, sharing articles and socializing with other people globally. Subsequently, such a high-volume, high-velocity surge of Twitter data generated at each second have the potential of being utilized for significant analytical and interpretation purposes. The objective of this paper is to demonstrate an easy and simple solution, called Tweet-Analyzer. We propose a system to extract real-time Twitter data and to represent the trending Twitter hash tags and active users on a bar graph. Tweet Analyzer also makes use of the user's current location coordinates to represent the tweets on a world map. The proposed system can be easily deployed and used for various real-world applications such as job search, news updates, and business intelligence.
4. Twitter produces a massive amount of data due to its popularity that is one of the reasons underlying big data problems. One of those problems is the classification of tweets due to use of sophisticated and complex language, which makes the current tools insufficient. We present our framework HTwitt, built on top of the Hadoop ecosystem, which consists of a MapReduce algorithm and a set of machine learning techniques embedded within a big data analytics platform to efficiently address the following problems: (1) traditional data processing techniquesare inadequate to handlebig data;(2) data preprocessing needs substantial manualeffort; (3) domain knowledge is required before the classification; (4) semantic explanation is ignored. In this work, these challenges are overcome by using differentalgorithms

combined with a Naive Bayes classifier to ensure reliability and highlyprecise recommendations in virtualization and cloud environments. These features make HTwitt different from others in terms of having an effective and practical design fortext classification in big data analytics. The main contribution of the paper is to propose a framework for building landslide early warning systems by pinpointing useful tweets and visualizing them along with the processed information. We demonstrate the results of the experiments which quantify the levels of overfitting in the training stage of the model using different sizes of real-world datasets in machine learning phases. Pharmaceutical innovation faces challenges. Research merges quantum computing and machine learning to revolutionize drug discovery, simulation, and safety assessment for expedited progress.[19] Our results demonstrate that the proposed system provides high-quality results with a score of nearly 95. The detailed survey provided in paper [13]. Cyberattacks surge. Cybercriminals seek efficient channels to spread malware via images. JPEGVigilant, a machine learning method, identifies malicious JPEGs using 10 derived properties.[18]

III.PROPOSED SYSTEM

A) Problem Statement:

Given a status or a comments of that status, classify whether that is of positive, negative, or neutral sentiment. For status or comments conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen and average count will be given at last.

B) Block Diagram:

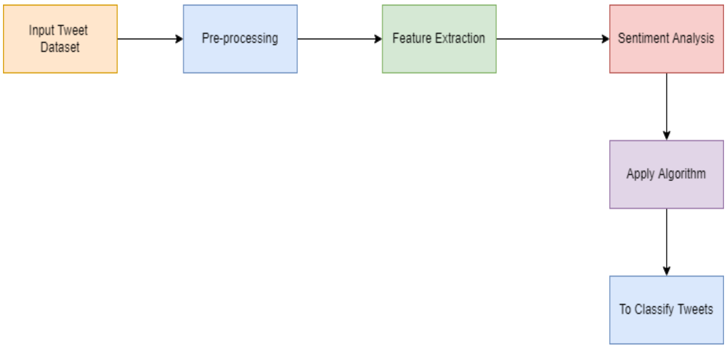


Fig:Architectural Diagram

C) Software Requirement:

RAM : 8 GB
Processor : Intel i5 Processor
IDE :Spyder
Coding Language : Python Version 3.8
Operating System : Windows 10(64 Bit)

D) Hardware Requirement:

Speed : 1.1 GHz
Hard Disk : 40 GB
Key Board : Standard Windows Keyboard
Mouse : Two or Three Button Mouse
Monitor : LCD/LED

E) Algorithm/Workflow of system:

- Start
- Identify the Data
- SVM,NLP
- Check the target

IV. METHODOLOGY IMPLEMENTATION

The model explained here can be extended to improve user experience, provide additional functionalities and optimize processing power. Machine Learning techniques are simpler and efficient than Symbolic techniques. These techniques can be applied for twitter sentiment analysis. Classification accuracy of the feature vector is tested using different classifiers like Nave Bayes, SVM, Maximum Entropy and Ensemble classifiers. All these classifiers have almost similar accuracy for the new feature vector .

1. Firstly we gather CSV dataset.
2. After data collection done then we prepare data using preprocessing.
3. Training:-After preprocessing done we have to train dataset by using SVM ,NLP algorithm.
4. Testing we use train model for testing and detect sentiments positive or negative and for accuracy of data we use confusion matrix showing the accuracy.

❖ Support Vector Machine (SVM) is a popular machine learning algorithm for sentiment analysis. Here's a high-level overview of how you can use SVM for sentiment analysis in a software project using social media data:

1. Data Collection: Gather social media data containing text (e.g., tweets, Facebook posts, reviews).
2. Data Preprocessing: Clean the data by removing special characters, numbers, and stopwords. Tokenize the text and convert it into numerical features using techniques like TF-IDF (Term Frequency-Inverse Document Frequency).
3. Labeling: Label the data as positive, negative, or neutral based on the sentiment expressed in the text.
4. Splitting Data: Split the data into training and testing sets for model evaluation.
5. Model Training: Train an SVM model using the training data. SVM tries to find the hyperplane that best separates the positive and negative samples.
6. Model Evaluation: Evaluate the model using the testing data to measure its performance in predicting sentiment.
7. Deployment: Integrate the trained SVM model into your software project to analyze sentiment in real-time social media data.

❖ Natural language processing (NLP) algorithms to analyze and classify text data into positive, negative, or neutral sentiments. One popular approach is to use a deep learning mode such as a recurrent neural network (RNN) or a transformer model like BERT. Here's a general outline of how you can use such models for sentiment analysis:

1. Data Collection: Gather social media data containing text (e.g., tweets, Facebook pos reviews).
2. Data Preprocessing: Clean the data by removing special characters, numbers, and stopwords. Tokenize the text and convert it into numerical representations suitable for chosen deep learning model.
3. Labeling: Label the data as positive, negative, or neutral based on the sentiment expressed in the text.

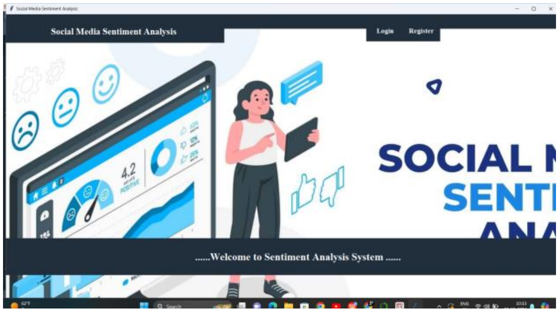
4. Splitting Data: Split the data intotraining and testing sets for model evaluation.
5. Model Training: Train a deep learning model (e.g., RNN, LSTM, BERT) using the trainig data. Use pre-trained embeddings (e.g., GloVe, Word2Vec) orfine-tune the embedding during training.
6. Model Evaluation: Evaluate the model using the testing data to measure its performance in predicting sentiment. Use metrics like accuracy, precision, recall, and F1-score.
7. Deployment: Integrate the trained NLP model into your software project to analyze sentiment in real-time social media data.

V. RESULT

Input to the program is always gives the sentiments with accuracy chart and sentiments of human being by analyzing the data.

Input 1 : This page shows the dashboard of project which consists of tabs like login and resister.

Output 1:


Social Media Sentiment Analysis
Social Media Sentiment Analysis
Login
Register
4.2
POSITIVE
SOCIAL M
SENTI
ANA

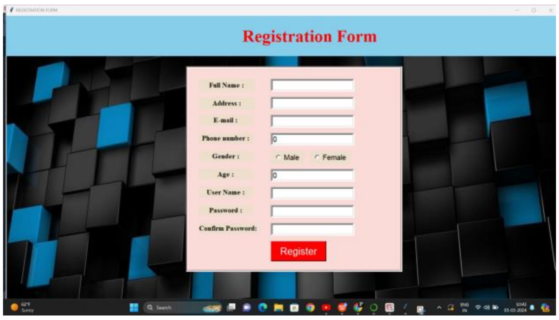
......Welcome to Sentiment Analysis System ......

Fig: output 1

Input 2: Registration form consist of some text bar for registering the details of user for database after entering details register button is used.

Expected output:All the entered data should accept and save in database.

Output 2:



Registration Form

Full Name :

Address :

E-mail :

Phone number : 0

Gender : ○ Male ○ Female

Age : 0

User Name :

Password :

Confirm Password:

Fig: output 2

Input 3: Enter right username and password.

Expected output: Check username and password is correct or not.

Output 3:

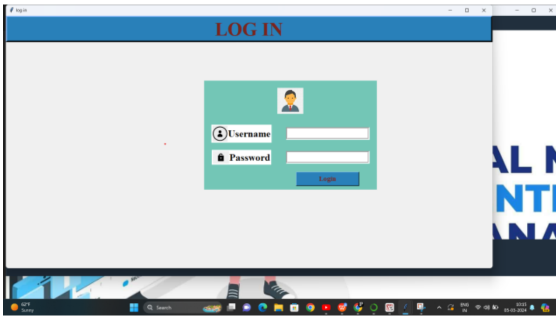


Fig: output 3

Input 4:This page shows the control panel and some functions for test the data.

Output 4:

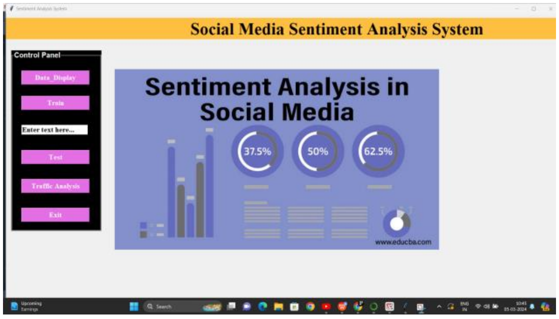
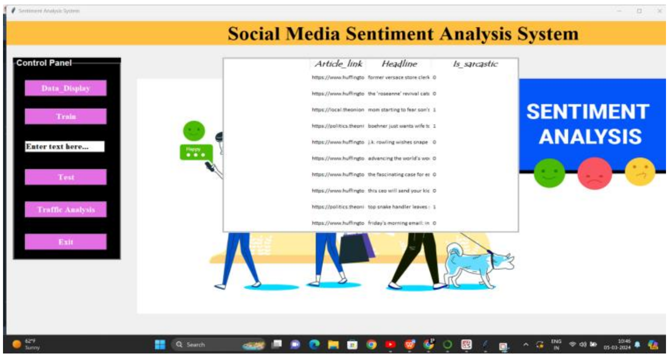


Fig: output 4

Input 5: Data display shows the trained data after entering sentiments data it shows the result.

Expected output:Analysing the data and give review like positive review and negative review.

Output 5:



Article_link	Headline	Is_sarcastic
https://www.huffingto	former versace store clerk	0
https://www.huffingto	the 'roseanne' revival cat	0
https://local.theonion	mom starting to fear son's	1
https://politics.theoni	boehner just wants wife to	1
https://www.huffingto	j.k. rowling wishes snape	0
https://www.huffingto	advancing the world's wo	0
https://www.huffingto	the fascinating case for e	0
https://www.huffingto	this ceo will send your kic	0
https://politics.theoni	top snake handler leaves	1
https://www.huffingto	friday's morning email: in	0

Fig:Output5

Output 5:

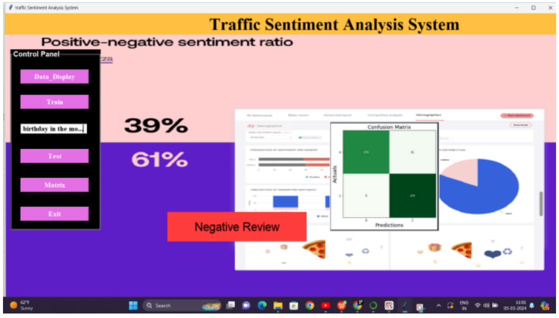


Fig:Output6

Output 5:

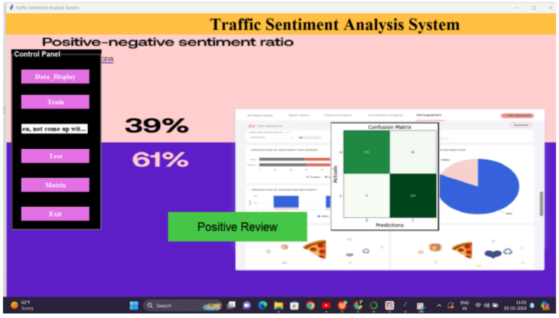
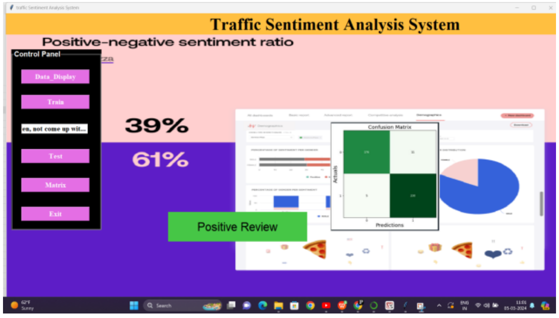


Fig:Output7

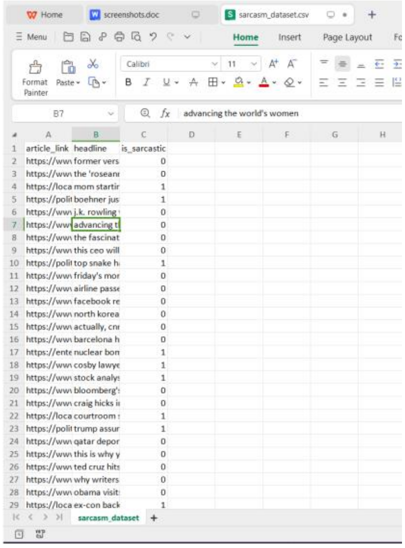
Input 7:Showing the accuracy of trained data.

Output 7:



Input 8:Showing the trained data in excelfile.

Output 8:



advancing the world's women

	A	B	C
1	article_link	headline	is_sarcastic
2	https://wwv	former vers	0
3	https://wwv	the 'roseanr	0
4	https://loca	mom startir	1
5	https://polit	boehner jus	1
6	https://wwv	j.k. rowling	0
7	https://ww	advancing tl	0
8	https://wwv	the fascinat	0
9	https://wwv	this ceo will	0
10	https://polit	top snake h	1
11	https://wwv	friday's mor	0
12	https://wwv	airline passe	0
13	https://wwv	facebook re	0
14	https://wwv	north korea	0
15	https://wwv	actually, cnr	0
16	https://wwv	barcelona h	0
17	https://ente	nuclear bon	1
18	https://wwv	cosby lawye	1
19	https://wwv	stock analys	1
20	https://wwv	bloomberg's	0
21	https://wwv	craig hicks i	0
22	https://loca	courtroom	1
23	https://polit	trump assur	1
24	https://wwv	qatar depor	0
25	https://wwv	this is why y	0
26	https://wwv	ted cruz hits	0
27	https://wwv	why writers	0
28	https://wwv	obama visit	0
29	https://loca	ex-con back	1

Fig:Output8

VI. CONCLUSION

Tweeter Analyzer" is presented that is implemented using simple programming concepts of Python and JavaScript. Tweeter Analyzer is capable of finding out the top ten trending hashtags and users at any given point in time and plotting them against their frequency using a bar graph.

- The model explained here can be extended to improve user experience, provide additional functionalities and optimize processing power. Machine Learning techniques are simpler and efficient than Symbolic techniques. These techniques can be applied for twitter sentiment analysis.
- Classification accuracy of the feature vector is tested using different classifiers like Nave Bayes, SVM, Maximum Entropy and Ensemble classifiers. All these classifiers have almost similar accuracy for the new feature vector .

VII.REFERENCES

[1]. ChaimaaLotfi,SwethaSrinivasan,MyriamErtz,ImenLatrous "Web Scraping Techniques and Applications: A Literature Review " on 2023.

[2]. Khin Than Nyunt,NawThiriWaiKhin "Web for career analysis based on youtube data APIs using web content mining abstract" on 2022.

[3]. Ajay Sudhir,NaveenGhorpade,Rohith S, S Kamalesh, Rohith R, Rohan B S "Web Scraping Approaches and their Performance on Modern Website"on 2022.

[4]. Chiapponi, Marc Dacier,OlivierThonnard,MohamedFangar,MattiasMattsson,VincentRigal "An industrial perspective on web scraping characteristics and open issues" on 2022.

[5]. DipaliShete,SachinBojewar ,AnkitSanghvi "Survey Paper on Web Content Extraction and Classification" 0n 2021.

[6]. Roopesh N, Akarsh M S, C. NarendraBabu "An Optimal Data Entry Method, Using Web Scraping and Text Recognition" 0n 2121.

[7]. Eric C. Dallmeier "Computer Vision-based Web Scraping for Internet Forums" on 2021.

[8]. ERDINC¸ UZUN "A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages" on 2019.

[9]. VidhiSingrodia, AnirbanMitra "A Review on Web Scrapping and its Applications" on 2019.

[10]. Rabiyatou DIOUF, EdouardNgor SARR, Ousmane SALL, Babiga BIRREGAH, Mamadou BOUSSO, SenyNdiaye ´ MBAYE "Web Scraping: State-of-the Art and Areas of Application" on 2019.

[11]. Gunawan, R., Rahmatulloh, A., Darmawan, I., and Firdaus, F. (2019). Comparisonof web scraping techniques: regular expression, HTML DOM and Xpath. In International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018) Comparison (Vol. 2):283-287.

[12]. Jadhav, A., Shinde, A., Nanavare, N., Ranmode, G., &Gavali, A. B. (2018). RFID based secure smart school bus system. IAETSD Journal for Advanced Research in Applied Sciences, 5(3), 127-134.

[13]. Prof. G. G. Sayyad, Ms. PrachiPrakashBarbole, Ms. SandhyaBaluDhainje, Mr. GauravRajendraRayate, Mr. RohanShivajiRokade, "A Study on Machine LearningTechniques for Sentiment Analysis of Users in Urban Cities for Transportation Using Social Media Data", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9 Issue 10, pp. 29-34, September-October 2023.

[14]. Karve, S. M. ., Kakad , S. ., SwapnajaAmol, Gavali, A. B. ., Gavali , S. B. ., &Shirkande, S. T. . (2024). An Identification and Analysis of Harmful URLs through the Application of Machine Learning Techniques. International Journal of Intelligent Systems and Applications in Engineering, 12(17s), 456–468.

[15]. Paigude, S., Pangarkar, S. C., Mahajan, R. A., Jadhav, P. V., Shirkande, S. T., &Shelke, N. (2023). Occupational Health in the Digital Age: Implications for Remote Work Environments. South Eastern European Journal of Public Health, 97–110.

[16]. Aaglave, K. N., Jadhav, S. S., Khatib, A. F., &Khurangale, R. L. (2023). A Survey on the Web Scraping: In the Search of Data.

[17]. Nalawade, V. S., Ashok, G. K., Hanumant, B. A., &Reshma, G. (2021). ENCRYPTION THEN COMPRESSION BASED SYSTEM USING GRAYSCALE BASED IMAGE ENCRYPTION FOR JPEG IMAGES.

[18]. Ekatpure, J. N., Kharade, N., Korake, D., Kshirsagar, D., & Mind, R. (2023). JPEG Vigilant: AI-Powered Malware Image Detection.

[19]. Ekatpure, J. N., Jadhav, P., Gavali, R., Kale, P., & Padasalkar, S. (2023). Pharmaceutical Data Optimisation Using Quantum Machine Learning.