International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 3 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

TEXT CLASSIFICATION ON TWITTER DATA.

[1] Dr. N. Bhanupriya, [2] Anumandla Prasanna Laxmi
[1] Assistant Professor,[2] Student
Department of MCA
Sree Chaitanya College of Engineering, Karimnagar

ABSTRACT:
Twitter has become one of the most influential social media platforms for sharing opinions, news, product reviews, and public discussions. The massive volume of user-generated content on Twitter presents valuable opportunities for extracting meaningful insights through automated text analysis. Text classification is a fundamental Natural Language Processing (NLP) task that involves categorizing textual data into predefined classes based on content and contextual information. This paper presents a text classification framework on Twitter data using machine learning techniques. The proposed system collects tweets from various domains and applies preprocessing methods such as tokenization, stop-word removal, stemming, and feature extraction to transform raw textual data into structured representations. Machine learning algorithms are employed to classify tweets into relevant categories, including sentiment classes, topic groups, or informational categories. The framework leverages textual features and linguistic patterns to improve classification accuracy and support automated analysis of large-scale social media data. Experimental results demonstrate that machine learning-based text classification effectively categorizes Twitter content with high accuracy and efficiency. The proposed approach provides a scalable and intelligent solution for social media analytics, sentiment monitoring, and information management.

I. INTRODUCTION

Sentiment Analysis which means to analyse the underlying emotions of a given text using Natural Language Processing (NLP) and other techniques to extract a significant pattern of information and features from a given large corpus of text. It analyses the sentiment and attitude of the author towards the topic of the subject mentioned in the text. This text can be a part of any document, post on social media or from any database source. Sentiments can be classified as objective or subjective, positive or negative or neutral.This classification can be either lexicon-based or machine learning based. Lexicon based classification makes use of already existing dictionary which has pre-assigned scores to each word and those scores are used to calculate the overall sentiment expressed in the sentence whereas in machine learning based classification, a model is trained using some ML algorithm using some labelled data and then use that model to predict a class for a new text. Twitter is nowdays easily one of the most popular microblogging platforms and millions of users express their views publicly on Twitter making it a rich source of information on public opinions and thus, helpful in sentiment analysis on any topic. In this paper, lexicon and machinelearning based approaches have been employed to reveal the prevailing sentiments of tweets. Textblob, SentiWordNet and Word Sense Disambiguation are giving the correct sense of a word in a given context for Lexicon-based Sentiment analysis while among the machine learning based algorithms MNB, LR, SVM and RNN Classifier have been used. In this paper, a comparison has been presented in terms of accuracy in predicting the sentiment of a given tweet. An ensemble

approach has also been implemented on the datasets which involve majority voting of MNB, LR and SVM and the resultsare being compared with the rest of the approaches.

II. LITERATURE SURVEY

1)**A survey of sentiment analysis techniques**

AUTHORS: Harpreet Kaur, Veenu Mangat, Nidhi

Sentiment analysis is an application of natural language processing. It is also known as emotion extraction or opinion mining. This is a very popular field of research in text mining. The basic idea is to find the polarity of the text and classify it into positive, negative or neutral. It helps in human decision making. To perform sentiment analysis, one has to perform various tasks like subjectivity detection, sentiment classification, aspect term extraction, feature extraction etc. This paper presents the survey of main approaches used for sentiment classification.

2)**Machine Learning-Based Sentiment Analysis For Twitter Accounts**

AUTHORS: Geetika Gautam ; Divakar Yadav

The wide spread of World Wide Web has brought a new way of expressing the sentiments of individuals. It is also a medium with a huge amount of information where users can view the opinion of other users that are classified into different sentiment classes and are increasingly growing as a key factor in decision making. This paper contributes to the sentiment analysis for customers' review classification which is helpful to analyze the information in the form of the number of tweets where opinions are highly unstructured and are either positive or negative, or somewhere in between of these two. For this we first pre-processed the dataset, after that extracted the adjective from the dataset that have some meaning which is called feature vector, then selected the feature vector list and thereafter applied machine learning based classification algorithms namely: Naive Bayes, Maximum entropy and SVM along with the Semantic Orientation based WordNet which extracts synonyms and similarity for the content feature. Finally we measured the performance of classifier in terms of recall, precision and accuracy.

3)**Text Based Sentiment Analysis**

AUTHORS: Biswarup Nandi, Mousumi Ghanti, Souvik Paul

One of the most important parts of running business successfully is analyzing customer's opinion and sentiments[1]. In this paper, the paragraph of sentences given by the customer is accepted and after extracting each and every word, they are checked with the stored (database has been maintained here) parts of speech, articles and negative words. After checking against the database, CFG is used to validate proper formation of the sentences. Each sentences are delimited by `.' or `?' or `!'. Emotions[2] are detected as - positive, negative or neutral sentence. There are 3 types of cases-1. If the paragraph contains more positive sentences than negative, then overall result will be positive. 2. If the number of negative sentence is greater than positive sentence, then the overall result is negative. 3. If there are same numbers of positive and negative sentences in the input paragraph, then the result is neutral and if a sentence has been entered that is a normal statement neither positive nor negative, that will be also considered as neutral.

4)**Sentiment analysis of twitter data**

AUTHORS: Hamid Bagheri, Md Johirul Islam

Nowadays, people from all around the world use social media sites to share information. Twitter for example is a platform in which users send, read posts known as `tweets' and interact with different communities. Users share their daily lives, post their opinions on everything such as brands and places. Companies can benefit from this massive platform by collecting data related to opinions on them. The aim of this paper is to present a model that can perform sentiment analysis of real data collected from Twitter. Data in Twitter is highly unstructured which makes it

difficult to analyze. However, our proposed model is different from prior work in this field because it combined the use of supervised and unsupervised machine learning algorithms. The process of performing sentiment analysis as follows: Tweet extracted directly from Twitter API, then cleaning and discovery of data performed. After that the data were fed into several models for the purpose of training. Each tweet extracted classified based on its sentiment whether it is a positive, negative or neutral. Data were collected on two subjects McDonalds and KFC to show which restaurant has more popularity. Different machine learning algorithms were used. The result from these models were tested using various testing metrics like cross validation and f-score. Moreover, our model demonstrates strong performance on mining texts extracted directly from Twitter.

5)**opinion mining of news headlines using sentiwordnet**

AUTHORS: Apoorv Agarwal, Vivek Sharma, Geeta Sikka, Renu Dhir

Opinion Mining (also known as "Sentiment Analysis") is an area of text classification which continuously gives its contribution in research field. The main objective of Opinion mining is Sentiment Classification i.e. to classify the opinion into positive or negative classes. SentiWordNet is an opinion lexicon derived from the WordNet database where each term is associated with some numerical scores indicating positive and negative sentiment information. Up until recently most researchers presented opinion mining of online user generated data like reviews, blogs, comments, articles etc. Opinion mining for offline user generated data like newspaper is unconcerned so far despite the fact that it is also explored by many users. As a first step, this paper present opinion mining for newspaper headlines using SentiWordNet. Further, most of the researchers implement the opinion mining by separating out the adverb-adjective combination present in the statements or classifying the verbs of statements. On the other hand, in this paper we analyze each and every word in the News headline whether it is a noun, verb, adverb, adjective or any other part-of-speech. During experiment, python packages are used to classify words. Then SentiWordNet 3.0 is used to identify the positive and negative score of each word thus evaluating the total positive/negative impact in that news headline.

III. SYSTEM ANALYSIS AND DESIGN

EXISTING SYSTEM:

In The Existing system used xgboost for house price prediction. This study aims to explore the important explanatory features and determine an accurate mechanism to implement spatial prediction of housing prices in Beijing..,, based on the housing price and features data in Beijing, China. Our result shows that compared to traditional hedonic methods, machine learning methods demonstrate significant improvements on the accuracy of estimation despite that they are more time-costly. Moreover, it is found that XGBoost is the Less accurate model in explaining and predicting the spatial dynamics of housing prices in Beijing.

DISADVANTAGES OF EXISTING SYSTEM:

- IN Xgboost, you have to manually create dummy variable/ label encoding for categorical features before feeding them into the models. Catboost/Lightgbm can do it on their own, you just need to define categorical features names or indexes.
- Training time is pretty high for larger datasets.
- Moreover, it is found that XGBoost is the Less accurate model in explaining and predicting the spatial dynamics of housing prices in Beijing.

  Algorithm: XGBOOST.

PROPOSED SYSTEM:

The proposed method is based on the linear regression. This project is proposed to predict house prices and to get better and accurate

results.The data for the house prediction is collected from the publicly available sources. In validation, training is performed on 50% of the dataset and the rest 50% is used for testing purposes.

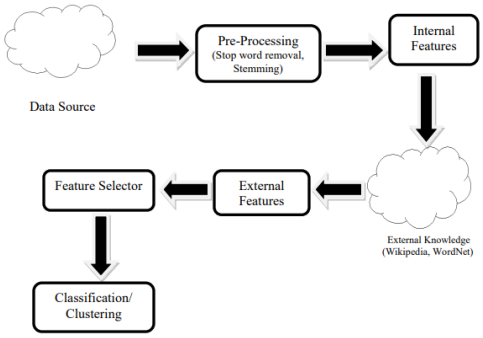
This technique splits the dataset into a number of subsets. At that point, it has been attempted for preparing on the entirety of the subsets; however, leave one (k-1) subset for the assessment of the prepared model. This strategy emphasizes k times with an alternate subset turned around for the preparation reason each time.

ADVANTAGES OF PROPOSED SYSTEM:

- The error free prediction provides better planning in the prediction of house price and other industries.
- This would be of great help for the people.
- This would be of great help to the people because the house pricing is a topic that concerns a lot of citizens whether rich or middle class as one can never judge or estimate the pricing of a house on the basis of locality or facilities available.
- Linear Regression is simple to implement and easier to interpret the output coefficients
- The ability to determine the relative influence of one or more predictor variables to the criterion value

Algorithm: Linear Regression (LR)

System Architecture:



IV. SYSTEM IMPLEMENTATION

Modules:

1. Data Extraction

An access token is used as authentication mechanism to get the data from twitter database. This access token is generated by login to your twitter account. Authorisation token that include API key are required to establish the connection and allow a search query. In the search query parameter of the twitter API a popular food brand is used. Other search parameters include returned tweet and language limit of 500 English tweets, and the result type of recent or popular tweets. This contains attribute and label types including tweet ID, username, number of retweets, original text, date and time it was created, language, and many others. The dataset was saved in MS Excel file.

2. Manual Data cleaning and Sentimental Labelling

saved data is set to be reviewed by human labeler whose task was to filter irrelevant data and retaining those which are significant to the experiment. The labeler was instructed to the label the text as the positive, negative or neutral with the following guidelines. Out of the 500 tweets, 352 neutral tweets were discarded since they serve no purpose in the analysis. The remaining dataset of 148 tweets were discarded since they serve no purpose in further experiment. The sentiment labeling is manually conducted by two (2) experts who classified each tweet as either positive or negative.

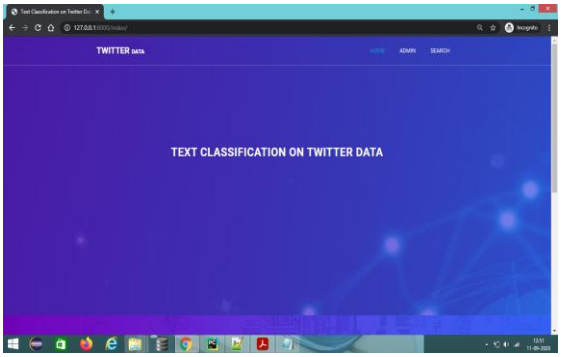
3. Feature Extraction

A pre-processing stage is needed in order to clean the dataset in order to train the model. Five operations were applied at the pre-processing stage. They were as follow transform Cases, Tokenize, Filter Operator, Filter Token, Stem Porter to extract the useful feature of the data. The following operations are described as follow – a) The Transform cases operator converts the characters in a document to lowercase so that the words, “Hello” and “hello” are the same. b) The Tokenize operator splits the texts of a document into a sequence of words (or tokens) by setting

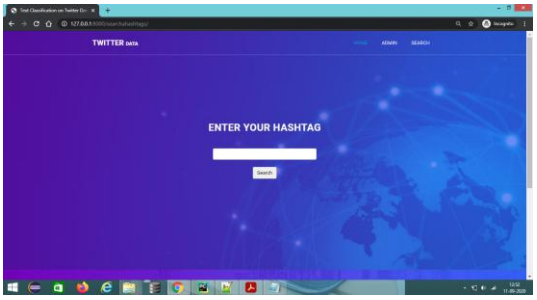
the nonletters mode as splitting points. c) The Filter Stop words operator removes the unwanted words from a document such as is, are, the, of, etc. These words are commonly used in the text but provide no content information. d) The Generate n-Grams administrator makes nlength of tokens in a report. The administrator will check words that much of the time tail each other. In content examination, it is basic that tokens are gathered so as to separate all the more significance. Single tokens, for example, "wellbeing", "living", "summer", and "breeze" may give little data instead of combined (or 2 gram) tokens, for example, "healthy_living" and "summer_breeze". In this manner, it will be clearer the setting of the related terms.

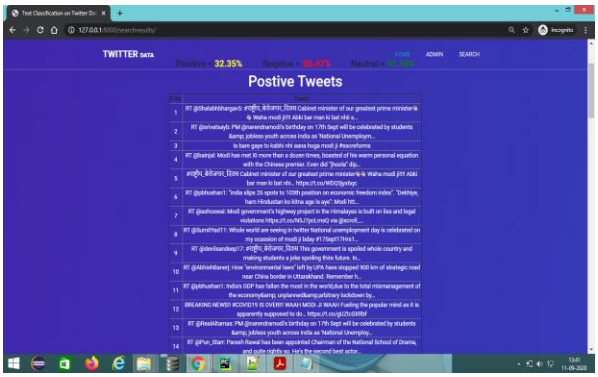
SCREEN SHOTS

Home page:

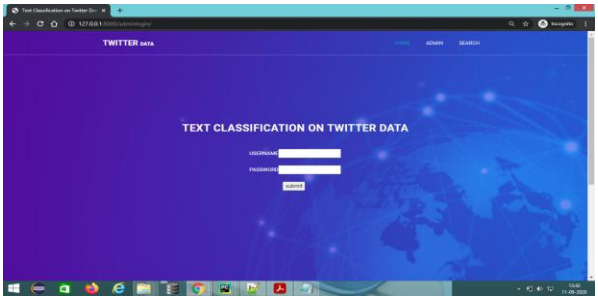


Search:

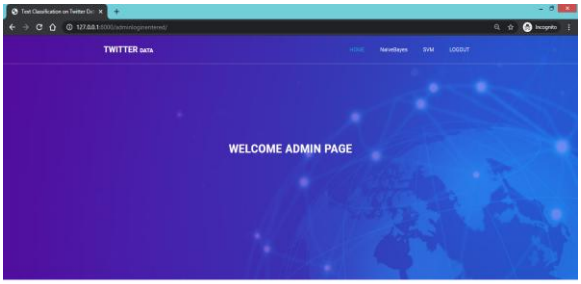




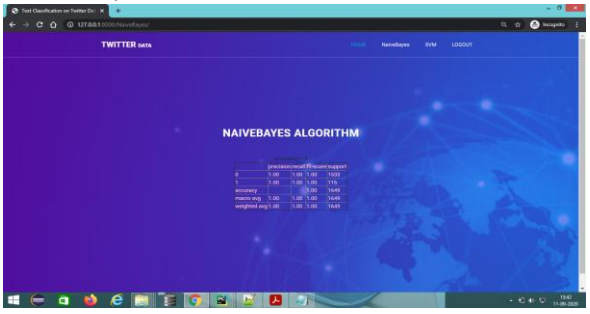
Admin:



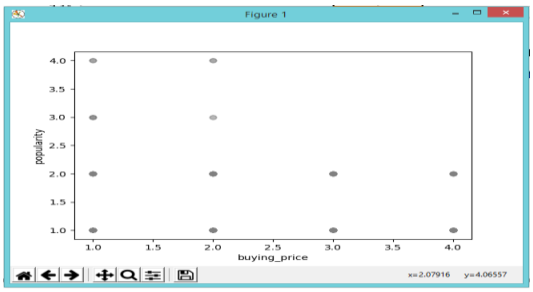
Admin home:



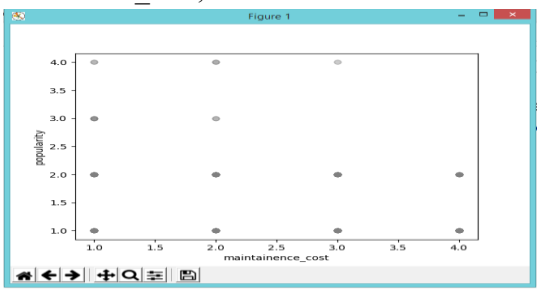
Naive Bayes:



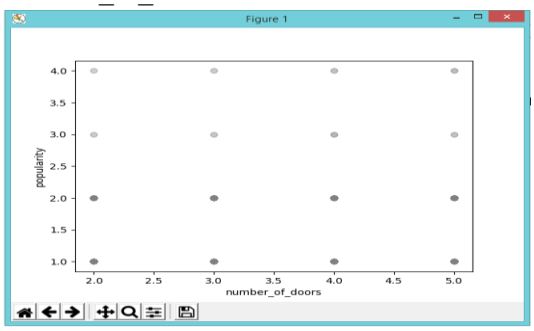
Svm:



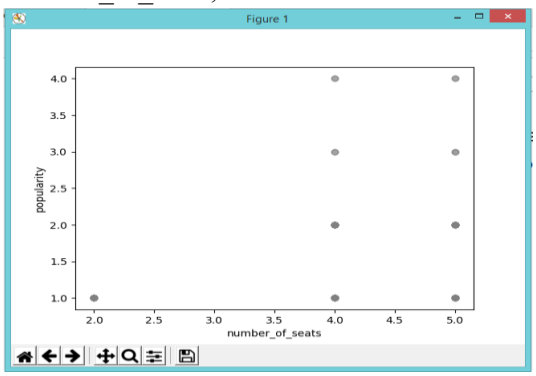
Maintaince_cost;



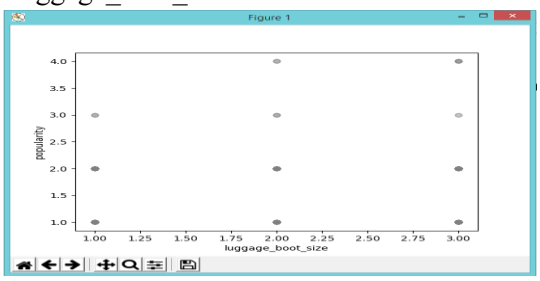
Number_of_doors:



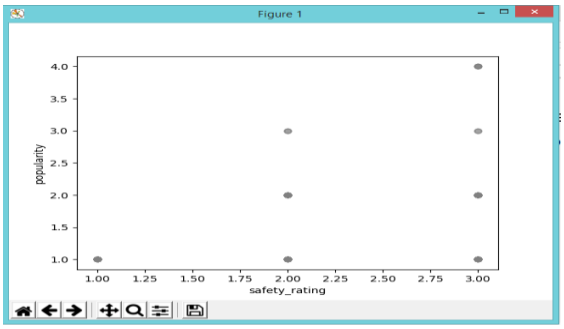
Number_of_seats;



Luggage_boot_size:



Safety_rating:



V. CONCLUSION

This study presented a machine learning-based framework for text classification on Twitter data. The proposed approach utilizes Natural Language Processing techniques and machine learning algorithms to analyze, process, and categorize large volumes of tweets into predefined classes. By employing effective preprocessing and feature extraction methods, the system successfully captures linguistic and contextual information necessary for accurate tweet classification. Experimental observations indicate that machine learning models achieve high classification accuracy and efficiently handle the dynamic and unstructured nature of Twitter data. Furthermore, automated text classification supports various applications, including sentiment analysis, trend identification, opinion mining, and social media monitoring. The proposed framework contributes to improved information organization, faster content analysis, and enhanced decision-making based on social media insights. Future research may focus on integrating transformer-based language models, multilingual text analysis, deep learning architectures, and real-time social media analytics to further improve classification performance and adaptability across diverse Twitter datasets.

REFERENCES

[1] Harpreet Kaur, Veenu Mangat, Nidhi, “A survey of sentiment analysis techniques”, February 2017,2017 International Conference on

I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).

[2] Ali Hasan, Sana Moin, Ahmad Karim And Shahaboddin Shamshirband, "Machine Learning-Based Sentiment Analysis For Twitter Accounts", February 2018.

[3] Biswarup Nandi, Mousumi Ghanti, Souvik Paul, "Text Based Sentiment Analysis", November 2017, 2017 International Conference on Inventive Computing and Informatics (ICICI).

[4] Bhagyashri Wagh, Prof. J. V. Shinde, Prof. P. A. Kale, "A Twitter Sentiment Analysis Using NLTK and Machine Learning Techniques", December 2017,International Journal of Emerging Research in Management & Technology.

[5] Hamid Bagheri, Md Johirul Islam, "Sentiment analysis of twitter data", Iowa State University .

[6] Apoorv Agarwal, Vivek Sharma, Geeta Sikka, Renu Dhir, " Opinion mining of news headlines using SentiWordNet", March 2016, 2016 Symposium on Colossal Data Analysis and Networking (CDAN).

[7] Rincy Jose, Varghese S Chooralil, "Prediction of election result by enhanced sentiment analysis on Twitter data using Word Sense Disambiguation", November 2015, 2015 International Conference on Control Communication Computing India (ICCC).

[8] Umar Farooq, Tej Prasad Dhamala, Antoine Nongaillard, Yacine Ouzrout, Muhammad Abdul Qadir, "A word sense disambiguation method for feature level sentiment analysis", December 2015, 2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA).

[9] Lesk Algorithm-Wikipedia

[10] Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani, "Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python", May 2017,International Journal of Computer Applications.

[11] Support Vector Machines(SVM) — An Overview

[12] Understanding Logistic Regression, (https://www.geeksforgeeks.org/understandinglogistic- regression/)

[13] Shadi Diab, Al-Quds Open University, Ramallah, Palestine, "Optimizing Stochastic Gradient Descent in Text ClassificationBased on Fine-Tuning Hyper-Parameters Approach.", December 2018,International Journal of Computer Science and Information Security (IJCSIS).

[14] Fenna Miedema, Prof. dr. Sandjai Bhulai, "Sentiment Analysis with Long Short-Term Memory networks" , 2018, Vrije Universiteit Amsterdam.

[15] Mochamad Ibrahim, Omar Abdillah, Alfan F. Wicaksono and Mirna Adriani, "Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation", November 2015, IEEE International Conference on data mining workshop(ICDMW).

[16] Ajay Deshwal and Sudhir Kumar Sharma, "Twitter Sentiment Analysis using Various Classification Algorithms", September 2016, International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO).

[17] Dan Li, Jiang Qian, "Text sentiment analysis based on long short-term memory", December 2016, 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI).

[18] Khin Zezawar Aung, Nyein Nyein Myo, "Sentiment analysis of students' comment using lexicon based approach", June 2017, 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS).

[19] Diego Tumitan, Karin Becker, "Sentiment-based Features for Predicting Election Polls: a Case Study on the Brazilian Scenario", August 2014, 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT).