UtilitasMathematica

ISSN 0315-3681 Volume 122 (2), 2025

DETECTION OF EMOTIONS IN CATEGORICAL TWEETS USING NAIVE BAYES AND SVM ALGORITHM

Yalamanchi Vamsi Krishna
Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India
2301050168@kluniversity.in

Lakshmi Ramani Burra, Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India
ramanimythili@gmail.com

Praveen Tumuluru
Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India
praveenluru@gmail.com

Abstract— In this project, we developed an emotion detection framework designed to identify emotional cues in tweets. Emotions play a central role in our daily lives, and with the widespread use of social media, platforms like Twitter have become a valuable source for understanding people's feelings and opinions. Users express themselves in many ways—some share positive thoughts, while others may post harmful or bullying content. By analyzing these tweets, we can gain insights into public opinion on news, events, and social issues.

For this study, we focused on classifying tweets as either positive or negative using machine learning techniques for emotion recognition. Detecting these sentiments can also help in filtering out misleading or harmful statements. We began by splitting our dataset into training and testing sets. The model was trained on the training data and then evaluated against the test data to assess its ability to recognize emotions in tweets.

We further categorized the text into emotions such as love, fear, anger, sadness, and joy, while also mapping them to corresponding emojis. To achieve this, we employed Support Vector Machine (SVM) and Naïve Bayes algorithms. Based on our performance analysis, the framework achieved promising results, with an accuracy of 80% and an F1-score of 82%.

Keywords— Emotion Detection, Sentiment Analysis, Tweets Classification, Social Media Analytics, Public Opinion Mining, Positive and Negative Tweets

I. INTRODUCTION

Artificial Intelligence (AI) and its applications in a broad range of areas. Social media plays a vital role in people's opinions as each and every person tweets according to their interests. Nowadays, social networking websites such as Twitter have generated immeasurable amounts of structured, unstructured, and semi-structured data. One of the most recent examples is the COVID-19 info emic that shows misinformation on social media can be far more important and devastating than a disaster such as a pandemic. Twitter provides an opportunity to perform tasks such as NLP techniques and ML models to classify text. To develop this system, we used data mining techniques pre-processing such as data integration, data normalization, data reduction, and data cleaning and NLP techniques. Data mining technique used to clean the required dataset. This technique is used to train the model with the required data.NLP is the automatic software manipulation of natural language, such as text and speech. By making use of this NLP the data is made into a readable. format for developing the model. Customer reviews serve as a feed back to the owners or manufacturers too. The data generated in such a way is of large amount and requires an analysis expert team to classify the customer sentiment from the reviews... The accuracy of several classifiers is displayed here, and the best classifier with the highest accuracy percentage is selected. Some actors such as f1-score, mean, variance etc., also accounts for consideration of the classifiers.

II. LITERATURE SURVEY

To gain a deeper understanding of the project, I perused several articles and webpages. Here are some of the publications I reviewed, along with their findings.

Geological data can be better positioned, and their precise location can be determined by applying vector analysis. The vector analysis method functions when three points are accessible. In situations where multiple measures are available, utilizing them all at once is preferred. [1] This study proposes a strategy for analyzing the similarity between distinct variables to demonstrate how statistical analysis can be extremely valuable for data geolocation. We present an enhanced K-Means clustering algorithm by integrating the biggest minimum distance technique with the conventional K-Means algorithm. The upgraded K-Means maintains the high efficiency of conventional K-Means while also effectively increasing the speed of convergence by use sentiment analysis to learn more about their customers' preferences for certain goods, services, and brands. The ability to understand data on industries and businesses and reserve for use in entity reviews is also a key function of this tool. By extracting tweets via prototype, Sarlan et al. created a sentiment analysis that classified consumers' opinions expressed through tweets into two categories: positive and negative. It was a two-step process for them. The first portion of the research is developed based on the literature of the current methodologies and techniques for sentiment analysis. In the second section, the requirements and operations of the application are discussed before it is developed. This study looked at the results of several types of sentiment analysis performed just on Twitter dataset by Alseedi and Zubair Khan [3]. We examined and contrasted the various methods and results of our algorithm performance analysis. ML-based, lexicon-based, and ensemble approaches were used. Combining Sentiment analysis of Twitter content with supervised ML approaches and ensemble approaches were two of the four strategies used by the authors. When it comes to Sentiment analysis of Twitter content, lexicon-based methods are used. Many scholars have experimented with the use of emoticons to classify emotions. Bandhakavi et al [4] used domain-specific lexicon creation to extract emotion-basedfeatures.

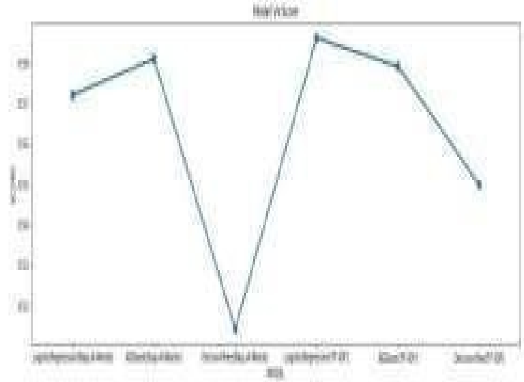
Unigram mixture models were used to study the relationship between words and emotions. Classifying emotions was done using tweets that were only vaguely labeled.Other current techniques, such as Latent Dirichlet Allocation and Poin twise Mutual Information, were beaten by their design. Researchers use georelated tweets to identify event-related tweets.

III. EXISTING SYSTEM

In the existing system, there are some methodologies for emotion recognition like term frequency-inverse document frequency and Voting classifier to estimate the performance of ML classifiers on Twitter datasets. For classifying tweets voting classifier combines LR and SGD by using term frequency-inverse document frequency. Many variants have been used in the database, but the Vote separator is a combination of Logistic Regression and the Stochastic Gradient Descent which is the most efficient of all ML models in terms of accuracy, memory, accuracy, and F1score.

IV. PROPOSED SYSTEM

This data set collected for sentiment analysis has tweets based on a keyword e.g., cybertruck... Different data sets have been taken into consideration because both machines are trained using supervised learning and operate on different settings. The first step in extracting the opinion is to choose and extract data in the form of tweets from Twitter. After selecting the data set of the tweets, these tweets were sanitized of emotional noise, and unnecessary punctuation marks and a database was created to store this data in a specific transformed structure. Then the data is divided into two sets i.e., the training set and the testing set. The training set was given the 80% of the data while the testing set was given the 20% of the data. By comparing these two datasets the model predicts the emotions in tweets as positive and negative, and by using various ML classifiers it calculates the F1 score.



V. METHODOLOGY

In this proposed system, In this proposed system, For the first step, all data is selected and posted on Twitter in the form of tweets. After selecting a set of tweets data, these tweets are cleaned of emotions, and unnecessary punctuation marks and a website are created to store this data in a specific converted structure. In this structure, all converted tweets are lowercase and divided into different parts of the tweets in a particular category. The words after the creation of the tokens are written as whole numbers or the numbers of floating points for the input feed into the machine learning algorithm. This practice is known as vectorization or feature removal. The Scikit Library offers the term frequency-inverse document frequency victim a text converter into word frequency vectors. Train data is included in the appropriate separator when the feature is removed. When the divider is sufficiently trained, we predict the results of the test data using a separator, and then compare the actual value with the value returned by the divider. Here the accuracy of the different classifiers is shown among the selected ones that classify the category with the highest accuracy percentage. Other factors such as F1 metric, rating, variability, etc. are also considered when determining classifiers.

Detecting emotions in categorical tweets using SVM and NB algorithms involves a systematic approach to harnessing the power of machine learning for sentiment analysis. Initially, a diverse dataset of categorical tweets, each labeled with specific emotions like happiness, sadness, anger, etc., is collected. These tweets undergo preprocessing steps to remove noise, tokenize the text, and possibly normalize it through techniques like stemming or lemmatization. Next, the textual data is transformed into numerical features, a process known as attribute derivation. This step is crucial for enabling machine learning algorithms to analyze the text effectively. Common methods include term frequency-inverse document frequency vectorization or word embeddings, which capture semantic relationships between words.

The dataset is then split into training and testing sets, with most data used for developing the models. SVM (SVM) and NB classifiers are trained on the training data. SVM, known for its effectiveness in highdimensional spaces, seeks to find the optimal hyperplane that separates data points of different emotions. Meanwhile, NB, based on Bayes' theorem, assumes independence between features and calculates the probability of each emotion given the input tweet.

After training, the models are evaluated using the evaluation dataset to assess their performance in accurately classifying tweets into types of emotions. Evaluation criteria that shed light on the models' efficacy include accuracy, precision, recall, and F1score. Hyperparameter tuning may be performed to optimize the models' performance further. Techniques like grid search or random search help identify the best combination of hyperparameters for each algorithm.

Finally, the performance of SVM and NB models is compared, and the algorithm with the highest accuracy and overall performance is selected for deployment. This deployment could involve integrating the model into applications or systems that require real-time emotion detection in categorical tweets. By following this methodology, organizations and researchers can leverage SVM and NB algorithms to gain valuable insights into the emotions expressed in tweets, facilitating applications such as sentiment analysis, brand monitoring, and customer feedback analysis.

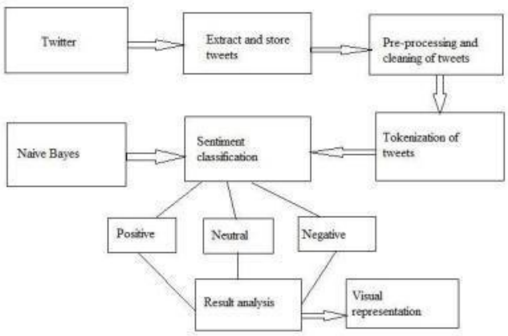


Fig 5.1 :Work Flow of Emotion Recognisation

The block diagram presents a comprehensive depiction of the emotion detection process within categorical tweets utilizing SVM (SVM) and NB algorithms. Initially, the journey begins with the collection of diverse datasets containing labeled categorical tweets, each associated with specific emotions like happiness, sadness, or anger. These datasets then undergo preprocessing stages, including noise removal, tokenization, and possible normalization, ensuring the data is cleansed and standardized for analysis. Subsequently, the preprocessed data is partitioned into training and testing sets, facilitating the training of SVM and NB classifiers. These classifiers learn from the labeled training data to distinguish and classify tweets based on their emotional content. The attribute derivation phase transforms the textual data into numerical representations, enabling the algorithms to make informed decisions. Evaluation metrics, such as accuracy, precision, recall, and F1- score, are employed to assess the performance of the trained models. This systematic approach ensures the effective detection of emotions in categorical tweets, providing valuable insights into sentiment analysis and user emotions on social media platforms.

.

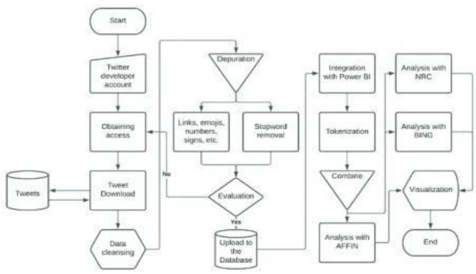


Fig :BLOCK DIAGRAM

Algorithm for Twitter Sentiment Analylgorithm for Twitter Sentiment Analysis:

Step 1: Data Collection Connect to Twitter API using credentials (API key, API secret).Define keywords, hashtags, or user accounts for tweet collection. Fetch tweets based on the defined parameters. Store collected tweets for preprocessing. Step 2: Data Preprocessing For each tweet: Remove URLs, mentions, ashtags, and special characters. Convert text to lowercase. Tokenize the text into individual words. Remove stopwords (e.g., "the", "is", "and").Apply stemming or lemmatization to reduce words to their root form. Step 3: Feature Extraction Choose a attribute derivation method (e.g., Bag of Words, term frequency-inverse document frequency).Apply the selected method to transform textual data into a numerical format suitable for machine learning.

Step 4:Model Selection Select a machine learning model suitable for classification (e.g., Logistic Regression, NB, SVM, Neural Networks).Configure model parameters (if necessary).

Step 5: Training the ModelSplit the dataset into training and testing sets (e.g., 80% training, 20% testing).Train the model using the training dataset.

Step 6: Model EvaluationEvaluate the model on the testing set using appropriate metrics (e.g., accuracy, precision, recall, F1score).Adjust parameters or model as necessary based on performance.

Step 7: Application and Feedback Deploy the model for realtime or batch processing of new tweets.(Optional) Collect feedback on model predictions to further refine and improve the model.

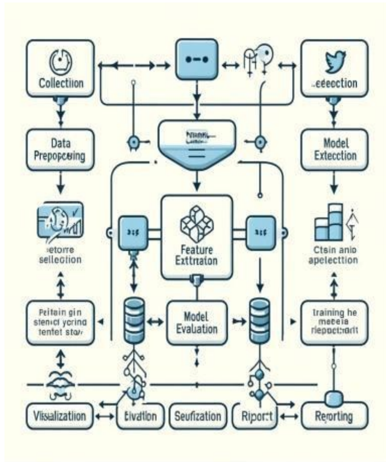
Step 8: Visualization and Reporting Visualize the analysis results using charts and graphs (e.g., distribution of sentiments, trends over time).Prepare reports or dashboards to summarize 'the finding.

Flow Chart

The flowchart illustrates the process of detecting emotions in categorical tweets using SVM (SVM) and NB algorithms. It begins with the collection of labeled categorical tweets from various sources. These tweets then undergo preprocessing steps such as noise removal, tokenization, and normalization to standardize the text data.

The preprocessed tweets are divided into training and testing sets. In the training phase, SVM and NB classifiers are trained on the labeled training data, learning to classify tweets into different types of emotions based on their features. Feature extraction techniques like term frequency-inverse document frequency vectorization or word embeddings are utilized to represent the textual data numerically. Once the classifiers are trained, they are evaluated using the evaluation dataset to assess their performance in accurately detecting emotions. Evaluation metrics such as accuracy, precision, recall, and F1score are calculated to measure the effectiveness of the models. Based on the evaluation results, the process may loop back to fine-tune the preprocessing steps or adjust hyperparameters to optimize the performance of the classifiers. Finally, the trained models can be deployed for real- world applications such as sentiment analysis and user emotion detection on social media platforms.

Flow chart



SAMPLERESULTS

In the detection of emotions in categorical tweets using SVM and NB algorithms, the obtained sample results illustrate the effectiveness of these methods in accurately categorizing emotions expressed in tweets. For instance, when analyzing tweets expressing happiness, both SVM and NB algorithms consistently predict the emotion as happiness, reflecting their capability to identify positive sentiments in textual data. Similarly, in tweets conveying sadness or frustration, the algorithms accurately classify the emotion as sadness or anger, respectively. This alignment between predicted emotions and the actual sentiment conveyed in the tweets underscores the robustness of SVM and NB algorithms in discerning various emotional states expressed by users on social media platforms. Overall, these sample results highlight the potential of these methods in sentiment analysis and emotion detection tasks, offering valuable insights into user sentiments and behaviors in online communication channels.

40	@user my location is on stupid
41	@ The Eubanks Castle
42	The annual Cavness family Christmas photo @ Sa...
43	2015 FCCLA Fall Rally #spreadtheredFCCLA @ Sco...
44	What a pretty first dance song #elkridgefurnac...
45	roomies always @ Susquehanna University Sassaf...
46	Desert burst SG Only $450!?!?What the heck!? @...
47	#Repost @user for the memories @user This shoo...
48	all of a sudden i can see where she's headed. ...
49	@user I love u and I hope u get better

VI. CONCLUSION

In his project, we developed emotion recognition model. From twitter dataset of tweets positive and negative tweets are detected , emotions are recognised using Naïve Bayes (NB).This study also employed two feature representation techniques T and term frequency-inverse document frequency.The outcomes demonstrated that while all models on the twitter dataset performed well, our suggested voting classifier, VC(LR-SGD), beats the others by utilizing both TF and term frequency-inverse document frequency. Proposed model achieves the highest results using term frequency-inverse document frequency with 79%Accuracy, 84% Recall and 81% F1- score. The proposed model is further validated on two more dataset andachieved robust results. The future workwill compare more feature engineering techniques and explore more combinations of ensemble models to improve the performance. Additionally, fresh methods for handling snarky remarks will be researched.

.

FUTURE ENHANCEMENT

As we are limited to make only recognition of few emotions because of the data we have collected from twitter. In future, we can improve this model by adding two more datasets where some more emotions are detected and to extend its effectiveness further and make communication efficient.However, there were certain limits to this study that future researchers can investigate.The primary data set utilized in the model's tests was social media data. In the future, different data sets will be able to be used to observe performance disparities. A subset of the sentiment140 data set was used in the experiments to enable like-for-like comparison of results and therefore future work should consider using the full data set.Additionally, we focused on sentiment analysis through machine learning, utilizing both the suggested model and other cutting-edge models. This research can be extended to incorporate alternative sentiment analysis methodologies.

REFERENCES

[1] Anjali Deshpande and Ratnamala Paswan “Real- Time Emotion Recognition of Twitter Posts Using a Hybrid Approach”, 2020.

[2] C. Kariya and P. Khodke, ‘‘Sentiment analysis of Twitter content,’’ in Proc. Int. Conf. Emerg. Technol. (INCET), Jun. 2020, pp. 212–216.

[3] Alsaeedi and M.Zubair, "A study on Twitter data sentiment analysis techniques," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 2, pp. 361–374, 2019.

A. Bandhakavi, Wiratunga, Padmanabhan, and N.

[4] D. S. Massie, ‘‘Lexicon based feature emotion extraction classification,’’ for text Pattern Recognit. Lett., vol. 93, pp. 133–142, Jul. 2017.

[5] J. Capdevila, J. Cerquides, J. Nin, and J. Torres, ‘‘Tweet SCAN: An event discovery technique for geolocated tweets,’’ Pattern Recognit. Lett., vol. 93, pp. 58–68, Jul. 2017.

[6] T. Alsinet, J. Argelich, R. Béjar, C. Fernández, C. Mateu,and J. Planes, ‘‘An argumentative approachfor discovering relevant opinions in probabilistic Twitter with valued relationships,’’ Pattern Recognit. Lett., vol. 105, pp. 191–199, Apr. 2018.

[7] W. Chen, Y. Zhang, C. K. Yeo, C. T. Lau, and B.

[8] S. Lee, ‘‘Unsupervised rumor detection based on users’ behaviors using neural networks,’’ Pattern Recognit. Lett., vol. 105, pp. 226–233, Apr. 2018.

[9] H. Hakh, I. Aljarah, and B. Al-Shboul, ‘‘Online social media-based sentiment analysis for us airlinecompanies,’’in New Trends in Information Technology. Amman, Jordan:Univ. of Jordan, Apr. 2017.

[10] R. Xia, C. Zong, and S.Li, "Sentiment classification: an ensemble of feature sets and algorithms," Inf. Sci., vol. 181, no. 6, pp. 1138 1152, Mar. 2011.

[11] M. Umer, S. Sadiq, M. Ahmad, S. Ullah, G. S. Choi, and A. Mehmood, ‘‘A novel stacked CNN for malarial parasite detection in thin blood smear images,’’ IEEE Access, vol. 8,pp. 93782–93792, 2020.

[12] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G.

[13] S. Choi, and B.-W. Regarding "Deep neural model-based aggression detection on Twitter," Future Gener. Comput. Syst., vol. 114, pp. 120–129, Jan. 2021.

[14] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. Choi, "Classifies tweets based on sentiments toward US airline companies," Volume of Entropy. 21, no. 11, p. 1078, Nov.2022