NATURAL LANGUAGE PROCESSING ALGORITHMS FOR PUBLIC SENTIMENT IN INDIA ON COVID-19 PANDEMIC

[1]S. VENKATESWARA RAO, [2]Dr. K.G.S VENKATESAN

[1]HOD, Dept. of CSE, Megha Institute of Engineering and Technology for Women, sabbineniv@gmail.com

[2]Assistant Professor, Dept. of CSE, Megha Institute of Engineering and Technology for Women, venkatesh.kgs@gmail.com

Abstract: *Sentiment analysis and opinion mining are research areas that investigate people's opinions, sentiments, evaluations, and emotions from written language. It is one of the most active research areas in natural language processing (NLP) and is also extensively studied in data mining, text mining, web mining. Sentiment analysis methods are being utilized in almost every industry and social domain because opinions are necessary to essentially all human actions and are key influencers of our actions. For this reason, when we need to make a decision, we usually attempt out other people's opinions. Nowadays, all groups of people believed in the impact of the COVID-19 pandemic. Various authors proposed different machine learning-based algorithms to detect sentiment opinions from the Twitter datasets. This research aims to detect sentiment polarity on COVID-19 to know the people's feelings, such as positive, negative, or neutral. In this research, a machine learning-based XGboost (Extreme Gradient Boosting) algorithm is applying to COVID-19 tweets and analyses the sentiment. We need to recognize how people are responding to the virus on Twitter. This research focuses on various traditional machine learning algorithms implemented for sentiment analysis and compares current algorithms.*

Keywords: *Natural language processing, sentiment analysis, opinion mining, machine learning, COVID-19.*

1. INTRODUCTION

Sentiment analysis or poll is a computational examination of opinions, values, and emotions on the speed of personalities and events and their characteristics. It has attracted many fans from academia and business due to its many difficult research issues and wide variety of programs (B. Liu 2010). Opinions count because whenever we need to outline them, we want to note the opinions of others. It always applies not only to humans but also to companies.

However, there were hardly any mathematical notes in the surveys before the discovery of Network Generation, as few texts of opinion were available. Later in life, when a person had to choose, he was forced to invite friends and family to complain. When employers wanted to find standard reviews of their products and services, they used behavioural surveys and focus groups. However, with the explosive growth of social media content on websites in the current era, this quarter has turned around. People can now post product reviews on business websites and express their views on almost anything on discussion forums, blogs, and social networking sites (Diana et al., 2009).

Diagnosis of emotions is usually not a single task but rather a multifaceted annoyance (Bing Liu 2010) that includes many subheadings, including item identification, feature exclusion, synonymous clustering, emotion orientation, and Integration (B. Pang et al. 2008). The survey covers strategies and practices that promise to enable a quick feedback framework. Here, the focus is on techniques that try to deal with new and difficult situations arising from the use of emotion-aware packages instead of those already in the full assessment based on additional conventional facts and observations. There is a need to create an integrated device to deal with all the multifaceted issues. Furthermore, the analysis of emotions is a popular topic because almost every person has a unique concept. All users use web platforms and regularly share their views, opinions, and comments in today's world. This type of data can be extracted and used successfully through various device awareness strategies.

Every low-level business today has felt the effects of the COVID-19 epidemic. The Coronavirus (COVID-19) has spread worldwide and has become a rapidly evolving epidemic. The virus first appeared and was diagnosed in the 2019 novel Corona Virus (2019-nCoV). The virus spreads through the square. Started in Wuhan, China [Zhu, N. Et al. 2020]. In early March 2020, the epidemic spread to Indonesia. Authorities play a key role in managing epidemics and are committed to neutralizing and preventing epidemics through various applications. Launched programs offer the provision, support and stabilization of clinical gadgets.

The primary function of emotion analysis is to categorize the polarity of a particular text within a report, sentence, or feature level, whether the record, the sentence, or the opinion expressed in the entity indicates that the given feature is strong, terrifying, Or incredibly neutral. In support, the sentiment evaluation also ranks the sentence on the premise of affective states along with "happy," "unhappy," and "angry.".

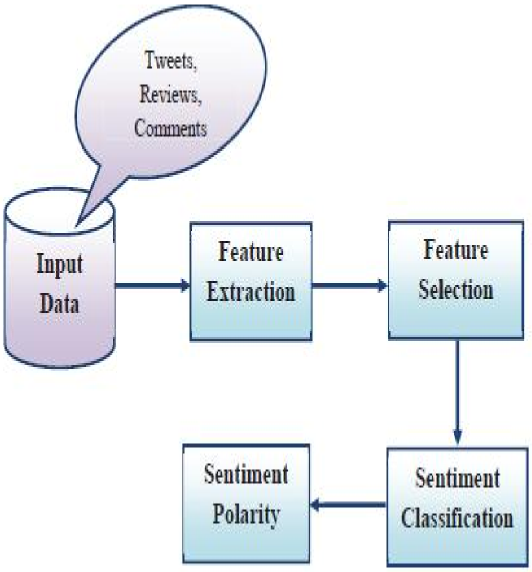

Fig.1 Process of sentiment analysis

Sentimental analysis begins with data entry, including tweets, reviews, or assessment of asset comments for emotional analysis. Beneficial households should be excluded from the income statistics, after which the required households will be identified. The category of emotions applies to the real characteristics by which emotional polarity is assessed.

1.2 CLASSIFICATION OF SENTIMENT ANALYSIS

Sentiment analysis is categorized into strategies: lexicon-based and machine-based retrieval of information.

The complete holistic approach works to support the number and weight of emotions. A core-based, glossary-based, and manual strategy approach has been incorporated with a focus on labelling. A completely holistic approach to a dictionary is classified as a corpus approach and a linguistic approach.

Machine learning is a category of algorithms that allows software program applications to be more accurate in predicting outcomes without explicitly programming. The main purpose of device learning is to develop an algorithm that can capture input information and expect output using statistical analysis.

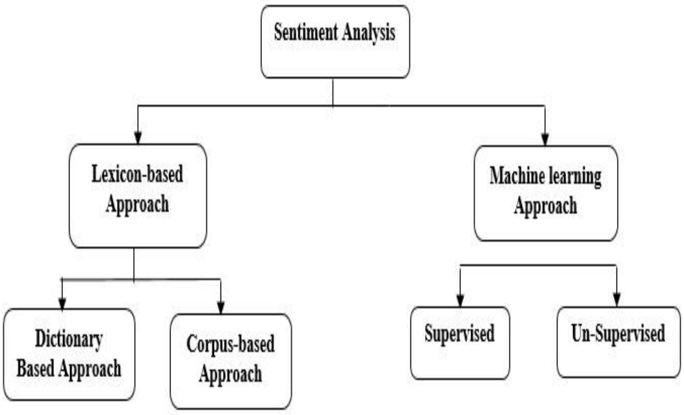

Fig.2 Classification of sentiment analysis

On the basis of machine learning sentimental analysis is classified as:

Unsupervised Learning: With hidden tags, untagged facts are categorized using unsupervised domains. Although there is no position for the student, the answers to mistakes or skills are not assessed.

Supervised Learning: Based on the general patterns, specific statistics are classified as the use of the supervised period. In the exam section, data is compiled, samples are examined, and academic data is used. These techniques are used to show and predict fact-based activities. Trained information is analysed, actions are developed for this period, and a complete pattern of events is examined.

2. LITERATURE SURVEY

Within the framework of the Literature Survey, several research papers were studied, providing the researcher with instant information about the research work. In this component, the relevant panels are primarily based on the author's reviews of new trends and emerging technologies related to each distinction.

Nikhil Yadav et al. [2020] Twitter, a micro-exercise running a blog, is a huge pool of public opinion expressed through various people, offers, organizations, products, and more. Emotion assessment is itself a public grievance test machine. Analysis of emotions, even in conjunction with Twitter, provides useful insights into what is being expressed on Twitter. This report highlights the specific strategies used to rank product reviews (which can be in the form of tweets). Bad or neutral. And use this assessment to evaluate the product market. The data used in this form, our online product reviews, were collected from Twitter and used to rank the best reviews.

Muvazima Mansoor et al. [2020] Effect of Coronavirus (COVID-19) infectious diseases on normal lifestyle. People from all over the industry have come to social media to express their well-known opinions and sentiments about this trend engulfing the industry during the storm. Twitter has seen a tremendous increase in tweets related to the unconventional Coronavirus. In addition, several device mastering models, including long-term memory (LSTM) and artificial neural networks for the type of emotion were performed, and their accuracy was determined. An analysis of research data was also completed for a set of facts that provides data on the number of cases filed continuously on some of the most affected international websites since the onset of the epidemic. From the onset of the epidemic to the present, it is possible to gauge the exchange of views on alternative issues.

Samuel et al.[2020] In addition to the coronavirus epidemic, another catastrophe has emerged in the form of widespread concern and panic due to incomplete and regularly misrepresented records. Therefore, the COVID-19 information crisis may need to be addressed and better understood, and public opinion scrutinized to implement appropriate policy messages and options.. They used a Naïve Bayes method to study 91% stable type accuracy for short tweets. Furthermore, we found that the logistic regression type method provides 74% cheaper accuracy with short tweets, and both methods performed poorly overall for long tweets. This document provides data on the increase in fear caused by the Coronavirus and describes the technologies, effects, barriers and skills involved.

Emadi, M et al. [2019] A comprehensive assessment of human emotions about a business, event, or character is important for business optimization, event analysis, and appreciation evaluation. In this study, we combine the results of traditional rankings and strategies based entirely on NLP to suggest a brand new way to gauge Twitter sentiment. The proposed technique used an ambiguous score to determine the significance of each ranking in the final decision. The offensive matrix is used in conjunction with the Coquet physical method, and the output of the classification needs to be included to generate the final tag. Our stories with specific emotional datasets on Twitter show that adding classification based on ambiguous requirements improves the general accuracy of the emotional class.

Kamel Ahsene djaballah et al. [2019] Terrorist organizations and their supporters use social media to promote terrorism. He expresses his views and beliefs by sharing his reviews on these social networks. In this newsletter, we support technology for finding extremist content on Twitter. To that end, we collected Arabic tweets related to the game of terrorism and categorized many of them into two emotional pieces of training: to provoke terrorism now or not. We use Word2vec and Word2vec with weight averages to represent our tweets. In addition, we use automated testing algorithms, especially SVM and random jungle, to anticipate sensations. We've done some experiments with a validation approach to validate our techniques. To test our effectiveness, we used three steps. The results show that Word2vec as a common weighting method is slightly more than the Word2vec approach.

Dorababu Sudarsa et al. [2018] an incredible feature of the wider public community has conveyed certain people are emotions innovatively. In addition, it is a broker with many records where clients can see the evaluation of unique clients. To do this, we first create the above data set, and then extract the feature from the dataset with more than one meaning, called capacitance vector, by posting the problem vector to this particular extent and Diagnose by factor. In particular, Naive Bayes, mostly entropy, and SVM rely entirely on the community of phrases along the threshold of semantic creation to extract synonyms and similarities with the textile properties of the material. When paused, we measure the classifier's general performance in the mirror phrase, accuracy and precision.

3. PROPOSED METHODOLOGY

Initially, the theoretical study of several sub-strategies for evaluating the sentiments of object identification, feature extraction, synonymous clustering, opinion orientation, integration, and classification are discussed. As mentioned above, the proposed research aims to develop an automated emotion-extracting machine to improve the overall performance of emotional analysis based entirely on a sub-technique of extracting a character based entirely on herbal language processing. Many researchers implement device awareness strategies to find emotional information from Twitter datasets. This study suggested an XGboost rating based on reading surveillance devices to find the emotional polarity of COVID-19 tweets. XGboost applied a more elaborate design formulation to control overfitting, giving it a more general performance than GBM (gradient boost system). It is one of the fastest-growing force tree programs, detecting potential damage to all possible partitions to create a new distribution.

XGBoost belongs to promotion algorithms that turn weekly learners into strong learners. One week of learning is significantly more than random feedback. Boosting is a later technique: trees are created one after the other using the information of previously planted trees. This method gradually learns from the records and tries to improve its prediction in subsequent iterations. XGboost is a gradually improved wood implementation designed for speed and versatility. It uses a gradient magnifying frame in the centre.

The conceptual model is performed in the form of a schooling version and a confirmation version.

Train the model

We can use different algorithms to train our version. However, we will use the XGBboost algorithm in this article. XGBoost is a Gradient Boosted Selection Wood development designed for performance and speed. XGBoost plays better than most predictive models. So, we're going to implement it to rank our tweets.

Testing the model on covid-19 tweets

This section will review our design and analyse emotions on covid-19 tweets. We need to see how people are responding to the virus on Twitter. The method of cleaning the tweet is similar to the previous one.

XGboost Algorithm steps

Step 1: Parallel Computing: It is allowed with parallel processing, i.e., when we operate XG boost, by default, it would use all the cores of our machine.

Step 2: Regularization: Regularization is a method utilized to evade overfitting in linear and tree-based representations.

Step 3: Missing Values: XGBoost is intended to manage missing values inside. The missing values are used in such a way that if there exists any bearing in missing values, it is caught by the model.

Step 4: Flexibility: In addition to regression, rating, and rating issues, it also contributes to user-defined objective features. An objective function is used to measure the performance of a model given a specific set of parameters. In addition, it helps in the diagnostic matrix described by the person.

Step 5: Save and Reload: XGBoost gives us a function to save and reload our fact arrays and models later. Assuming we have a large collection of facts, we will save the version and use it in the future instead of wasting time doing the calculations.

Step 6: Tree Pruning: Unlike GBM, where pruning stops with the observation of a passive stop, XGBoost extends the tree to its maximum depth and then cuts it down until the prevention feature is minimized.

SYSTEM ARCHITECTURE

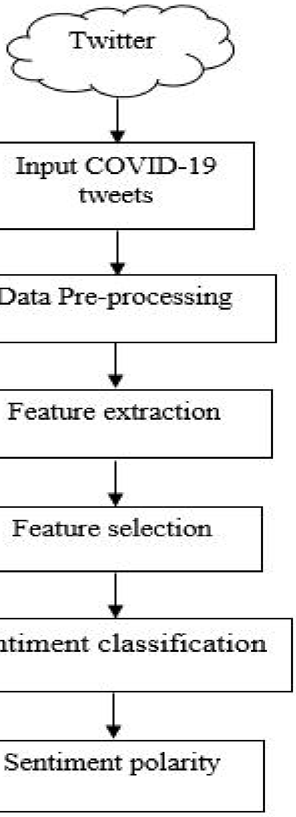

Fig. 3 Proposed model architecture

As shown in figure 3, the proposed model divided into various stages of data pre-process, feature extraction, feature selection, sentiment classification, and sentiment polarity.

Each stage described below,

Data pre-process

The tweets include a series of comments about the Covid 19 statistics that people express differently. The amplifier dataset used in this work has already been highlighted. The labelled estate set offers an awesome, good, or neutral value, allowing you to quickly analyze the information. There is a significant risk of data incompatibility and redundancy with polarity. Preferential logs affect the effects, and therefore, to increase the quality, pre-processing is done on the estate.

Feature extraction

Optimized data set configured after pre-processing has many distinct features. The feature extraction technique extracts the factor from the dataset. Next, the adjective is used to indicate good, bad, or neutrality for the duration of a sentence to determine the critique of those who use the Unigram form.

Feature selection

Feature selection refers to finding as much important information as possible within the extracted information to facilitate the processing and evaluation of available facts. The contribution of feature selection techniques in building an honest model is important. The main purpose of using feature selection is to minimize the number of key factors taken into account when creating a model to implement a range of features. The statistics needed to create a classified model usually contain the information needed to create the model.

Sentiment classification

The type of sentiment diagnosis is used to determine the form of emotion contained within a given statement. Classification algorithms obtain results by declaring the polarity of the statements to be excellent, awesome, or neutral. Ranking algorithms are capable of processing a large number of records. In the textual content category, the available texts are compared to a set of predefined sentences based entirely on the polarity of the sentence to be decided.

Sentiment polarity

The final result of sentiment analysis is received as the willpower of the polarity of emotions. Polarity scoring refers to labelling the text of a tweet as an expression of positive, negative, or neutral emotions. After educating the classifier, the emotional poles of the test information set are usually predicted. The Scikit Metrics module measures the accuracy, precision, memory and F1 score of a classifier.

The proposed system will be simulated. Simulation can be managed using a programming language such as Python or appropriate tools.

The overall performance of the proposed machine can be reviewed and compared with the current emotional analysis system's equivalent parameters, including performance, accuracy, and many more.

4. RESULTS AND DISCUSSION

When we tried to find the 'Sentiment' column, we found that most people have standard feelings about several issues that give us hope during epidemics. Very few people have a very negative view of Quaid 19.

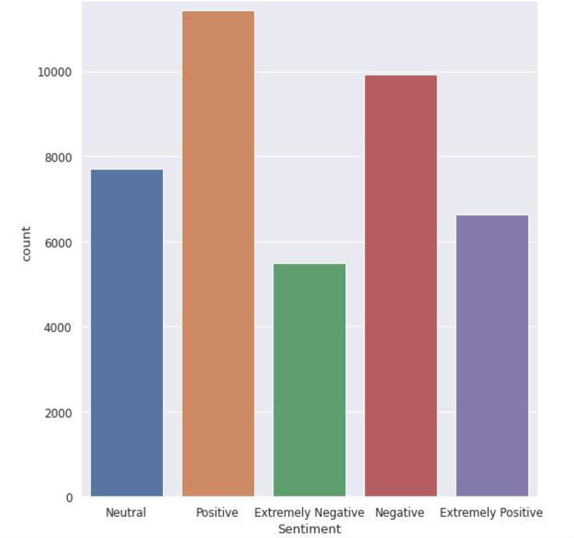

Fig.4 sentiment analysis on Covid-19 pandemic dataset

We analysed several papers on this situation and discovered interesting approaches to understanding the pandemic's short-term and long-term effects impacts.

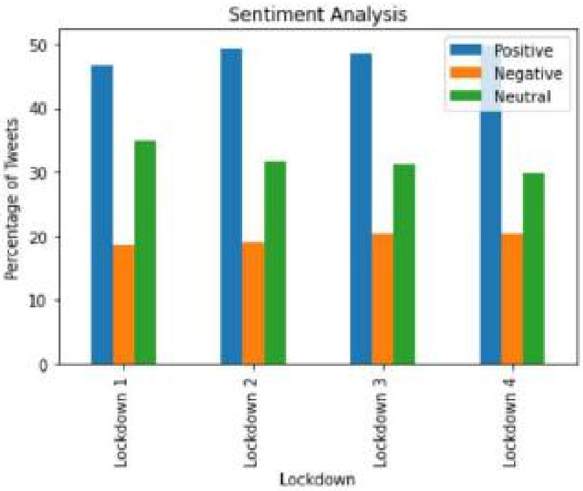

Fig. 5. Sentiment Category Distribution over 4 Lockdowns

5. CONCLUSION

The sentiment analysis approach classifies whether a text includes positive, negative, or neutral sentiments. One of the problems with sentiment analysis in Twitter stakes is its 140 qualities per post. There's not much space to complete a proper sentence after decreasing the length of a user's Twitter handle. This paper examined various machine learning algorithms used to find Twitter messages' sentimental polarity in earlier work. This study applies a machine learning-based XGboost (Extreme Gradient Boosting) algorithm to COVID-19 tweets and analyses the sentiment. The experiment was performed on the training data with 7613 tweets with varying text lengths and, similarly, the testing data with 3263 tweets of covid-19.

REFERENCES

1. *Bing Liu (2010), "Sentiment Analysis: A Multi-Faceted Problem", IEEE Intelligent Systems, pp.1-5.*
2. *Diana Inkpen, Fazel Keshtkar, 2009, "Using Sentiment Orientation Features for Mood Classification in Blogs", IEEE, pp. 1-6*
3. *L. Lee and B. Pang, 2008, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135*
4. *Zhu, N. et al. (2020) "A Novel Coronavirus from Patients with Pneumonia in China, 2019" , pp. 727–733.*
5. *Muvazima Mansoor, V R Badri Prasad, 2020, "Global Sentiment Analysis of COVID-19 Tweets Over Time", pp.1-7.*
6. *Samuel, Jim, et al, 2020, "Covid-19 public sentiment insights and machine learning for tweets classification.",p.314*
7. *Murad, H. R and Ahmad, A. R, 2020, "The Impact of Social Media on Panic during the COVID-19 Pandemic in Iraqi Kurdistan: Online Questionnaire Study", p. e19556.*
8. *Nikhil Yadav and Omkar Kudale, 2020, "Twitter Sentiment Analysis Using Machine Learning for Product Evaluation", pp.181-185.*
9. *Emadi, M and Rahgozar M, 2019, "Twitter sentiment analysis using fuzzy integral classifier fusion".*
10. *Prasadu Peddi (2019), "AN EFFICIENT ANALYSIS OF STOCKS DATA USING MapReduce", ISSN: 1320-0682, Vol 6, issue 1, pp:22-34.*
11. *Dorababu Sudarsa, L. Jagajeevan Rao, 2018, "Sentiment Analysis for Social Networks Using Machine Learning Techniques", pp.473-476.*
12. *Prasadu Peddi (2019), Data Pull out and facts unearthing in biological Databases, International Journal of Techno-Engineering, Vol. 11, issue 1, pp: 25-32.*