ISSN No:-2456-2165

An Analytical Insight of Omicron Sentiments by N-Gram Using Machine Learning

N. Narasimha Rao[1]
[1]Assistant Professor, NRI Institute of Technology, A.P, India-521212

V.Srujan[2]
[2]UG Scholar, Dept. of IT, NRI Institute of Technology, A.P, India-521212

A. Praneeth Surya[3]
[3]UG Scholar, Dept. of IT, NRI Institute of Technology, A.P, India-521212

D. Siva Teja[4]
[4]UG Scholar, Dept. of IT, NRI Institute of Technology, A.P, India-521212

Abstract:- The capacity to assess and forecast a variety of topics, including commercial requirements, environmental needs, election patterns (polls), governmental needs, etc., may be added to social media as an intelligent platform. This inspired us to start a thorough investigation of public thoughts and opinions on the COVID-19 epidemic on Twitter. The fundamental training data were gathered from tweets. Based on this, we have produced research using ensemble deep learning algorithms to forecast Twitter views more accurately than earlier works that do the same task. An N-gram stacked auto encoder supervised learning technique is used to extract features first. The collected features are subsequently used in a classification and prediction process using an ensemble fusion strategy comprising certain machine learning algorithms, including decision trees (DT), support vector machines (SVM), random forests (RF), and K-nearest neighbors (KNN). Using both mean and mode approaches, all individual findings are combined/fused for a superior forecast. The N-gram stacking encoder we suggest using in combination with an ensemble machine learning strategy surpasses all other known competitive techniques, including bigram auto encoders and unigram auto encoders. The public has a great deal of trust in government policy during the third wave, and they support all measures taken to contain the epidemic, including widespread participation in vaccine programmes.. The study's findings may be summarised by saying that people are getting past their fear of the disease.

***Keywords*:-** *Omicron Sentiment Analysis, N-Gram, Analysis, Social Media, Omicron, Tweets, Twitter, Big Data, Data Analysis.*

I. INTRODUCTION

Internet users have been growing quickly over the past ten years, and with the Covid-19 epidemic, social media became the preferred medium for expressing public sentiment. They are utilising the free microblogging website Twitter to impulsively share their ideas, happiness, and sadness. In order to forecast public opinions on socially relevant problems, researchers are very interested in evaluating public sentiment using data science techniques including natural language processing and machine learning approaches. Twelve well-known machine learning algorithms are utilised in the suggested research paper to analyse public opinions. Commonly used words are represented as n-grams; three of these n-grams—Unigram, Bigram, and Trigram—are gathered here, and predictions are made using the data .Today's online media has developed a reputation for its ability to switch as well as advertise. People divide their pricey opinions, assessments, and experiences on responsive destinations with the hope that others would profit from these. Twitter is one of these platforms where the general public communicates its opinions in brief terms, like 140 characters. Twitter serves as the corpus for open mining and sentiment analysis. These audits continue to be for anything and everything other than management, including movies, financial transactions, educational institutions, legal matters, and a great deal more. People provide their unbiased opinions about anything they wish in order for this audit to be seen as more comprehensive and real.

To complete this entire framework, five basic advancements are necessary. The first step is choosing how to prepare the data based on the type of concern. The second step is preprocessing the data to remove irrelevant information such as URLs, customer names, shoptalk vocabulary, imagery, and so on. [fig:7.1]. The third step is to establish associations through Twitter knowledge computation. Naive Bayes and Support Vector Machine are used for the alliance of tweets interested in different classes. The final step is to reveal the advance results.

II. TECHNOLOGIES USED

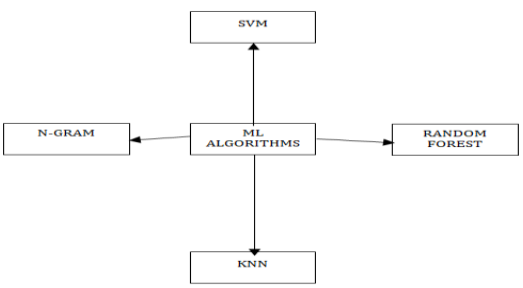

Fig 1 Technologies Used

ISSN No:-2456-2165

➢ *KNN*

KNN is a non-parametric, slow learning method. In order to forecast the categorization of the next sample point, it leverages data from many classes. KNN is non-parametric since the model is distributed from the data and no assumptions are made about the data being investigated.[Fig:2]

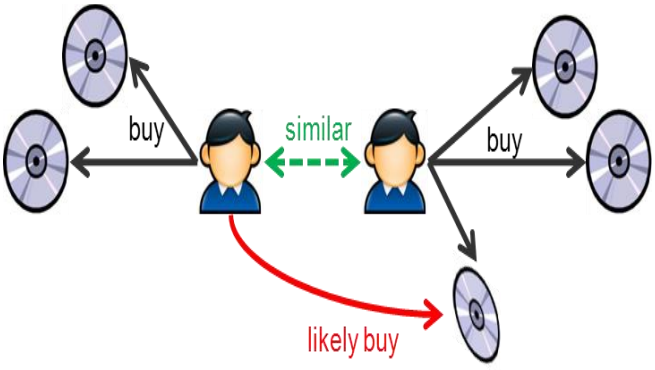

Fig 2 KNN

➢ *Random Forest*

A popular classification and regression approach is Random Forest. We may claim that the Random Forest Algorithm is one of the most significant algorithms in machine learning since classification and regression are the most significant parts of machine learning. The ability to categorize observations accurately is useful for a variety of commercial applications, such as determining whether a certain user will purchase a product or if a loan would fail or not.[Fig:3]

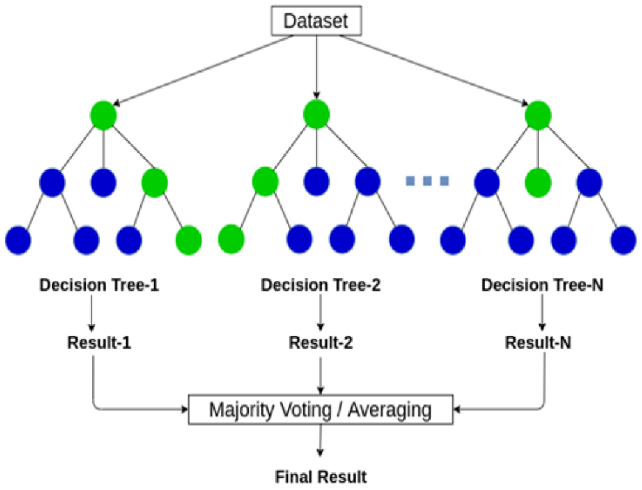

Fig 3 Random Forest

➢ *SVM:*

SVMs could offer a learning technique that is applicable to both regression and classification. A fast algorithm that produces favorable outcomes for a multitude of educational assignments is classified. It is not based on probability. A binary linear classifier that takes a set of input data and predicts, for every given input, which of the two available outcomes it belongs to.Classes are made up of the input. The support vector is composed of the training examples that are used for its formation. apparatus.[Fig:4]

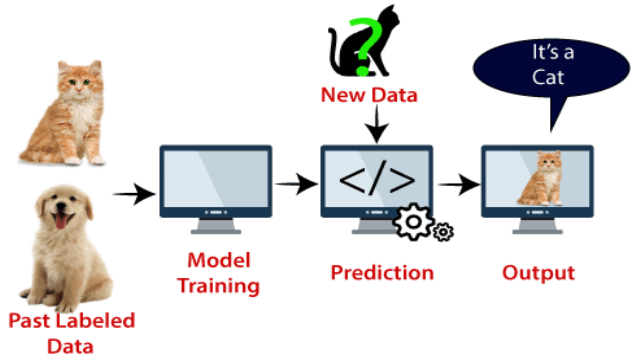

Fig 4 SVM

➢ *N-GRAM:*

I believe that N-gram is the simplest concept to comprehend in the entire field of machine learning. A combination of N words in a row is called an N-gram. For illustration, "Medium blog" is a two-word combination (a bigram), "A Medium blog post" consists of four words (a 4-gram), and "Write on Medium" has three words (trigram). That was quite dull and uninspiring. Indeed, yet we still have to take into account the likelihood associated with n-grams, which is quite intriguing.[Fig:5]

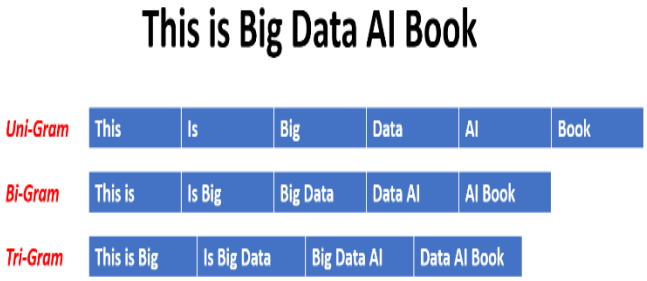

Fig 5 N-GRAM

III. SOFTWARE REQUIREMENTS SPECIFICATION

SRS is a captures complete description about how the system is expected to perform. It is usually signed off at the end of requirements engineering phase . It defines how software system will interact with all internal modules, hardware, communication with each other programs and human user interactions with a wide range of real like scenarios.

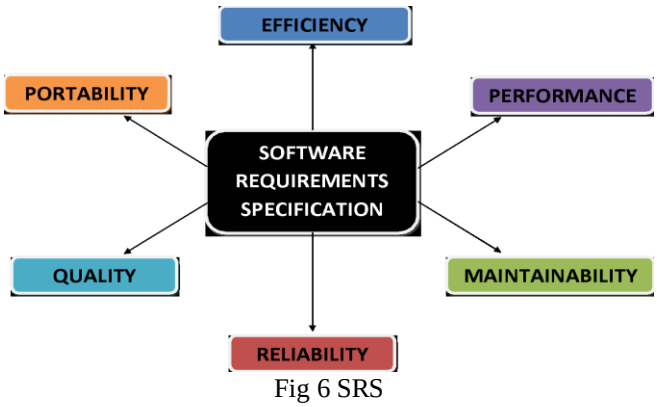

Fig 6 SRS

➢ *Reliability*

It would be more dependable and maintain the web application's updated information current. Once logged in, our username and password are hidden from view and are not visible to other users of the online application.

➢ *Quality*

This project has higher quality, and students, instructors, and administrators may access it from anywhere via the internet.

➢ *Maintainability*

The administrator would cleanly maintain the programme to keep the data secure and error-free.

➢ *Efficiency*

Downloading the information and answering questions would be more efficient for students, and instructors may upload the data as well.

➢ *Portability*

It would run without cost in any browser on any platform.

➢ *Performance*

Performance is higher because it would have provided excellent service to both instructors and students.

IV. EXISTING SYSTEM

We have seen evolution in covid-19 during the last two years. There are several varieties of covid, including omicron and delta. Omicron analysis only uses information or data that has been personally collected or observed by humans. Omicron analysis was manually updated. We employed machine learning with the n-gram approach to get around this.

➢ *Disadvantages of Existing System:*

- *Time Consuming Process.*
- *Disregards the Overall Context of the Text .*
- *Can be Reductive in its Approach.*

V. PROPOSED SYSTEM

To analyse the omicron data, we employ machine learning using Python. Using Python, do Omicron Sentiment Analysis When people were discussing the Omicron version on Twitter, the dataset that we are utilising for the job of Omicron sentiment analysis was first gathered. It may be acquired from Kaggle. So let's begin by importing the required Python modules and the dataset in order to do the Omicron sentiment analysis operation. All other known competitive approaches, such as the bigram autoencoder and the unigram autoencoder, are outperformed by our suggested strategy, which incorporates an N-gram stacking encoder into an ensemble machine learning scheme.

➢ *Advantages of Proposed System:*

- *Yields rich insights as it include qualitative and quantitative analysis.*
- *Easily replicable since it follows systematic procedures.*
- *Relatively inexpensive.*

VI. SYSTEM ARCHITECTURE

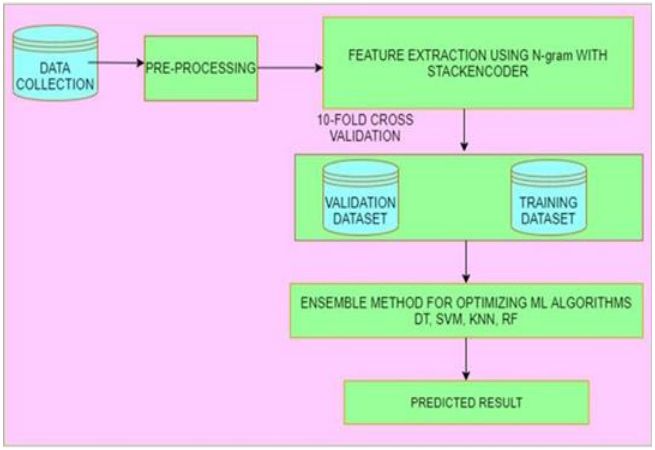

Fig 7 System Architecture

The goal of sentiment analysis is to determine automatically whether or not a particular textual item expresses opinions like positive or negative on an important issue. The valuable information from the text is retrieved using latent semantic analysis (LSA). Offline sentiment and semantic models have been developed for analysis in order to assess machine learning methods and get the best answers. The K closest neighbour (KNN), Random Forest (RF), Decision Tree (DT), and Support Vector Machine AI models have been utilised (SVM). In the proposed study, we used ensemble learning, combining the predictive models from the DT, SVM, RF, and KNN, and then combining all of the predictions using statistics like the mode or mean to get better outcomes.

➢ *Data Collection:*

Twitter data was gathered for this study using open-source data from the IEEE website [32]. This publicly accessible dataset included tweets from across the world that had been filtered using the terms "coronavirus," "_covid," "-covid-19," "sarscov2," "#covid19," "#covid 19," "2019-ncov," "#2019ncov," "sarscov2," "#covid," "sars cov2," etc. Tweet IDs were only accessible as of March 20, 2020 [28]. According to the IEEE website, tweet objects are what are used to collect information about tweets and extract the tweet ID.[Fig:7]

➢ *Data Preprocessing:*

A key component of the social media network idea analysis system is pre-processing the data. That is in the latent semantic analysis and sentiment analysis of Twitter's streaming data. The text data that may be accessed via Twitter are largely jumbled and loud. Data preparation is essential to provide the best outcome. [Fig:8].

ISSN No:-2456-2165

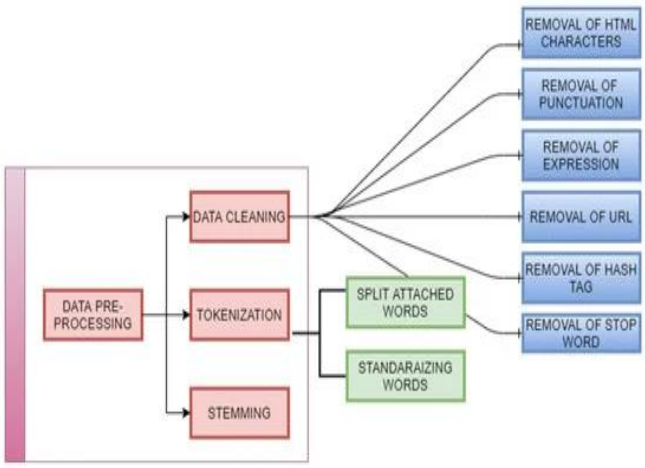

Fig 8 Data Preprocessing

- *Data Cleaning:*

The following procedures are used to eliminate undesirable material during this phase:

✓ *Removal of HTML Characters:*

HTML elements like >& that are contained in the original data are frequently seen in online data. We can transform these things into typical HTML tags by utilising Python's HTML parser.

✓ *Punctuation Elimination:*

All punctuation that is in accordance with the priority must be removed. Examples of necessary punctuation that must be kept in place are ".", ",", and "?" Other punctuation marks must be dropped.

✓ *Expressions that have been Removed:*

The text data may also include human expressions. These phrases should be deleted since they often don't relate to the text's subject matter.

- *Tokenization:*

Tokenization is the process of dividing lengthy words or strings into smaller units called tokens. There are two steps to it.

✓ *Split Attached Words:*

The initial phase is creating text data in a loose structure in the social network. The majority of tweets include phrases like "its epidemic," "totally lockdown day," and similar expressions. These things can be divided into their regular forms.

✓ *Standardizing Words:*

The textual information is not formatted properly, e.g., "loveeee u," "miss u," etc. These phrases need to be divided into their correct sections.

- *Stemming:*

The words are being returned to their original form through this procedure. That reduces the amount of words in the text from the root to the word type. As an illustrative example, the words "Jumping" and "jumped" will be removed in favour of the word "jump."

➢ *Feature Extraction:*

Feature extraction in the analysis of textual data is difficult. Text feature extraction is a method of representing a text message by removing text information from a vast volume of text processing. The procedure of feature extraction, which involves lowering the feature space dimensions, has been successfully implemented. We eliminate uncorrelated features during feature extraction. To recognise sentiment and perform latent sematic analysis at the word level, we provide an unique deep learning approach in this proposed study employing stacked encoders. The distributed word vector representation from the "n" gram is used as input in the newly suggested model, and the resulting continuous word vectors are merged with stacked auto encoder for fine-tuning word embeddings.

➢ *Stacked Autoencoder:*

- Words that are often used are spread in n-gram string data.
- The representation is then transformed into a reduced vector using the "SA" technique of the attacked autoencoders.
- With machine learning techniques like decision tree (DT), support vector machine (SVM), random forest (RF), and K-nearest neighbour (KNN), sentiment analysis and latent semantic analysis are applied (KNN).
- The quick forecast is produced by applying the ensemble approach to the ML model discussed above. The feature extraction framework is depicted in Figure.

VII. FUTURE SCOPE

This study has a number of restrictions. In order to find trends and the frequency of keywords used in this study, keywords connected to Omicron are worth considering. It's possible that the list of chosen keywords isn't full. Sentiment analysis can have an impact on future research in information systems, public mental health, and policy formulation. This study provides a helpful analysis to pinpoint the characteristics that lead to posting both good and negative tweets. We cannot claim that social media is only to blame for societal responses and the emotional impact it has on individuals. This is just based on correlational study; there are many more aspects that connect to psychological effect.

VIII. CONCLUSION

In our upcoming research, we will be utilizing more sophisticated deep learning techniques and multiple classifiers to further enhance the precision (e.g., surpassing 90%) of sentiment analysis on social mediaposts related to covid19. This is the method of evaluating the emotions Related to the Omicron form of coronavirus. The World Health Organization has labeled this novel form of coronavirus as a variant of concern. I trust that you enjoyed reading this piece on sentiment analysis of Omicron using machine learning techniques.

REFERENCES:

[1]. M. Lenzerini, "Data integration: A theoretical perspective," in PODS, 2002, pp. 233– 246.
[2]. D. Caruso, "Bringing Agility to Business Intelligence," February 2011, Information Management, http://www.informationmanagement.com/nfodirect/2009191/business intelligence metadata analytics ETL data management-10019747-1.html.
[3]. R. Hughes, Agile Data Warehousing: Delivering world-class business intelligence systems using Scrum and XP. IUniverse, 2008.
[4]. Y. Chen, S. Alspaugh, and R. Katz, "Interactive analytical processing in big data systems: A cross-industry study of map reduces workloads," Proceedings of the VLDB Endowment, vol. 5, no. 12, pp. 1802–1813, 2012.
[5]. M. Singh, A.K. Jakhar, S Pandey Sentiment analysis on the impact of coronavirus in social life using the BERT model
[6]. T. Vijay, A. Chawla, B. Dhanka, P. Karmakar Sentiment Analysis on COVID-19 Twitter Data 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)
[7]. G. Matošević, V. Bevanda Sentiment analysis of tweets about COVID-19 disease during pandemic 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)
[8]. xxx Apoorv Agarwal BoyiXie Ilia Vovsha Owen Rambow Rebecca Passonneau, "Sentiment Analysis of Twitter Data", Columbia University, Newyork.
[9]. K.-W. Fu, H. Liang, N. Saroha, Z. T. H. Tse, P. Ip, I. C.-H. Fung, How people react to Zika virus outbreaks on Twitter? A computational content analysis, Am. J. Infect. Control 44 (2016) 1700–1702.
[10]. D. Thorpe Huerta, J. B. Hawkins, J. S. Brownstein, Y. Hswen, Exploring discussions of health and risk and public sentiment in Massachusetts during COVID-19 pandemic mandate implementation: A Twitter analysis, SSM Popul. Health 15 (2021) 100851.
[11]. Chakraborty, K.; Bhatia, S.; Bhattacharyya, S.; Platos, J.; Bag, R.; Hassanien, A.E. Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. *Appl. Soft Comput.* **2020**, *97*, 106754. [**Google Scholar**] [**CrossRef**] [**PubMed**]
[12]. Shahsavari, S.; Holur, P.; Tangherlini, T.R.; Roychowdhury, V. Conspiracy in the time of corona: Automatic detection of COVID-19 conspiracy theories in social media and the news. *J. Comput. Soc. Sci.* **2020**, *3*, 279–317. [**Google Scholar**] [**CrossRef**] [**PubMed**]
[13]. Havey, N.F. Partisan public health: How does political ideology influence support for COVID-19 related misinformation? *J. Comput. Soc. Sci.* **2020**, *3*, 319–342. [**Google Scholar**] [**CrossRef**] [**PubMed**]
[14]. Pinter, G.; Felde, I.; Mosavi, A.; Ghamisi, P.; Gloaguen, R. COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. *Mathematics* **2020**, *8*, 890. [**Google Scholar**] [**CrossRef**]
[15]. Twitter: Standard Search Api. 2020. Available online: **https://developer.twitter.com/ en/docs/ tweets/ search/ overview** (accessed on 20 April 2020).

BIOGRAPHIES

Mr. N. Narasimha Rao is currently working as assistant Professor in Information technology at NRI Institute of technology, Pothavarappadu, Agiripalli, Krishna(dist), India. Completed B.tech in Lakireddy Bali reddy college of engineering and M.Tech in Vikas group of institutioins

V.Srujan is currently studying B. Tech with specification of Information Technology in NRI Institute of Technology. He done a summer internship project.

A. Praneeth Surya is currently studying B.Tech with specification of Information Technology in NRI Institute of Technology. He done a summer internship project .He completed two NPTEL courses.

D.Siva Teja is currently studying B.Tech with specification of Information Technology in NRI Institute of Technology. He done a summer internship project. He completed one NPTEL course.