Real-Time Twitter Trends Analysis Using Latent Dirichlet Allocation and Machine Learning

Ayush Ranjan
Department of Computer Science and Engineering
Sharda University, Greater Noida, India
ayush02sonu1@gmail.com

Sandeep Kumar
Associate Professor ,
Dept. of Computer Science and Engineering
Sharda University, Greater Noida, India
sandeep.csengg@gmail.com

***Abstract*:**

As far as social media is concerned twitter has become a major source of public data in the form of tweets. Twitter is an emerging source of large textual data(big data).People can easily express their opinions , reviews, interests and tastes about a particular event, topic ,product etc., occurring worldwide. This makes twitter a good source of valuable data which can be further used to perform sentiment analysis, know trending topics of public interests and public opinion on a particular product or event which can be beneficial for business growth and political parties to know public choices and they can take actions accordingly. The extraction of meaningful insights from the vast and dynamic stream of social media data is greatly facilitated by the implementation of topic modeling and sentiment analysis on Twitter tweets. The identification of recurring themes or subjects within a collection of tweets, which is the essence of topic modeling, serves to reveal patterns in discussions, thereby aiding in the comprehension of prevalent topics and trends. Concurrently, sentiment analysis, through the evaluation of the emotional tone of tweets, enables the discernment of whether the expressed sentiments are positive, negative, or neutral. The combination of these techniques provides researchers and businesses with valuable perspectives on public opinions, emerging issues, and user sentiments, thereby empowering them to make informed decisions and develop effective engagement strategies in the ever-evolving landscape of social media. But twitter data(tweets) are unstructured ,contains noisy data, URL, stop words, re-tweets video ,emoji etc., which need to be cleaned and preprocessed to perform proper sentiment analysis and use it to extract meaningful effective insights from it. This paper focuses on various methods of topic modeling and discovering latent topics ,text-mining approaches ,micro-blogging methods used in various researches .This paper focuses on latent dirichlet allocation method of topic modeling and text-mining to discover latent topics in tweets, micro-blogging , text-mining approaches .A proper survey is done on previous researches on this topic in this article.

Keywords:-LDA, Text-Mining, Topic Modeling, Micro-Blogging, Machine Learning, Sentiment Analysis, NLP.

I.INTRODUCTION

Twitter functions as a dynamic barometer of global discussion, exerting a profound influence on our understanding of current trends, public viewpoints, and shared areas of interest. Through real-time interactions, users participate in concise dialogues, thereby rendering the platform an invaluable instrument for measuring public sentiment. The timeliness of tweets metamorphoses Twitter into a prompt and responsive gauge, thereby unveiling societal concerns, breaking news, and cultural phenomena [1]. Its concise format and trending hashtags expedite the swift identification of popular conversations, thereby affording individuals, enterprises, and policy-makers immediate insights into the constantly evolving landscape of public discussion and interests. The global platform of Twitter enables individuals and entities to disseminate their thoughts, updates, and initiatives, thus enhancing their influence and public image. The rapid dissemination of information through tweets contributes to increased awareness and influence. Businesses utilize the power of Twitter to effectively reach their target audience and establish a strong brand presence [2]. By sharing concise marketing messages through tweets, companies can promote their products, provide updates, and engage directly with customers. The use of hash tags further amplifies visibility and fosters brand loyalty through interactive campaigns.

Twitter serves as a direct communication channel between businesses and their customers, allowing for prompt responses to queries and concerns. Active engagement with the audience enhances customer satisfaction and trust. The real-time nature of the platform enables businesses to promptly adapt and refine their strategies based on immediate feedback [3].

Twitter has emerged as a primary source for breaking news and real-time information, with journalists, news outlets, and eyewitnesses using tweets to report events as they unfold. The platform's reach and speed make it an invaluable tool for the rapid dissemination of information, particularly during times of crises and emergencies. Twitter acts as a barometer of public sentiment, reflecting the collective opinions and attitudes of users. By analysing trending topics and popular hash tags, businesses and policymakers can gauge public opinion, identify emerging trends, and adjust their strategies accordingly [4]. This real-time feedback loop aids in staying attuned to societal shifts and evolving preferences. Politicians, governments, and activists utilize Twitter as a virtual town hall to engage with the public, share policy updates, and mobilize support. Hash tags and re-tweets become powerful tools for organizing movements and raising awareness about social and political issues, fostering transparency and accountability. Twitter provides a platform for educators, institutions, and experts to share knowledge and engage in discussions, creating a global educational community [5]. From sharing research findings to hosting live Q&A sessions, the platform facilitates educational outreach and networking opportunities for academic institutions and professionals. Entertainment industry professionals, such as actors, musicians, and filmmakers, leverage Twitter to connect with fans, offer exclusive glimpses behind-the-scenes, and promote their work. The platform serves as a valuable tool for engaging with the audience and promoting the entertainment and pop culture industry .The vast amount of data generated through tweets offers valuable insights into market trends, consumer behaviour , and competitor activities, making it a goldmine for business intelligence. Advanced analytics and sentiment analysis tools help extract meaningful patterns from this data, facilitating data-driven decision-making and strategic planning for businesses [6].

Topic modelling and micro-blogging, particularly on platforms like Twitter, have revolutionized the manner in which we engage with information online. The influence of utilizing topic modelling techniques for extracting trending topics within the realm of Twitter tweets is profound, impacting various fields such as marketing, journalism, public opinion analysis, and more [7].

Impact of Topic Modelling on Twitter Trends:

1. Enhanced Content Discovery: The implementation of topic modelling algorithms, such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF), enables the identification of prevailing themes within a vast stream of tweets. This augmentation of content discovery is achieved through the categorization and organization of tweets around specific topics, aiding users in navigating through the abundance of information and concentrating on what is most pertinent to them [8].

2. Real-Time Insights:

The success of micro-blogging platforms like Twitter thrives on real-time interactions. Topic modelling greatly facilitates the extraction of trending topics in real-time, thereby providing users with immediate insights into on-going discussions and emerging themes. This capability is of immense value to businesses, journalists, and researchers seeking to remain up-to-date with current events and public sentiments [9].

3. Strategic Marketing and Branding:

For businesses, comprehending trending topics on Twitter is of utmost importance for strategic marketing and branding purposes. Through the identification of the most discussed subjects, companies can tailor their content to align with popular trends, thus increasing the likelihood of engagement and visibility. This strategic approach aids in maintaining relevance and capturing the attention of a wider audience [10].

4. Improved User Engagement:

The success of micro-blogging platforms greatly hinges on user engagement, and the ability to tap into trending topics plays a pivotal role in increasing interactions. By incorporating popular hash tags and themes into their tweets, users can actively participate in on-going conversations, expand their reach, and connect with a broader audience. This heightened engagement, in turn, contributes to the virality and visibility of their content [11].

Applications of Topic Modelling and Micro-blogging in Twitter Tweets:

1. News and Journalism:

News outlets effectively utilize topic modelling to sift through the extensive volume of tweets and identify emerging news stories. By comprehending trending topics, journalists can swiftly determine which issues are resonating with the public, enabling timely and relevant reporting. Additionally, the use of Twitter hash tags aids in the aggregation of user-generated content, providing a diverse range of perspectives on news events [12].

2. Public Opinion Analysis:

Understanding public sentiments and opinions is of utmost importance for businesses and policymakers. Topic modelling enables the extraction of prevalent themes from Twitter conversations, offering valuable insights into prevailing attitudes [13].

3. Event Monitoring and Planning:

During events, conferences, or product launches, monitoring Twitter trends provides organizers with real-time feedback. By analysing trending topics, event organizers can gauge the success of their initiatives, identify areas of interest, and adapt their strategies in real-time. This dynamic approach enhances the overall experience for participants and ensures the event remains relevant in the digital landscape [14].

In this article, we have discussed various methods used in topic modelling, text-mining ,micro-blogging discussed about various research works in introduction, literature review, research methodology and finally conclusion and discussion concludes the paper.

II. LITERATURE REVIEW:

Gaining comprehension of the American perception on COVID-19 via Twitter. Elucidation regarding the fluctuations in sentiment throughout the course of the pandemic. The analysis of topical patterns through topic modelling demonstrates a preoccupation with the economy, politics, and the dissemination of the virus. The popularity of sentiment analysis has increased in critical events as it allows for the extraction of valuable information. The existing body of literature examines sentiment analysis in the context of disasters and crises, with a specific emphasis on the COVID-19 pandemic [1]. Reviews research articles on tweets classification and clustering. Focuses on sentiment analysis, topic detection, and machine learning algorithms [2]. Enhances sentiment analysis and trend detection in tweets. Utilizes machine learning algorithms for analysing large volumes of data [2]. Discussions were centred around matters of justice, the charges against the suspect, an online petition, and instances of police brutality. Sentiment analysis revealed that Twitter users expressed neutral sentiments and provided objective perspectives. The word cloud depicted the public's demand for justice and the call to press charges against the suspect. The prevailing themes surrounding the murder of Gregorio were explored on Twitter. A detailed analysis was conducted on 1045 tweets to examine the sentiment, subjectivity, and themes present [3]. The TS-LDA model demonstrates superior performance compared to baseline methods in the task of identifying interesting tweets. The proposed approach specifically emphasizes the re-ranking of highly re-tweeted tweets based on their level of interestingness. Prior research on identifying interesting tweets has been examined through the lens of topical analysis. The ultimate goal is to identify tweets that are engaging for a wider audience. In terms of topic extraction, the novel TS-LDA model surpasses the performance of baseline methods. The primary focus remains on re-ranking the most-re-tweeted tweets based on their interestingness [4]. The LDA method is employed for the purpose of topic modelling on Twitter data. The method utilizes four distinct topics. The goal of clustering in data analysis is to identify shared characteristics and group data accordingly. The LDA method is particularly useful for

accurately clustering Twitter data when performing topic modelling. It optimizes the processing of tweet data for the purpose of extracting topics. Python programs are utilized for both data processing and clustering within these topics [5]. Four clusters of public reactions can be observed, namely Understanding, Action planning, Hope, and Reassurance. Within a span of 4.5 weeks, there is a shift from individual-centric topics to topics that focus on the community. The analysis of public reactions to COVID-19 was conducted using Twitter data. To efficiently extract public opinions, structural topic modelling was employed. Various aspects were investigated, such as problem-focused, emotion-focused, promotion focus, and prevention focus. These investigations led to the identification of the aforementioned clusters in public reactions. This approach enables policymakers to promptly respond to public opinions and reactions. Additionally, it aids healthcare workers in providing mental health services for individuals experiencing anxiety and loneliness [6]. The classification of tweets into positive, negative, or neutral sentiments is undertaken. The sentiment analysis on Twitter data employs a Dictionary Based methodology. A comparison is made between exams, apprenticeships, etc., to assess overall performance. The development of scoring criteria for different techniques is carried out. The classification of tweets into positive, negative, or neutral sentiments is facilitated for organizations. The extraction of sentiments from tweets for analysis is achieved through a specific method. A structured approach is provided for analysing sentiments expressed on Twitter [7]. Compared machine learning and rule-based approaches in sentiment analysis. Used SVM algorithm for classification process in sentiment analysis . Compares machine learning and rule-based approaches in sentiment analysis. Rule-based approach uses sentiment dictionary to determine sentiment of sentences. Enhances sentiment analysis using machine learning and rule-based approaches [8]. The system determined the demand for talent and displayed candidate ranks to meet hiring requirements. Selected sentences for analysis from resumes for openings in computer engineering. Used text mining tools to investigate interview robots in the hiring process. Based on resume analysis, the system paired recruiters with job candidates. using text-mining, natural language processing, and web crawling techniques. determined the need for skill and provided applicant rankings for particular roles. Recruiters and job candidates are matched by the system according to demand. raises the degree of fit between open positions and qualified applicants [9]. Reactions to booster doses were bad in more than half of the tweets.
Posts on social media emphasized how younger people do not require vaccinations. Examined Indian opinions on COVID-19 booster shots on social media. Social media data analysis is crucial for anticipating and averting catastrophes. Comprehending public opinion facilitates the effective execution of health policies. LDA algorithms make hidden subjects in posts on social media visible [10]. Public responses can be divided into four clusters: comprehension, action planning, hope, and reassurance. In 4.5 weeks, discuss themes that are community-centred rather than self-centred. Social media text mining effectively gathers thoughts and responses from the public. allows decision-makers to react quickly to abnormal conduct. examined how the public was responding to COVID-19 using data from Twitter. used twint , a web scraping tool, to gather pertinent tweets. It allows decision-makers to react swiftly to the beliefs and actions of the people. It helps medical professionals deliver mental health services effectively [11]. The TS-LDA model finds more interesting tweets than baseline techniques. Tweets are ranked automatically for a large audience based on their level of interest. Examines earlier research pertaining to the paper. Finding captivating tweets to engage a larger audience .The new TS-LDA model performs better in topic extraction than baseline techniques .Reorder the most re-tweeted tweets according to their level of interest [12]. Compared to LDA, MF-LDA has a higher coverage rate and less confusion . Average P, R, and F for the HTT algorithm are 85.64%, 84.97%, and 85.66%. MR and FR in the HTLCM model are less than 18%. Empirical use on a female driver event demonstrates the models' efficacy. focuses on tracking and topic extraction in micro-blogging [13]. It used measures for accuracy and Alpha-reliability to assess inter-annotator agreement .discusses relevant research on KE and NA solutions. assesses inter-annotator agreement with cutting-edge measures [14]. ROUGE-L values: 0.982 for forecasts and predictions, and 0.662 for pertinent material. It focuses on analysing financial news for applications that scan markets. Help investors identify pertinent financial news events. Improve the way financial news is analysed for forecasts and predictions. Potential uses for investment strategy extraction in the financial sector [15]. Singapore had five themes, while Hong Kong had twelve. While the subjects discussed shared similarities, Hong Kong's topics were more focused. special subjects related to listing and host management for airbnb users .Concentrate on leveraging latent Dirichlet allocation to glean information from Airbnb reviews.

compared Airbnb reviews from 93,571 Singaporeans and 185,695 Hong Kong people. 12 topics in Hong Kong and 5 subjects in Singapore were found. Derive insights from latent Dirichlet allocation in Airbnb reviews.
For suggestions from managers, compare Airbnb evaluations from Singapore and Hong Kong[16]. Not as bad as they thought the election campaign was. Compared to mainstream parties, extreme parties speak more aggressively. Twitter is not utilized for actual politics, but rather as an emotion thermometer. Not much study examines specific tweets after processing large data sets. Analysis of political discourse during the elections in Madrid is the main subject of the study. recognizes hostile language used in Twitter political discussions. demonstrates that radical parties speak more aggressively than moderate ones. Rather than being used for politics, Twitter is employed as a sentiment barometer [17]. In India, attitudes on the COVID-19 vaccination are positive. discovered unfavourable attitudes toward vaccine skepticism and reluctance. Diverse attitudes throughout the population show the importance of focused communication. Summary of research on social media by COVID-19. Instruments and methodology for sentiment analysis.

III. RESEARCH METHODOLOGY

1. Data collection :

One can use the robust features of the Python package snscrape to gather and save Twitter tweet data in a CSV file. Snscrape makes it simple to scrape tweets and extract information such as the timestamp, username, and content. The information can be arranged and written to a CSV file for easy storage and analysis when the desired data has been extracted. This simplified procedure makes it easier to manage Twitter data effectively, opening it up for additional investigation and use in a range of applications, such as trend monitoring and sentiment analysis. Crucially, snscrape streamlines the process of gathering and storing data, improving accessibility to insightful information on Twitter.

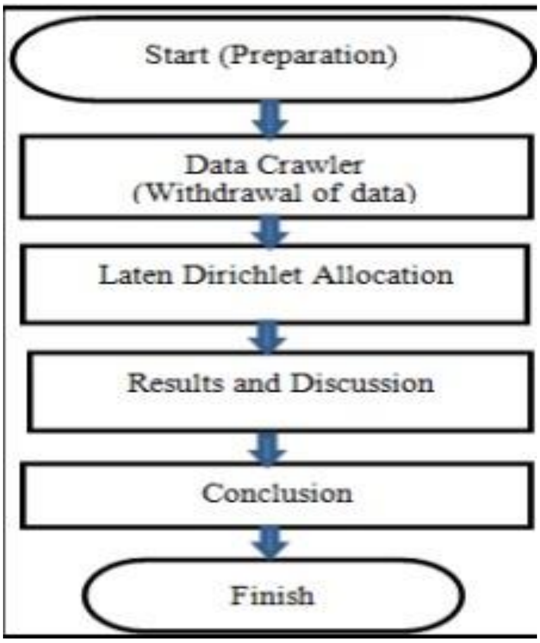

Fig. 1: Research Stage [8].

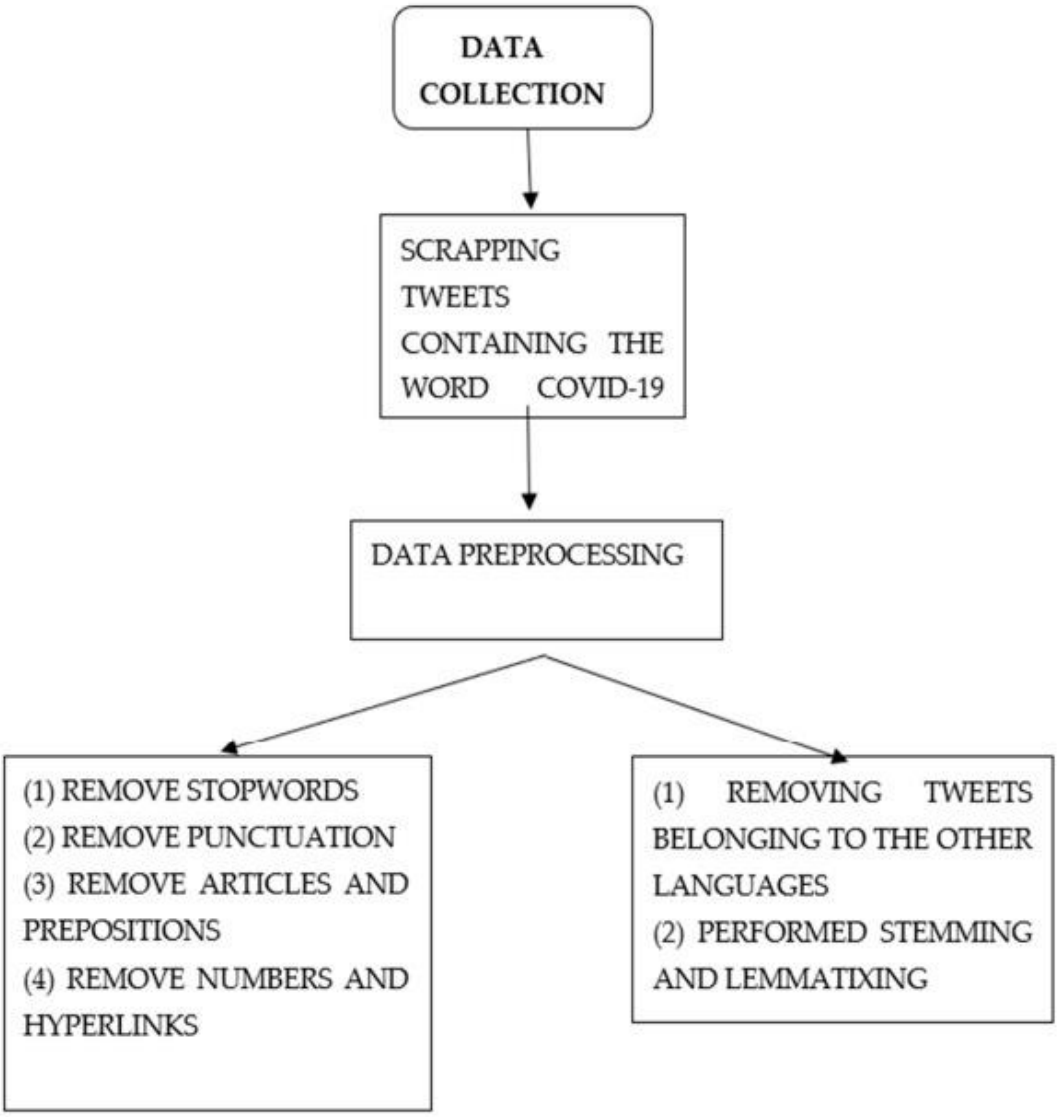
DATA COLLECTION
SCRAPPING TWEETS CONTAINING THE WORD COVID-19
DATA PREPROCESSING
(1) REMOVE STOPWORDS
(2) REMOVE PUNCTUATION
(3) REMOVE ARTICLES AND PREPOSITIONS
(4) REMOVE NUMBERS AND HYPERLINKS
(1) REMOVING TWEETS BELONGING TO THE OTHER LANGUAGES
(2) PERFORMED STEMMING AND LEMMATIXING

Fig. 2: Data Collection and Data Pre-processing [15].

2. Data pre-processing:

Preparing data for Twitter tweets requires a number of crucial processes.First data cleaning takes care of things like eliminating mentions, URLs, and special character.

The elimination of common words with minimal semantic meaning is known as stop words removal.

Stemming helps with consistency by breaking words down to their most basic form.

Lemmatization improves analysis accuracy by simplifying words to their dictionary or form.

By combining these methods, one can improve the original Twitter data, cut down on noise, and enable more insightful analysis. is pre-processing ensures more accurate and informative findings by streamlining following activities like sentiment analysis and topic modelling and improving the quality of the dataset.

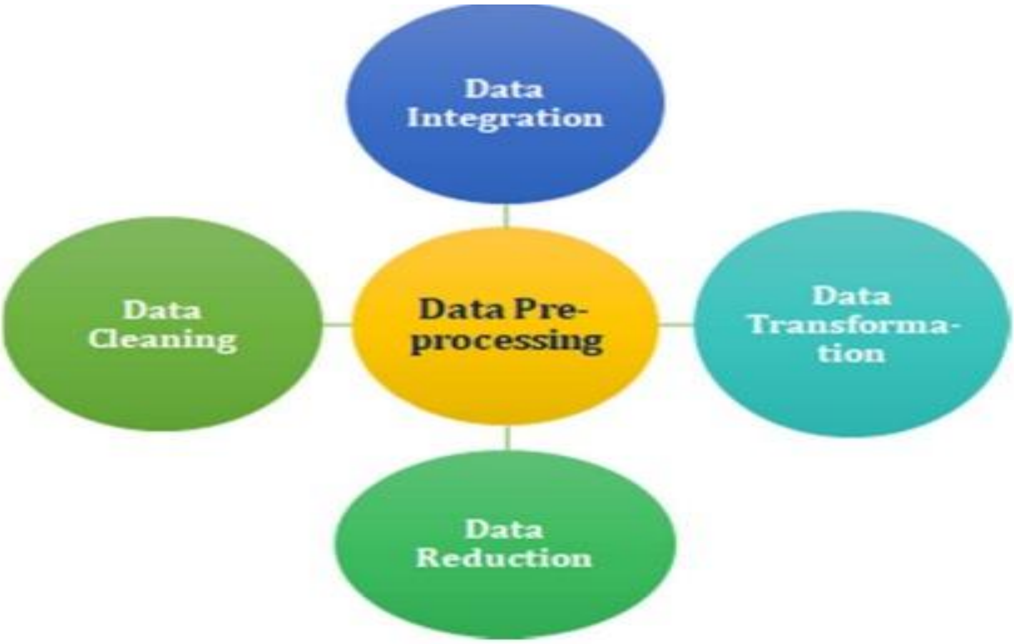

Fig. 3: Data Pre-processing Steps [6].

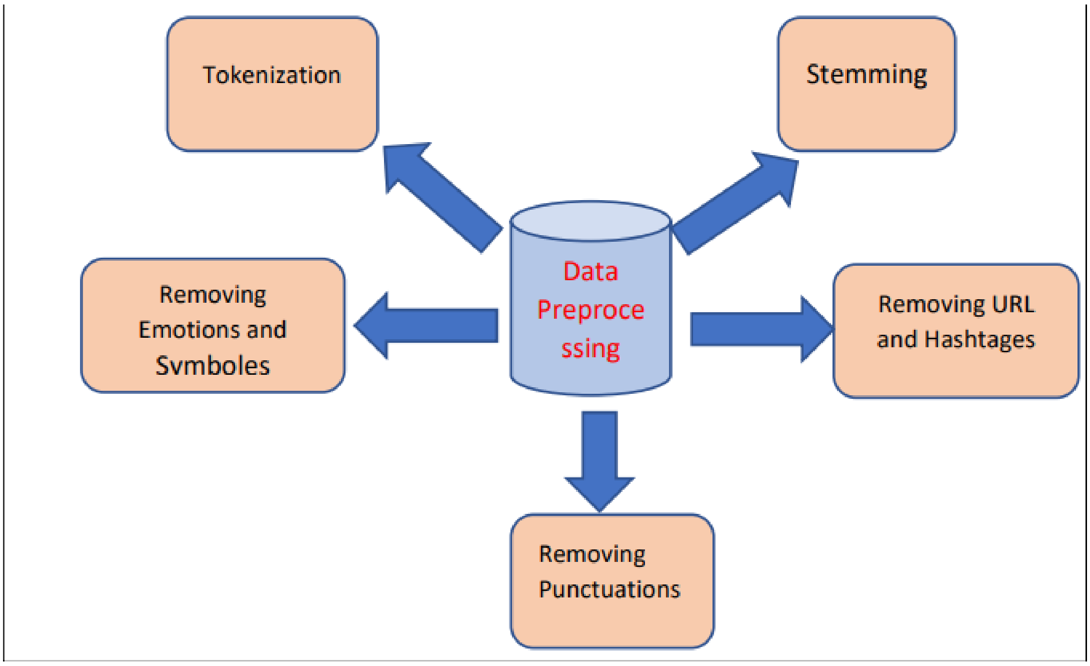

Fig. 4: Data Pre-processing Steps [1].

3. LDA method:

A probabilistic model called Latent Dirichlet Allocation (LDA) is used to find topics in huge text collections. The technique posits that documents are mixtures of latent subjects from which words are probabilistically derived. By evaluating the likelihood of words belonging to particular subjects and the distribution of topics across documents, LDA iteratively improves its topic assignments. As a result of this iterative procedure, coherent subjects and their frequency in the dataset are revealed. Document-topic and topic-word distributions are among the outputs of LDA that offer a thorough comprehension of latent themes in the corpus and facilitate perceptive analysis in domains like information retrieval and natural language processing.

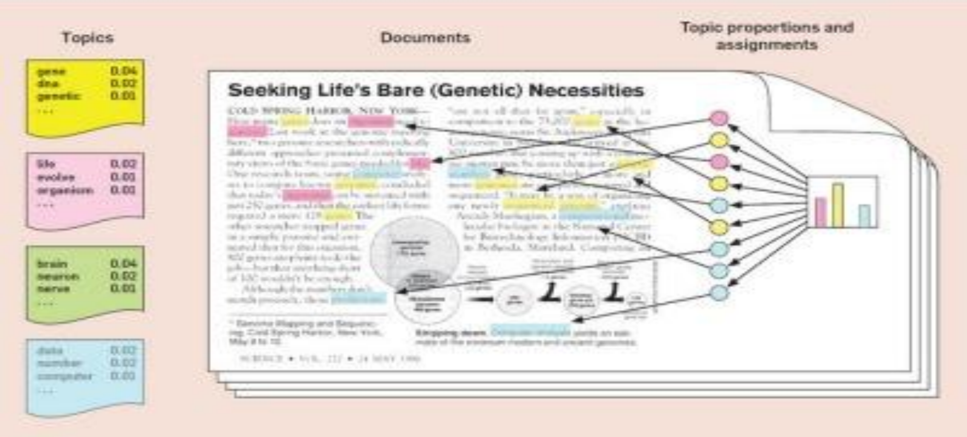

Fig. 5: Probabilistic Topic Model [8].

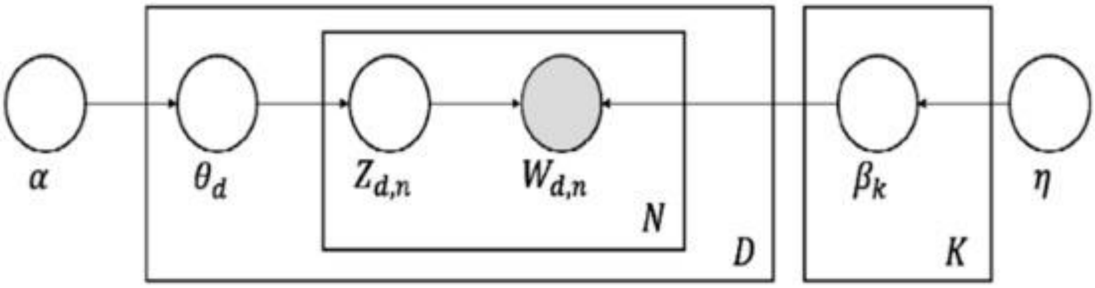

Fig. 6: Graphical Representation of LDA Model [15].

4. Bag-of-words:

Natural language processing (NLP) uses the Bag of Words (BoW) method, which disregards word order and grammar and portrays text as an unordered bundle of words. The complete corpus is used to generate a vocabulary, and each text is encoded as a numerical vector with each dimension representing a distinct word and its frequency represented by its value. BoW is useful for tasks like sentiment analysis and document categorization since it streamlines text analysis by concentrating on word occurrence. BoW serves as the foundation for more complex NLP models, allowing for effective and scalable text representation even when it loses contextual information.

5. Sentiment Analysis:

Sentiment analysis on Twitter data involves evaluating the emotional tone expressed in tweets, determining whether sentiments are positive, negative, or neutral. Using natural language processing techniques, algorithms analyse words, phrases, and context to discern user opinions. This process aids businesses, researchers, and policymakers in gauging public sentiments, monitoring brand perception, and identifying emerging trends. The dynamic and real-time nature of Twitter makes sentiment analysis a valuable tool for understanding the collective mood, allowing swift response to customer feedback, crisis management, and informed decision-making based on the prevailing sentiments within the Twitter verse.

In the above steps ,first twitter data is collected using snscrape module of python and the pre-processed cleaned ,stemming, lemmatization , stop words removal, LDA method is applied and then performed sentiment analysis.

IV. DISCUSSION AND CONCLUSION

The economy, politics, and viral transmission were the main subjects of conversation on Twitter. LDA algorithms aided in locating hidden subjects within posts on social media. AI-powered interviewing system lessens subjectivity-related talent loss. When it comes to topic modelling accuracy, LDA beats LSI approach. TS-LDA model exceeds baseline methods in terms of finding noteworthy tweets. Reordering the most re-tweeted tweets determines the rating of interesting tweet. Reordering the most re-tweeted tweets determines the rating of interesting tweets. The themes that emerged from the public's responses were "Understanding," "Action planning," "Hope," and "Reassurance." There was a clustering of public responses around "Understanding," "Action planning," "Hope," and "Reassurance." Topics should transition from self-centred to community-centred in 4.5 weeks. Public sentiment and opinions can be effectively extracted through social media text mining. Allows decision-makers to react quickly to abnormal conduct. Text mining on social media effectively gathers thoughts and responses from the public.

Facilitates prompt responses from policymakers to abnormal conduct. Twitter is utilized less for actual politics and more as a sentiment barometer. At last, LDA in sentiment analysis on Twitter provides benefits by revealing subjects that are concealed within tweets. It improves comprehension of a range of topics. Because it takes into account the larger context of conversations, this approach promotes a more accurate sentiment analysis and helps researchers and businesses understand public opinion at a more detailed level.

Future works include, introducing more enhanced machine learning methods to increase the accuracy of the model and this model should also include languages other than English like Arabic etc.

V. REFERENCES

[1]. Alsadan, N. M. (2023). Sentiment analysis and trend detection of tweets using machine learning techniques. International Journal of Computer Applications, 184(51), 1-6.

[2]. Du, Y., Yi, Y., Li, X., Chen, X., Fan, Y., & Su, F. (2020). Extracting and tracking hot topics of micro-blogs based on improved latent Dirichlet allocation. Engineering Applications of Artificial Intelligence, 87, 103279.

[3]. Du, Y., Yi, Y., Li, X., Chen, X., Fan, Y., & Su, F. (2020). Extracting and tracking hot topics of micro-blogs based on improved latent Dirichlet allocation. Engineering Applications of Artificial Intelligence, 87, 103279

[4]. García-Méndez, S., De Arriba-Pérez, F., Barros-Vila, A., González-Castaño, F. J., & Costa-Montenegro, E. (2023). Automatic detection of relevant information, predictions and forecasts in financial news through topic modelling with latent Dirichlet allocation. Applied Intelligence, 53(16), 19610-19628.

[5].García-Méndez, S., De Arriba-Pérez, F., Barros-Vila, A., González-Castaño, F. J., & Costa-Montenegro, E. (2023). Automatic detection of relevant information, predictions and forecasts in financial news through topic modelling with latent Dirichlet allocation. Applied Intelligence, 53(16), 19610-19628

[6]. Hamed, S. (2021). Real-time sentiment analysis of Twitter data. International Journal for Research in Applied Science and Engineering Technology, 9(5), 856-862.

[7]. Kiatkawsin, K., Sutherland, I., & Kim, J. (2020). A comparative automated text analysis of Airbnb reviews in Hong Kong and Singapore using latent Dirichlet allocation. Sustainability, 12(16), 6673.

[8]. Negara, E. S., Triadi, D., & Andryani, R. (2019). Topic modelling Twitter data with latent Dirichlet allocation method. 2019 International Conference on Electrical Engineering and Computer Science (ICECOS).

[9]. Shurrab, S., Shannak, Y., Almshnanah, A., Khazaleh, H., & Najadat, H. (2021). Attitudes evaluation toward COVID-19 pandemic: An application of Twitter sentiment analysis and latent Dirichlet allocation. 2021 12th International Conference on Information and Communication Systems (ICICS).

[10]. SV, P., Lorenz, J. M., Ittamalla, R., Dhama, K., Chakraborty, C., Kumar, D. V., & Mohan, T. (2022). Twitter-based sentiment analysis and topic modeling of social media posts using natural language processing, to understand people's perspectives regarding COVID-19 booster vaccine shots in India: Crucial to expanding vaccination coverage. Vaccines, 10(11), 1929. [11]. Tampus- Siena, M. (2022). Analyzing the discussion of Gregorio murder on Twitter using text mining approach. Computers in Human Behavior Reports, 8, 100248. [12]. Vishwakarma, A., & Chugh, M. (2023). COVID-19 vaccination perception and outcome: Society sentiment analysis on Twitter data in India. Social Network Analysis and Mining, 13(1).

[13]. Yang, M., & Rim, H. (2014). Identifying interesting Twitter contents using topical analysis. Expert Systems with Applications, 41(9), 4330-4336

[14]. Yang, M., & Rim, H. (2014). Identifying interesting Twitter contents using topical analysis. Expert Systems with Applications, 41(9), 4330-4336

[15]. Shurrab, S., Shannak, Y., Almshnanah, A., Khazaleh, H., & Najadat, H. (2021). Attitudes evaluation toward COVID-19 pandemic: An application of Twitter sentiment analysis and latent Dirichlet allocation. 2021 12th International Conference on Information and Communication Systems (ICICS

[16]. Hamed, S. (2021). Real-time sentiment analysis of Twitter data. International Journal for Research in Applied Science and Engineering Technology, 9(5), 856-862. [17]. Du, Y., Yi, Y., Li, X., Chen, X., Fan, Y., & Su, F. (2020). Extracting and tracking hot topics of micro-blogs based on improved latent Dirichlet allocation. Engineering Applications of Artificial Intelligence, 87, 103279.

[18]. Negara, E. S., Triadi, D., & Andryani, R. (2019). Topic modelling Twitter data with latent Dirichlet allocation method. 2019 International Conference on Electrical Engineering and Computer Science (ICECOS).