ANALYSIS ON SOCIAL MEDIA-A CASE STUDY

Abstract

As social networks gain more and more traction, the study of social network analysis has grown in significance and interest. A method for analyzing social systems using networks and graphs is called social network analysis. Since social network information is unstructured, it is essential to extract the organized information in order to make use of the crucial information. The research focused on social network data extraction enables the systematic utilization of such a vast volume of dispersed, unstructured data. To improve the accuracy of seeing the structured data scattered across the social network, natural language processing is used. The main goal of monitoring is to glean important information from texts submitted by uninformed social media users. It analyses natural language text to obtain information about a variety of entities, connections, or occurrences. The Natural Language methods are being thoroughly investigated through this research. In this study article, text mining systems—the cornerstone of Natural Language Processing—are used to assess data from social networks and investigate sentiment analysis processing through machine learning.

Keywords: Social media network, text mining, classification, sentiment analysis.

Authors

Pankaj Kumar Jain
Assistant Professor,
CSE, GGCT
pankajjain@ggct.co.in

Shailendra Chourasia
Assistant Professor,
CSE, GGCT
shailendrachourasia@ggits.org

I. INTRODUCTION

A phenomenon that has attracted billions of users has emerged as a result of the recent globalization of social media. Through social networking sites, individuals can create their own content to share in a variety of media, including information, private words, images, audio, and videos. As a result, social computing is recognised as a growing area of research and study that incorporates a wide range of topics, such as Web semantic, artificial intelligence, natural language processing, network analysis, and big data analytics [1].

Online word-of-mouth is influencing people's opinions and choices more and more these days, which has caught the attention of many. Social media platforms can be broadly categorized as content-sharing websites, forums, blogs, and microblogs. Through social media platforms like Facebook, Instagram, Foursquare, Flickr, and YouTube, people can exchange information, messages, images, videos, and other types of content. Web user forums (like StackOverflow, CNET forums, and Apple Support) are places where people post specialised information, questions, or solutions. On blogs (like Gizmodo, Mashable, Boing Boing, and many others), people can post messages and other content and share knowledge and ideas. Only brief texts for news and opinions can be shared on microblogs (like Twitter, Sina Weibo, and Tumblr).7].

The use of social media has changed greatly in recent years. Ideas can be exchanged between people with the touch of a mouse. One of the most popular social media platforms today is Twitter. Some of the subjects get a lot of interest. Some of these pieces of information circulated around a large group of individuals, and this process is referred to as "going viral." A company must examine these tendencies and then take the required actions to advance the company. The research community places a lot of importance on social media trends.

Table 1: Platforms for social media and their features

Type	Characteristics	Examples
Social Network	In order to exchange user-generated content, a user can create a web page on a social networking website and connect with a friend or other acquaintance.	MySpace, Facebook, LinkedIn, Meetup, Google Plus+
Blog Comments	An online journal known as a blog allows its author to post content and have it displayed in reverse chronological order. Blogs are typically run by an individual or group. User comments on blogs or online newspaper articles are referred to as blog comments.	Online publications, Huffington Post, Business Insider, and Engadget
Microblogs	Similar to a blog, but with less content, is a microblog..	Twitter, Tumblr, Sina Weibo, Plurk
Social News	Social news encourages its audience to share and vote on content as well as submit news stories.	Digg, Slashdot, Reddit
Media Sharing	A website that lets users upload films and photos for sharing with others.	YouTube, Flickr, Snapchat, Instagram, Vine

Social networking has become a global phenomenon that has changed how people interact and communicate. Practically every aspect of our lives is influenced, including communication, politics, healthcare, and social connections. sharing online with people who have similar interests your blog entries, emails, photos, and videos. Also, it empowers individuals to aggressively express their ideas or viewpoints. [3].

A subfield of artificial intelligence in computer science is called natural language processing. The main focus of natural language processing is communication between humans and computers. Both Natural Language Processing and Natural Language Understanding are components of NLP.

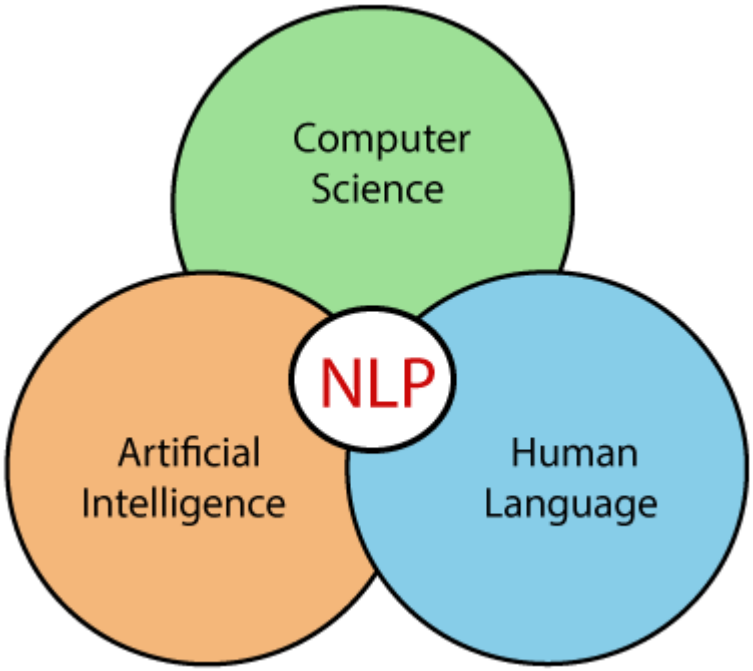

Figure 1: NLP as hybrid concept

This paper is broken up into five sections: section 1 offers an introduction, section 2 explores a topic related to natural language processing, section 3 explores open-source NLP libraries and challenges they face when used with social networks, section 4 offers text mining techniques, and section 5 offers a summary of the essay.

II. NATURAL LANGUAGE PROCESSING (NLP)

On social media engagement platforms, there are many alternatives. Text posts are one of the most important. Over the past 25 years, the natural language processing (NLP) of traditional media, such as printed news and articles, has gained popularity as a research area. Using expertise from computer science, artificial intelligence, and linguistics, NLP usually allows computers to extract meaning from natural language input.

NLP for social media text is a new study area that calls for either modifying existing NLP techniques to these types of texts or inventing whole new techniques appropriate for information extraction and other social media-related tasks. Natural language processing is the practise of manipulating natural language by a computer (NLP).

Any written or spoken language that has human origins is considered to be a natural language. [Wik21] gives the following definition of NLP: In order to study how computers and human language interact, the field of "natural language processing," or NLP, combines linguistics, computer science, and artificial intelligence. Its main goal is to develop methods for teaching computers to process and analyse enormous amounts of natural language data. As a result, a computer can now "understand" the contents of documents, including the subtleties of their language when applied in particular situations. It might be simple to compare different writing styles by counting word frequencies at one extreme. On the other hand, NLP requires "understanding" complete human statements, at least to the point of being able to respond, according to [BKL09].

The first stage in social network data collection is identifying issues and opportunities. The information may be kept as text or it may be dynamically collected online and analysed in real time to satisfy specific requirements. The next stage is the SASM pipeline, which consists of specific NLP tools for data processing and social media analysis. Social media data are large, noisy, unstructured datasets. SASM transforms social media data into messages that are relevant and understandable using social information and knowledge. SASM then analyses the data from social media to produce social media intelligence. Social media intelligence can be made available to users or provided to decision-makers in order to foster better awareness, communication, and decision-making.

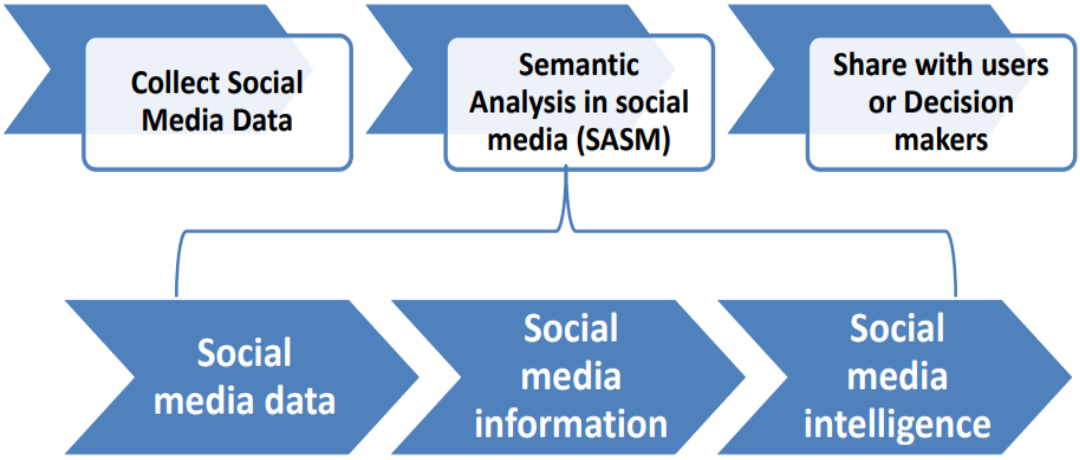

Figure 2: A framework for semantic analysis in social media, where NLP tools transform the data into intelligence

Data visualisation techniques could be used to complete the presentation of the SASM-analyzed data. These techniques have been developed in a variety of scientific fields, including statistics, computer science, linguistics, and library science. Text mining approaches deal with unstructured text and concentrate on automated analysis of textual data as a kind of natural language. Despite the lack of a universal definition for text mining, the

general method of analysis is accepted. Text mining is also linked to Natural Language Processing (NLP), which is concerned with the study of natural languages. Software options are accessible for analysing social media applications due to the requirement of applying automatic techniques for textual data analysis and extracting pertinent information. In order to understand people's reactions and behaviours, text mining technologies are used to find and analyse posts, likes, and followers in online social networks. Additionally, it demonstrates the variety of viewpoints and ideas on various subjects. Data gathering, pre-processing, content analysis, discovery, and integration are all essential steps in the text mining process.

Ambiguity and Uncertainty in Language

Ambiguity, a concept widely used in natural language processing, is the ability to have multiple possible interpretations. To put it simply, uncertainty has the ability to be understood in multiple ways. Language is quite ambiguous in general. In NLP, there are the following categories of ambiguity:

Lexical Ambiguity: Lexical ambiguity refers to a word's lack of clarity. For instance, using the word "silver" as a noun, adjective, or verb.

Syntactic Ambiguity: When a sentence is parsed in many ways, this form of ambiguity develops. Take the phrase "The man saw the girl with the telescope," for instance. It's unclear if the man saw the girl carrying a telescope in person or if he looked through his telescope to see her.

Semantic Ambiguity: This kind of ambiguity exists when the words themselves have potential for misinterpretation. In other words, semantic ambiguity happens when a statement employs an ambiguous word or phrase. Due to its semantic ambiguity, the phrase "The car hit the pole while it was moving" can be understood as either "The car hit the pole while it was moving" or "The car hit the pole while the pole was moving."

Anaphoric Ambiguity: This kind of ambiguity arises due to the use of anaphora entities in discourse. For example, the horse ran up the hill. It was very steep. It soon got tired. Here, the anaphoric reference of “it” in two situations cause ambiguity.

Pragmatic Ambiguity: This form of ambiguity occurs when a sentence might have several meanings depending on its context. Pragmatic ambiguity, to put it simply, is when the statement is not specific. For instance, the phrase "I like you too" might be interpreted in a variety of ways, such as I like you (just as you like me) or I like you (just as someone else does).

1. Automatic summarization, chunking, part-of-speech tagging, named entity recognition, named entity disambiguation, fact/relation extraction, word meaning disambiguation, and sentiment analysis are NLP techniques that are crucial for social network monitoring.

   Automatic Synthesis With the use of a computer programme, a text document is condensed through the process of automatic summarization, which results in a summary that keeps the key ideas from the original text.

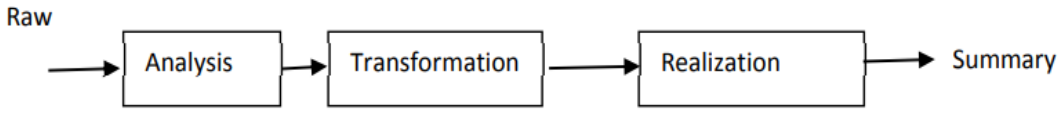

Figure 3: Process of Auto Summarization

2. **Chunking:** Chunking is the primary entity detection technique. Chunking selects a subset of the tokens as opposed to tokenization, which omits whitespaces. Tokenization prevents the created fragments from the original text from overlapping. A form of software called parts-of-speech tagging examines text in a language and identifies each word's part of speech, such as a noun, verb, or adjective, to name a few. Typically, computational applications use more exact One of the tags used to identify parts of speech is "noun-plural." Depending on the category or categories that dictionaries assign to a given term, a word may come under more than one category. Run, for example, can be used as both a verb and a noun. In order to clear up this uncertainty, taggers employ "Probabilistic Information."Identification of Named Entities Finding and categorizing Named Entities in text into pre-defined categories, such as the names of people, organizations, locations, expressions of times, quantities, monetary values, percentages, etc., is the goal of recognition, a subtask of information extraction.

3. **Named Entity:** The process of tying together the identities of entities that are mentioned in the text is known as named entity disambiguation. Although it recognizes names' references rather than their occurrence, it differs from named entity extraction in this regard. It is vital to have a database of the things that names can be related to.

4. **Relation Extraction:** The facts or relationships that exist between the specified entities in the required categories can then be extracted after we have identified the named entities in a text. Finding and separating the semantic connections between things in text or relations is the aim of fact extraction. The entities are then used to fill out a specified template.

   In NLP, stemming and lemmatization are two often used approaches. Both normalize words, but do so in various ways.

   a. By stemming, a word is truncated to its stem word. For instance, "friend" will be substituted for phrases like "friends," "friendship," and "friendships." Stemming might not provide us with a grammatical, dictionary word for a specific set of words.

   b. Lemmatization, as opposed to stemming, discovers the dictionary word rather than truncating the original word. Lemmatization algorithms frequently need a dictionary of the language to be able to accurately classify each word since they extract the correct lemma from each word.

5. Word disambiguation based on sense when a word has more than one meaning, this open NLP and ontology topic determines the appropriate sense of the word in the phrase. Depending on their existing knowledge of the subject, a person may rapidly understand the meaning of a phrase. The aspect of the term is difficult for a machine to recognize, though. The text's word ambiguities can be reduced using this way. For illustration: Word Net is a comprehensive English lexical database that is free to use.

6. Sentiment analysis is a natural language processing (NLP) technique that recognises, extracts, and counts the user's sentiment towards the content presented.

III. OPEN-SOURCE NLP LIBRARIES

The algorithmic foundation of NLP in practical applications is provided by NLP libraries. To build up or deploy servers and infrastructure, it offers a free API.

- **Apache OpenNLP:** This machine learning toolkit is open source and offers natural language text. It offers several services, including part-of-speech tagging, named entity extraction, summarization, searching, information grouping, tokenization, natural language production, feedback analysis, and more. It offers a command-line interface where models can be trained and tested with some predefined models.
- **Natural Language Toolkit (NLTK):** This well-known Python toolkit offers modules for parsing, categorising, tokenizing, stemming, sematic reasoning, and other text processing tasks. It offers friendly user interfaces for more than 50 corpora and lexical resources, including WordNet.
- Stanford NLP is a collection of NLP tools that includes named entity recognition, sentiment analysis, part-of-speech tagging, and more. It offers tools for statistical NLP, deep learning NLP, and rule-based NLP, all of which are widely utilised in business, academia, and government.
- **MALLET:** This Java programme offers Latent Dirichlet Allocation, document classification, clustering, topic modelling, information extraction, and more features.
- Genism-gensim is a powerful library for semantic modelling that may be used in a variety of applications.
- Stanford's CoreNLP A variety of human language technology technologies make up CoreNLP. It seeks to make it simple and effective to apply linguistic analysis methods to a document.
- SpaCy is a Python-based open-source NLP library. It allows you to create applications that process and comprehend massive amounts of text because it is specifically intended for use in production environments.

Challenges in NLP

Informal Language: On social networking sites, people post messages that are noisy, casual, and that lack punctuation, spelling errors, capitalization, and grammatically accurate sentences.

- Part-Of-Speech tags make it harder to extract information from social networks.
- **Brief Contexts:** Social networks, like Twitter, have minimal duration requirements. As a result, the user utilises more acronyms in their posts to provide more detailed information. Due to the posts' briefness and the co-references among the feeds, it is challenging to resolve specified entities.
- **Noisy Scant Contents:** Social network posts by users don't always contain pertinent information. Filtering is needed as a preliminary processing step to clean up the input postings stream.
- **Information About Entities:** As people typically utilise social networks to express information about their everyday activities, events, or happenings, the knowledge base does not contain information about entities. The methods used for information

extraction connect the entities involved in the retrieved information to a knowledge base. A new suit for information extraction from social network posts is required.

- **Unknown Contents:** Not every information on social networks can be relied upon. Contributions from users often contain information that is inconsistent with reliable outside sources. It is challenging to deal with the uncertainty in the retrieved relations and facts.

IV. CASE STUDY NLP BASED SOCIAL MEDIA ANALYSIS

The methodology has been developed based on the objective of the study, that is, to explore the recent text mining techniques applied in social networks analysis. The study is to explore the following aspects:

- The text mining techniques that most used in social networking.
- The text mining algorithms that most applied in social networking.
- The tools used to perform the opinion mining / sentiment analysis.

1. **Data Collection:** This is the procedure for methodically acquiring and analysing data, which then makes it possible to respond to pertinent inquiries and assess results. It addresses the issue that outdated information can be searched for a few days and the earlier versions cannot be located.

2. We used the Twitter API to accumulate tweets. Appropriate hashtag (#) can be used to filter and gather the specific class of tweets from the enormous collection. These hashtags were elected after investigating the top trendingtags from online assets and Twitter itself (Table-2).

Table 2: Trending Hashtag on Tweeter

Twitter Hashtags	
#NewEducationPolicy2020	#Facebook
#Socialmediamarketing	#Onlinebusiness
#digitalmarketing	#love
#socialmedia	#photography
#marketing	#startup
#branding	#ecommerce
#business	#motivation
#marketingdigital	#success
#seo	#webdevelopment
#onlinemarketing	#fashion
#contentmarketing	#content
#instagram	#facebookmarketing

3. **Pre-Processing:** This step refers to the processing of raw data to deliver a podium for data analysis. Preparing text data for future processing is the goal of the pre-processing stage. Pre-processing the data is a crucial step because it determines how effectively the subsequent processes will work. The tweets must be syntactically corrected as needed.
   - Eliminate Unicode strings and noise

- Swapping URLs and user mentions
- Convert the text to lower
- Replacing slang and abbreviations
- Removal of Special symbols
- Removing numbers
- Swapping negations with antonyms
- Eliminating punctuation
- Eliminating stop words
- Removal of web links
- Spelling correction

The following steps are carried out during text pre-processing: • Tokenization, which involves dividing the text into sentences and the sentences into words. We delete the punctuation and lowercase the words.

Tokenization is the process of fragmenting a text into smaller components known as tokens. Tokens can be letters, numbers, or commas. Another name for it is word segmentation.

Beds and chairs are examples of certain sorts of furniture.

The packages and the details of their installation are as follows −sent_tokenize package, word_tokenize package, WordPunctTokenizer package [11].

Stemming: For grammatical reasons, language has many varieties. variations in the sense that both English and other languages have a variety of word forms. Examples include democratic reform, democracy, and similar concepts. It is essential for machine learning projects that the machines recognise how many different words, like the ones stated above, have the same fundamental structure. As a result, when performing text analysis, it is very beneficial to remove the words' root forms from the text. The several stemming packages that are provided by the NLTK module include PorterStemmer, LancasterStemmer, and SnowballStemmer.

Lemmatization applies a morphological analysis to words by taking into account a language's entire lexicon in addition to word reduction. For instance, third-person terms are transformed to first-person.

- Normalization is the process of turning a list of words into a more standardised order. This is helpful when getting text ready for processing later. Other operations can interact with the data and won't have to deal with problems that might jeopardise the process by converting the words to a standard format. For instance, making all terms lowercase will make searching easier.
- Word embeddings, also known as vectorization, is a technique used in natural language processing (NLP) to convert words or phrases from a lexicon into a corresponding vector of real numbers. This vector is then used to identify word

predictions, word similarity, and semantics. Vectorization is the process of turning words into numbers.C). Data Analysis

A feature is a sizeable body of information that can be interpreted as a particular that aids in problem-solving. Excellence and magnitude are important qualities since they have an emotional impact on the results produced by the selected model. Feature extraction is an important phase in the sentiment classification process since it involves extracting relevant information from the text data and directly affects the model's performance. Feature extraction increases the accuracy of learned models by using features derived from the input data. The process of extracting features from tweets is the selection of useful terms [9].

- Unigram features: Each word is individually assessed to see if it qualifies as a feature.
- N-gram features, which consider several words simultaneously.

Usage of a group of words with specific positive or negative meanings is known as using an external lexicon.

Wordcloud: A Word Cloud is an interactive visualisation that dynamically changes the size and colour of words based on how frequently they are used in social posts in your social wall feed [8].

Figure 4: Word cloud of social media
(Ref.https://www.flickr.com/photos/bitsfrombytes/45219986531)

In order to draw a conclusion, we can put into practise eight elements derived from the social cognition theory as intangible features causing user behaviour, such as trust and distrust, belief and expectation, benefit and risks, positive and negative social impact, influencing study, policy and execution, ethics, trends, etc.

4. **Sentiment Analysis Tools:** By using classification, it is possible for different statistical categories to be unintentionally labelled into multiple groupings depending on certain characteristics of the data. The basic approach is to assess the emotional intent of the sentence and identify the characteristics of the product using word frequency statistics. By creating a classifier using labelled training data, we proposed a novel way for categorising eight features from the user's behaviour aspect. For these crucial components, our recommended strategy is novel [22].

First, a sentiment categorization based on the analysis of public conduct views is established.

Second, we relate the technology of natural language process to a novel perceptual wisdom and grouping.

Important proof for our approach comes from the ability to identify emotions based on user actions by fusing machine learning with a lexical vocabulary. Many characteristics can be utilised to predict a person's behaviour, as was already mentioned.

In order to analyse user behaviour, we place a strong emphasis on categorising a notion that may be utilised to represent user intentions and feelings surrounding e-word. In user behaviour analysis, we may also think of this as an emotional multi-classification problem. The majority of earlier research on emotion grouping concentrated on positive and negative classification, or binary classification. Nonetheless, it is frequently more appropriate to announce additional specific supporting data regarding various classification methods. Because there are multiple elements that can influence whether users accept a new technology or not, these influencing factors are broken down into different categories. Eventually, a meta-classifier that combines the predictions of various classifiers into a final classification result can be created using a variety of classification techniques.

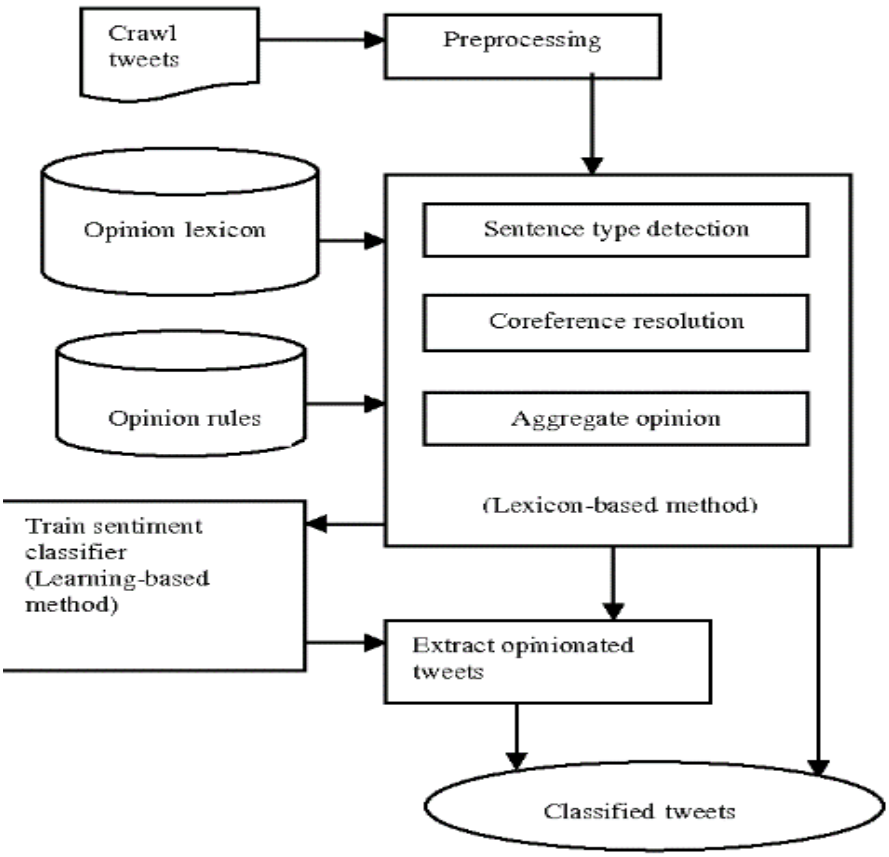

Figure 5: Architecture of sentimental analysis

The development of sentiment analysis can be improved by employing a hybrid technique. The programming language chosen for usage while putting this tactic to the test is Python. After pre-processing, the proposed algorithm consists of three steps. The first step, which is concerned with determining the optimum operating constraints for the classifier, deals with the lexicon-based model. On the other hand, in the second section, the learning-based paradigm is discussed and its implementation is reviewed. It is mentioned as the hybrid model to be examined in the third and final section.

To determine the polarity of a tweet, the lexicon-based method makes use of pre-built lexicon assets that contain divergence of sentiment arguments, such as SentiWordNet (SWN) 3.0 [10]. The lexicon-based method leverages pre-built lexicon assets containing divergence of sentiment arguments, such as SentiWordNet (SWN) 3.0 [10], to formulate a tweet's polarity.

To determine the polarity of a tweet, the lexicon-based method makes use of pre-built lexicon assets that contain divergence of sentiment arguments, such as SentiWordNet (SWN) 3.0 [10]. The lexicon-based method leverages pre-built lexicon assets containing divergence of sentiment arguments, such as SentiWordNet (SWN) 3.0 [10], to formulate a tweet's polarity.

In the combined lexicon-based and machine learning approach, the outcome from the lexicon-based method was used in the machine learning method as a training data to train the two classifiers (Naïve Bayes, SVM classifiers). Then, the same testing data (unlabelled data) was used to inspect the effectiveness of the two classifiers which provided Consequence.

The features of achieving foundational literacy, elasticity, multidisciplinarity, universal acceptance, education, importance of positive belief, understanding, inspiration, and life-threatening thinking, integrities and human & legal values, multilingualism, widespread use of technology, and topographical challenges can all be analysed in different ways [8].

The accuracy score evaluates how well a SA model predicts sentiment discussion after applying a number of machine learning classifiers (Naive Bayes, SVM, DT, and ME). By dividing the total number of events by the true positive and true negative rates, it is calculated. The results confirmed that applying feature selection techniques leads to improved performance.

V. CONCLUSION AND PERSPECTIVES

User-generated text content and text-based exchanges on social networking platforms frequently include a diversity of writing styles, subjects, and languages in addition to errors, freeform emojis, and abbreviations. Due to the variety of content and language, using social media data for NLP tasks is more challenging. Some of the cleaning techniques listed, such as emoji removal, hyperlink removal, language detection and translation, and typo correction, have been demonstrated to be effective in priming and pre-processing language of this kind. When applied to text before other Natural Language Processing (NLP) approaches like named-entity identification and key phrase extraction, these techniques result in cleaner output.

In addition to text data, social media data also includes statistics that reflect how people interact with and are interested in various kinds of information. Inferences drawn from named-entity recognition (NER) and key phrase extraction NLP approaches, when combined with these statistics, have been proven to be useful for trend analysis, analytics, and identifying correlations and affinities between user involvement with social media.

Even though the current research's findings are reliable, more can be done. First, in the situation of cross-domain, the model's accuracy cannot be guaranteed. Second, training the model is challenging when the information is insufficient, and larger and more varied databases are required to increase the model's accuracy. Finally, category labels must be manually annotated, which involves heavy manual labour.

REFERENCES

[1] Anna Atefeh Farzindar, Diana Inkpen," Natural Language Processing for Social Media" Third Edition, Springer Nature Switzerland AG 2022, DOI 10.1007/978-3-031-02175-6

[2] Hiba A. Abu-Alsaad, Rana Riad K. AL-Taie," NLP analysis in social media using big data technologies", TELKOMNIKA Telecommunication, Computing, Electronics and Control Vol. 19, No. 6, December 2021, pp. 1840~1846

[3] Huda N Alhazmi, Text Mining in Online Social Networks: A Systematic Review ,IJCSNS International Journal of Computer Science and Network Security, VOL.22 No.3, March 2022

[4] Jyotika Singh, Social Media Analysis using Natural Language Processing Techniques, PROC. OF THE 20th PYTHON IN SCIENCE CONF. (SCIPY 2021)

[5] CharuVirmani , Dr. Anuradha Pillai , Dr. Dimple Juneja," Extracting Information from Social Network using NLP", International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 4 (2017), pp. 621-630,

[6] Glen Coppersmith, Ryan Leary, Patrick Crutchley and Alex Fine," Natural Language Processing of Social Media as Screening for Suicide Risk" Biomedical Informatics Insights Volume 10: 1–11 © The Author(s) 2018 Article reuse guidelines: sagepub.com/journals-permissions DOI: 10.1177/1178222618792860.

[7] Hui Li, Qi Chen , ZhaomanZhong , Rongrong Gong , Guokai Han ,E-word of mouth sentiment analysis for user behavior studies, Information Processing & Management Volume 59, Issue 1, January 2022, 102784, https://doi.org/10.1016/j.ipm.2021.102784

[8] H.S.Hota,DineshK.Sharma,NileshVerma(2021), Lexicon-based sentiment analysis using Twitter data: a case of COVID-19 outbreak in India and abroad, https://doi.org/10.1016/B978-0-12-824536-1.00015-0

[9] Gamal, D., M. Alfonse, E.-S. M. El-Horbaty and A.-B. M. J. P. C. S. Salem (2019). "Implementation of Machine Learning Algorithms in Arabic Sentiment Analysis Using N-Gram Features." 154: 332-340

[10] A Comparison of Machine Learning Techniques for Sentiment Analysis,DOI: https://doi.org/ 10.17762/turcomat.v12i3.999.

[11] Aastha Tyagi, A Review Study of Natural Language Processing Techniques for Text Mining, International Journal of Engineering Research & Technology (IJERT), http://www.ijert.org ISSN: 2278-0181 , Vol. 10 Issue 09, September-2021

