JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

From Tweets to Trends: NLP -Driven Social Media Sentiment Analysis

[1]Sameera Jathar, [2] Bhavik Sachani, [3]Heet Kalaria, [4]Manasi Gohil,

[1]Student, [2]Student, [3]Student
[1]Department of Artificial Intelligence & Data Science,
[1]K J Somaiya Institute of Technology, Mumbai, India

Abstract: Natural Language Processing (NLP) is a disciplined of artificial intelligence that aims to understand and interpret human language in context. With the rise of social media, NLP can be used in social media analytics to enhance decision-making, analyses user sentiments, speed up data processing, and gain deeper knowledge of user behavior. This paper explores the application of NLP techniques in analyzing social media content to uncover trends, sentiment, and patterns.

Index Terms - **Natural Language Processing, sentiment analysis, social media analytics**

I. INTRODUCTION

Social media platforms have evolved into thriving hubs of information sharing, opinion expression, and sentiment revelation in the digital age. These social media platforms are constantly flooded with user-generated content that covers a wide range of human experiences, thoughts, and emotions. However, analyzing this amount of data is a difficult task. The sheer volume, unstructured nature, and linguistic oddities of social media data make manual analysis impossible and frequently wrong. Within this complex environment, the area of Natural Language Processing (NLP) emerges as a beacon of hope. Natural Language Processing is a branch of computer science, more particularly the field of Artificial Intelligence (AI) that is concerned with providing computers the ability to understand and interpret human language in a manner similar to that of humans. These technologies allow computers to process human language in the form of text or audio data, completely comprehend what is being said or written, and comprehend the intentions and sentiment of the speaker or writer. NLP powers algorithms that translate text between languages, react to spoken commands, and quickly evaluate massive amounts of information that is being generated [1] [2].

NLP offers an organized, automated method for extracting meaning and insights from this textual sea. NLP approaches cover a broad range of tasks, such as tokenization, sentiment analysis, topic modelling , and word embedding, among others, which are aimed at understanding the ambiguities and complexities of human language in digital form. One of the main techniques of NLP that is used for social media analytics is sentiment analysis. Sentiment analysis, commonly referred to as opinion mining, tries to identify and quantify the emotional tone or sentiment indicated in a text, such as a tweet, review, or news story. The main objective of sentiment analysis is to automatically categorize text as positive, negative, or neutral, offering insightful data on consumer feedback, public opinion, and brand perception. NLP is employed for sentiment analysis by first preprocessing textual data, including tasks like tokenization and removing punctuation. Sentiment lexicons or machine learning models are then used to assign sentiment labels (e.g., positive, negative, neutral) to text documents. These models learn patterns and associations between words and sentiments from labelled data during training.

Once trained, they classify the sentiment of new text data, enabling the automatic extraction of opinions or emotions. Along with social media analytics, sentiment analysis finds applications in customer feedback analysis, brand reputation management, and market research, providing valuable insights from textual information[3]. Because social media data is unique, NLP faces significant challenges in the domain of social media analytics. Firstly, social media information is typically noisy, full of slang, typos, and grammatical problems that provide challenges for NLP models trying to provide proper interpretation. Second, social media messages' sentiments can be surprisingly vague, including irony and sarcasm, confusing NLP algorithms that are supposed to classify them. Furthermore, real-time processing and analysis of the enormous volume of data generated across social media platforms necessitates highly sophisticated computing and resource skills. These complex problems highlight the difficulties NLP has when gathering knowledge from the complex and constantly changing world of social media data [4].

II. LITERATURE SURVEY

This book reviews the most up-to-date research on Natural Language Processing tools and techniques for extracting non-standard data from social media datasets that are readily available in large quantities (big data). It demonstrates how novel NLP approaches can incorporate relevant linguistic information into a range of industries, such as social media monitoring, healthcare, business intelligence, and industry. Additionally, it examines the evaluation criteria currently in use for social media applications and the most recent assessment campaigns and group projects that have been conducted with the use of new social media datasets. All of these activities are either co-ordinated by the Association for Computing Linguistics (ACL) or the NS&T (NS&T) through the text retrieval conference (TREC) or the text analysis conference (TAC).[5].

Social media has become a crucial aspect of our daily lives as a result of internet usage. The technique of recognizing and evaluating a piece of text to ascertain if its sentiment, ideas, perspectives, and feelings are good, negative, or neutral towards a certain problem, object, etc. using the Natural Language Tool Kit (NLTK). Social networking is becoming a need for people to keep in touch. The popular communication platform Twitter allows users to post their beliefs. People can post comments and brief remarks. An organization may examine Twitter sentiment to learn how people are talking about its image. There are multiple techniques for sentiment analysis, with diverse applications for various regions. Knowledge bases and machine learning are the two primary methods for assessing views. The tweets that were labeled in voting systems for this study's Twitter data were used. Tweets were processed beforehand using text mining. A vector space model was then built using the inverse document frequency and term frequency, and sentiment analysis was performed using the Random Forest Classifier, Decision Tree Classifier, and Logistic Regression techniques. Discussions of experiments lead to conclusions [6].

A.Sriteja, P. Pandey and V. Paudi presented a technique for identifying contentious news stories. This is accomplished by analyzing how individuals responded to news articles covering these topics on social media. Finding contentious news stories online is a current issue. It is especially helpful when discussing matters like a presidential election, governmental changes, climate change, etc. since it helps pinpoint the themes on which individuals have differing opinions. To do this work, we employ word matching and sentiment analysis. We demonstrate the use of our technique for identifying contentious issues during the 2016 US Presidential elections [7].

More and more people have started utilizing social media to sell their businesses in recent years. However, because social media platforms provide so much data, it may be challenging for businesses to properly assess and utilize this data. Natural language processing (NLP) techniques may be used to get insightful data for social media marketing by determining how people feel about postings on social media. In order to understand how people feel about things, this study article provides an outline of how NLP approaches are employed in social media marketing. The article discusses several NLP methods, including ones that rely on dictionaries, rules, and machine learning. The study also demonstrates a case study of how NLP may be used to social media posts to ascertain users' opinions about a particular brand or product. The case study's findings demonstrate that NLP is effective at determining sentiment and has the potential to significantly aid social media marketing. We discuss some of the drawbacks of utilizing NLP for sentiment analysis in social media marketing towards the conclusion of the study, as well as possible future directions [8].

From a business analytics viewpoint, it is critical to comprehend if consumers accept recently released mobile applications and to assess the elements that impact user acceptance given the growth of social media apps and the constant release of new apps. In order to solve this issue, this work proposes the first analytical model that combines natural language processing based on user evaluations with Technology Acceptance Model. The sentiment score for each review is obtained from processing user reviews in the app store with sentiment analysis and is used as quantitative data for the structural model. The user acceptability indicators are then classified using a TAM keyword vocabulary expanded by WordNet for each user review with a sentiment score and then processed into vectors of the same length in accordance with the time series. Finally, AMOS was used to perform regression estimates for the structural equation model. Five of the seven hypotheses were validated by the model's findings, while two were rejected. User attitude has a major impact on adoption among users. Additionally, consumers' sentiments are positively influenced by perceived utility and convenience of use. In addition, one of the elements affecting user adoption is contentment [9].

Many people in society display some type of prejudice or bias in their thoughts and communications, whether they do so consciously or unintentionally. Prejudice may be aimed against anything about which individuals might have an opinion and can be expressed in a variety of ways. The difference between prejudiced/biased thinking and having an opinion is that the latter may be erroneous because it is not necessarily founded on personal experiences. This essay outlines a method for analyzing a person's tweets on Twitter, one of the most widely used social media sites today, in order to spot any potential biases. A score of a user's possible bias is computed using sentiment analysis and other natural language processing technologies [10].

In recent years, identity fraud has become a major problem on online social networks. The goal of current research is to create technology that can identify identity fraud. The efficacy of the current tactics is debatable. We outline a study that uses clustering and classification approaches to identify identity theft. We outline conventional weaknesses in identity fraud detection for these techniques and make recommendations on how to improve their performance in practical settings. We first gather information from social media accounts and use preprocessing and filtering methods including Natural Language Processing (NLP), vectorization, dimensionality reduction, data standardization, etc. Based on the behavioral analysis and the traits of each profile, features are retrieved. To identify each profile, whether it is authentic or fraudulent, clustering techniques are applied[11].

The rapid expansion of information on social media made it necessary to assess user evaluations in order to understand the underlying feelings. This study suggests an assessment approach that incorporates aspect-related QoS factors gleaned from user feedback. In the pre-processing stage of our suggested model, tokenization, stemming, and stop-word removal are performed after review cleaning. A pre-processed set of word tokens is subjected to Stanford POS tagger's Parts of Speech (POS) tagging procedure[12]

III. METHODOLOGY

For sentiment analysis used for social media analytics, various NLP techniques are employed. Below discussed are the NLP steps and techniques used for social media sentiment analysis.

A. Data Collection

A dataset of social media posts from websites like Twitter, Facebook, and Instagram is needed for this study. To provide a thorough analysis, the dataset should include posts on a variety of subjects, such as politics, entertainment, and technology; etc.It must be ensured that there is availability of high-quality data. To get good results, the dataset must be labeled. Data can be uploaded using a live API, which enables you to get information that is readily accessible to the general public from websites like Amazon reviews,

Facebook, Twitter, or open-source data repositories like Kaggle. A .csv file can be used to manually upload data to the sentiment analysis API.

B. Data Preprocessing

Data preprocessing is an important step in natural language processing that involves cleaning and modifying unstructured text data to make it appropriate for analysis. Following are the preprocessing steps involved in NLP:

1. Text Cleaning: Special characters, URLs, and other non-alphanumeric characters should be removed from text.
2. Tokenization: The sentences are divided into smaller words called tokens. For example, the sentence 'He ate an apple' can be tokenized into 'He','ate','an','apple'.
3. Lower Casing: The words are converted into lower case. For example, 'BOY' is converted to 'boy'. Lower casing might sometimes lead to loss of information.For example, when working on any project that involves a person's emotions; using capital letters can indicate anger or excitement.
4. Stop Word Removal: Stopwords, or frequently used words, are omitted from the text because they don't contribute anything to the analysis. These phrases have little or no meaning. Some examples of stopwords include 'the','my','I','we','are',etc.
5. Lemmatization: Lemmatization is the process of breaking down a word into its root form for identification of similarities leading to better analysis. For example, the word 'studying' will be converted to 'study' after lemmatization.

C. Feature Extraction

Feature extraction is the process of transforming unstructured text data into a numerical representation that machine learning algorithms can comprehend and use. Because most machine learning algorithms require numerical input, this phase is essential. By extracting the necessary information from the text, feature extraction converts unstructured text data into a structured format.

1. Bag of Words (BoW)

The text is preprocessed using the Bag of Words model, which maintains track of the total number of times the most common terms are used in a text.

2. TF-IDF

TF-IDF (Term Frequency - Inverse Document Frequency) method examines the frequency of words to gauge how pertinent they are to a specific document.

3. N-grams

Contiguous groups of N items (often words) from a particular text are known as N-grams. They serve as features by capturing regional patterns. Unigrams (single words), bigrams (pairs of words), and trigrams (triplets of words) are popular options.

1. Part of Speech Tagging (PoS Tagging)
2. The process of classifying words in a text (corpus) in accordance with a particular part of speech, depending on the word's definition and context, is known as Parts of Speech (PoS) Tagging in natural language processing.
3. Word Embedding (Word2Vec, Glove, Fast Text).
4. A word is represented by a lower-dimensional numeric vector input called a word vector or word embedding. It enables the depiction of words with related meanings to be similar.

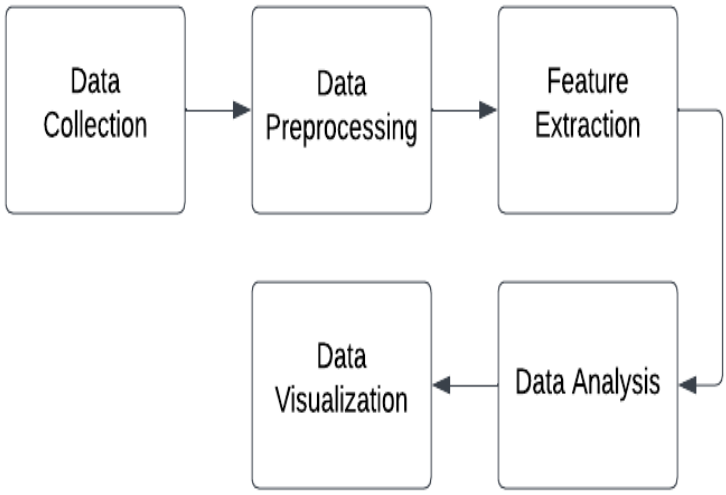

Fig. 1. Flowchart

IV. RESULT

The Twitter dataset available on Kaggle was used for this research. Analysis was done on this dataset using NLP techniques for preprocessing and various machine learning algorithms. The dataset used is a labeled dataset categorizing sentiments into three parts namely positive, negative and neutral. Other factors that were considered are the number of tweets, the time of the tweet (morning, noon or night), age group of the user, country of the user etc. The data was cleaned first by removing redundant data and handling the null or missing values. Further the data was preprocessed using various NLP techniques like tokenization, removing stop words and performing lemmatization to get the root words. The preprocessed data was further split into training and testing data for further analysis. The data was trained using various classification algorithms like Naive Bayes Classifier, Support Vector Machines and Random Forest algorithm. The same algorithms were used to test the testing dataset. The model

was efficiently able to predict the type of sentiment based on a given input. After analyzing the data, the results were visualized with the help of various libraries like seaborn and matplotlib.

Figure 2 is a column chart that shows the count of tweets according to the type of sentiment (count of positive, negative and neutral tweets). Column or bar charts are usually used for comparing data over a period of time. It is a graphical representation of categorical data.

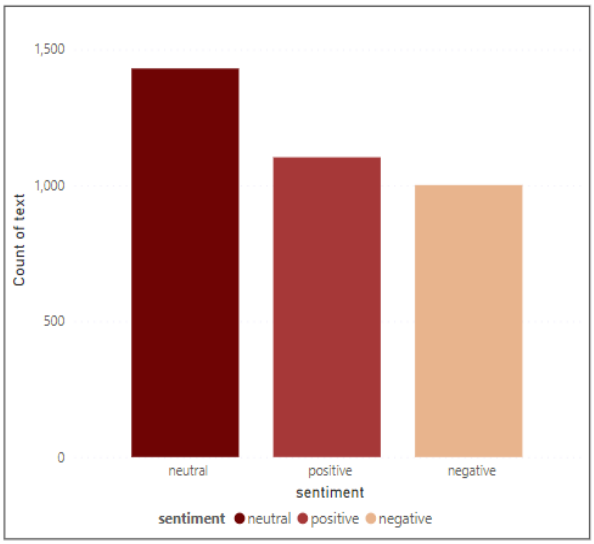

Fig. 2. Column Chart

Figure 3 is a doughnut chart that shows the relationship between the age group of the user and the sentiment of the tweet. These charts are used to represent the relationship between data of parts as a whole. They are popularly used to show the relational proportions between the data.

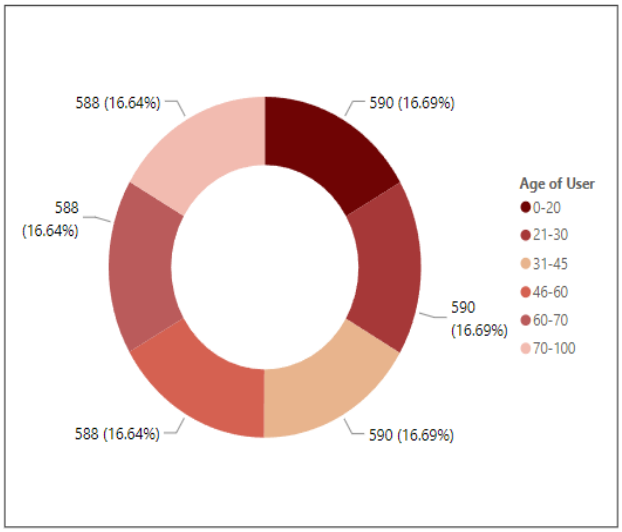

Fig. 3. Doughnut Chart

Figure 4 is a stacked column chart that shows the relationship between the type of sentiment(count of positive, negative and neutral) and the time of the tweet(morning, noon and night). Stacked column charts are used for comparing data over a period of time just as traditional column charts, but contain a composition of two or more variables.

All the machine learning models that we used were proven to be effective, but the performance and accuracy will completely depend on the type of dataset being analyzed. Support Vector Machines can be a good choice in case of high dimensional data while Naive Bayes has proved to be effective for smaller datasets and Random Forest can be used where robust performance is required. The choice of the classification algorithm will depend on the dataset size and complexity and the results will vary depending on the dataset.

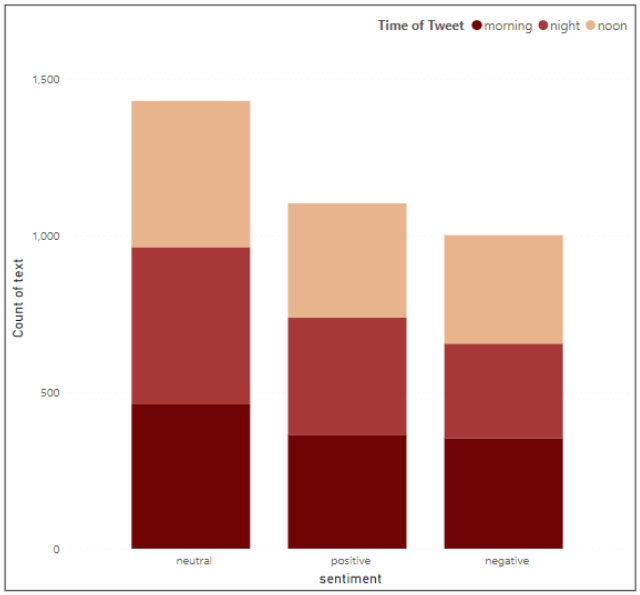

Fig. 4. Stacked Column Chart

V. CONCLUSION

This research paper explores the relationship between Natural Language Processing (NLP) and Social Media Analytics, highlighting its significant impact on understanding textual data generated across social media platforms. NLP techniques can be used to extract, decipher, and interpret sentiments, opinions, and trends within digital discourse. This has enabled businesses, researchers, and decision-makers to make informed choices in a data-driven world. However, the paper also highlights challenges such as data noise, sentiment ambiguity, ethical considerations, and the ever-evolving nature of language. The paper highlights the need for ongoing research, adaptability, and ethical data handling in NLP-driven social media analytics. As social media evolves, so too will NLP, blurring the boundaries between human language and machine understanding, enriching our understanding of digital discourse and shaping our engagement with the digital world.

REFERENCES

[1] IBM. 2023. What is Natural Language Processing? https://www.ibm.com/topics/natural-language-processing.

[2] Alua, M. Y. 2023. A Real-time Sales Dashboard using Machine Learning on the MERN Stack. IEEE 3rd International Conference on Intelligent Sustainable Systems (ICISS). pp. 303-308.

[3] Raj, N. 2023. Starters Guide to Sentiment Analysis Natural Language Processing. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/nlp-sentiment-analysis/

[4] Farzindar, A., Inkpen, D. 2015. Introduction to Social Media Analysis. In: Natural Language Processing for Social Media. Synthesis Lectures on Human Language Technologies. Springer, Cham. https://doi.org/10.1007/978-3-031-02157-2-1

[5] J. Singh and G. Singh. 2028. Sentiment Analysis of Social Media Reviews using QOS Parameterization. 2018. First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 255-260, doi: 10.1109/ICSCCC.2018.8703351.

[6] A. Sriteja, P. Pandey and V. Pudi, "Controversy Detection Using Reactions on Social Media," 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, 2017, pp. 884-889, doi: 10.1109/ICDMW.2017.121.

[7] K. K. Pandey, M. Thorat, A. Joshi, S. D, A. Hussein and M. B. Alazzam. 2023. Natural Language Processing for Sentiment Analysis in Social Media Marketing. 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). Greater Noida, India. pp. 326-330, doi: 10.1109/ICACITE57410.2023.10182590.

[8] Z. Liu. 2022. User Adoption Analysis of Social Media Applications Based on NLP and Technology Acceptance Model - A Case Study of Facebook. 5th International Conference on Data Science and Information Technology (DSIT), Shanghai, China. pp. 1-6, doi: 10.1109/DSIT55514.2022.9943887.