Sentiment Analysis on Social Media

Jyoti Yadav[1]

1 Montclair State University

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

Sentiment analysis is a technique that uses machine learning, natural language processing, and computational linguistics to understand people's feelings and opinions on social media. With the growth of social media users, there is an abundance of information shared in various forms including text, photos, audio and video. Sentiment analysis divides this information into positive, negative or neutral sentiment and has been divided into three levels: sentence, document and features. There are two main methods of sentiment analysis: machine learning and lexicon-based approaches. Sentiment analysis has applications in understanding public perception of eco-friendly transformation and air quality, predicting box office success, and analyzing public sentiment during global events such as pandemics. However, there are challenges in developing accurate sentiment analysis models for all languages.

Jyoti Yadav

yadavj2@montclair.edu

1. Introduction

In this advanced growing world, we can notice an enormous boost of user-generated data over various social media platforms and sentiment analysis. It has come across as dominant automation to get understanding from social media. Sentiment analysis is an analysis technique that studies people’s feelings and opinions and it has grown significantly over the years[1].

The sentimental analysis serves as an offshoot of machine learning(ML), DataMining(DM), NLP, and computational linguistics, which also have components from sociology and psychology[2]. Social media users are growing so fast that in 2019 there were up to 2.77 billion SM (social media) users worldwide and in 2022 the number increased to 4.59 billion. In relation to this increase, the kind of information shared and uploaded in the structure of photos, audio, video, and text is massive[1].

Social Media is overflowing with raw and up-to-date information and advanced machinery, especially ML and AI, that

allows data to be refined and turned into useful data that can be beneficial to the outside world[1].

The Paper will give a good understanding of the implementation of Sentiment Analysis over social media platforms by exploring related compositions between 2015 - 2022[3]. It is a method that applies Natural Language Processing(NLP) to abstract and transform the viewpoint from any means of social media information and then segregates it into positive, negative, or neutral sentiment. There have been many scholars working on and publishing papers and journals on sentiment analysis for more than 15 years and still growing promptly[1].

Sentiment analysis has been divided into three distinct levels, first is sentence-level analysis, second is document-level analysis and third is feature-level analysis[2]. The chunk of data available on the internet helps people to understand the user's views and attitudes on most affairs by analyzing the sentiment of multimodal data. Multimodal Sentiment analysis has special claim benefits over box office prediction, political elections or book reading in public, and many more[2].

Sentiment analysis has two main methods, one is the Machine learning

approach(MLA), and the other is the Lexicon-based approach. MLA includes algorithms to draw out and mark sentiment from information on the other hand Lexicon-based approach functions by calculating the constructive and destructive words in relation to the information[3].

Understanding the sentiment or perception of the general public towards eco-friendly forms of transportation, and their awareness of air quality can guide the policymakers to address the needs of the population accordingly. Especially during/after a global pandemic, people have used social media as a form of outlet to express their thoughts and opinions. The public perception of these topics plays an important role in the adoption of eco-friendly options. Studies have shown that there is an influence of environmental consciousness towards the adoption of environmentally friendly transportation options (Liu et al 2013; Liao et al. 2017).

Hence, understanding the sentiment of the tweets before and during/after the COVID-19 pandemic can help us understand the shift in the sentiment of people. Not only this, but we can also find out if there is a growing concern among the people regarding air pollution and the quality of air as well. For the purpose of this project, we train our model to predict the sentiments of tweets, it can help with future predictions to create long-term goals in terms of policymaking. We investigate how the perception of people has changed from January 2018 to December 2022 and analyze the differences in the sentiment of the people, if any, pre-pandemic (2018-2019) and post-pandemic (2020-2022).

2. Literature Review

Scientists and Scholars are working continuously to develop a functional and error-free model in sentiment analysis. They face many hurdles while building an accurate sentiment analysis model that works on all languages, not just English. To Develop the application of sentiment analysis most of the information or data needed was from social media, we have all information about any product, service, place, or event that supports sentiment analysis study[1].

The expansion of big data analytics in the current era raises the urge of marketing professionals to seek contemporary techniques with estimated label presentation in the case of label equity appraisal[1]. The main dispute in recent time operations is that they depend on techniques of conventional data gathering and investigation procedures like questionnaires and one on one or preliminary interviews that have notable hold. One of

the papers they bring in is a computational model that unites subject and sentiment organization to obtain powerful issues from client insight into social media[1].

Its dummy comes up with a narrative genetic algorithm that enhances a group of tweets in connotation logical groups, that behave as crucial conditions when penetrating the prevailing subject in an immense structure of information. To understand the logic of their dummy they have used the Uber matrix, from data gathered via Twitter. Its outcome prevails available to the client and manufactures cognizance for two elementary label equity dimensions: label consciousness, and label connotation[1].

Social media has sowed the seed of an abundant amount of data. On a regular basis, billions of users share posts and tweets and talk about their solar day[]. Because of this immense pursuit, the social media manifesto gives the leatest chance for research on the human aspect, knowledge diffusion, and impact circulation at a level that is hard to imagine or can say impossible to gather[1].

This social media information is considered the latest gem cache for data mining and predictive analytics. We are well aware of the fact that social media data vary from conventional data, and it is very interesting to know about its crucial way of study and its distinctive attributes[3]. In this paper, I am trying to look into information or data-gathering prejudice related to social media platforms[1].

In a certain way, I am trying to suggest a computational method that evaluates to know if it has unfairness or in the process social media platforms assemble its data or information available, to mark the unfairness from sample information without approaching the whole information and to reduce bias by creating information cluster programs that will help to increase the scope to reduce unfairness[1].

The latest type of information unfairness comes from Application Programming Interface(API) ambush that works on algorithms, information/data, and substantiate outcomes. This type of work helps us to understand how different types of information, ideas and characteristics of social media information/data can be largely considered and authenticated and in what way can be in touch arbitration mechanisms can be drafted to control pessimistic consequences. The procedure and discovery of this exertion can be thoughtful to understand various objectives of social media data/information[1].

One of the papers in 2017 explored the public point of view on a new academy meals program for childhood obesity avoidance, finding features with respect to those views

and recognizing feasible gender and geographical dissimilarity in the U.S. It has collected 14.317 pertinent tweets from 11.715 clients from the point of national policy ratification from Feb 9, 2010, through Dec 31, 2015. They appeal to belief mining methods to differentiate tweets into positive, negative and neutral groups and can regulate content analysis to

obtain awareness into features of belief turn of phrase earmark, receptacle, origin and purpose[1].

In the present day, there are many studies being conducted to explore machine learning techniques and language processing from the posts made on social media and other online platforms to assess the sentiment of the public regarding important issues. A 2018 study done to analyze 6,000 tweets about air quality and transportation found that the users were more likely to tweet negative sentiment towards air quality when they used public transportation, compared to those who used private vehicles, who were more likely to tweet to express a positive sentiment (Gurajala and Matthews 2018).

Some users were more likely to express their negative sentiment on transportation when they experienced delays in the services, but did not express positive sentiment tweets as much when they experienced better services (Das and Zubaidi 2021). Similarly, the other studies on people's sentiment toward air travel show that their sentiment became increasingly more negative after the COVID-19 pandemic (Field et al 2022).

Most sentiment analysis studies done in the past have shown that the sentiment of people changes based on the events and experiences they have. Since COVID-19 is a recent event, not many studies that show the change of sentiment in terms of both air quality and eco-friendly transportation have been done. With this review, we can state that our study dives deeper into this topic, and helps compare and analyze the changes in terms of people's sentiments through tweets towards these topics before and after the pandemic. Major steps for sentiment analysis Dataset:

- Converting text in the tweets in lowercase.
- Removing most common stop words such as a, about, above, etc.
- Removing non-character texts such as punctuations and emojis from the text in the tweets.
- Filter and remove repeated words, URLs, and numbers from the text.
- Tokenization was done to convert texts into tokens, that is, to split sentences into smaller units or words. This was done so meaning can be assigned to the word more easily.
- Stemming was done to extract the base form of the words by removing affixes from them. (Ex: words such as "likes", "likely" and "liked" returned as "like" after stemming).
- Term Frequency-Inverse Document Frequency Vectorizer (TF-IDF) was pre-owned to assess how relevant a term is in the corpus/text data, where TF-IDF vectorization is a process for calculating the TF-IDF score for every word.
- Suitable n-gram was created as necessary.

Here calculated the sentiment scores for different emotions using packages in R Studio and all these scores are then summed up to get the final sentiment scores for each keyword that we searched the sentiment scores for Keyword EVs. We used the same process to get sentiment scores for all the keywords.

Then, the preprocessed data was used in the predictive model. Here, use logistic regression, support vector regression (SVM), and Bernoulli Naïve Bayes for predictive analysis. This was done because we want to try simple to complex classifiers, and find the model that gives the best results with higher accuracy. For the predictive model, input will always be ($A \rightarrow b$), where A is text and b is sentiment.

3. Methodology

3.1. Social media mining (SMM) and SA (sentiment analysis)

We are aware of the fact that the internet consists of a large amount of information in the form of data and it is available in two forms: structured and unstructured texts[1]. By inspecting the unstructured text we got to know that it is a valuable knowledge item and possibly more crucial than drawing out structured data because of the total quantity of important data/information of nearly somewhat believable set hold in them.

Any business wants to have a mass or client viewpoint on their byproduct and assistance. Future clients want to have an idea about what the existing clients' viewpoint is prior to using the byproduct or using assistantship for the product. Additionally, we can say that sentimental analysis by another name is opinion mining, which can give important information/data by putting advertisements on web pages.

By going through the page we understood that people are showing beneficial or useful judgements or sentiments on the byproduct, then it is a possibility that putting the advertisement on that page is a positive idea. Anyhow, if clients show obstructive

judgments about the byproduct, then it's possibly not a good sign to list that product in an advertisement on that page. A good plan may be to put an advertisement for competitor byproducts.

Sentiment analysis can be used to examine positive, negative and neutral points of view of clients of specific brands or services. By considering the point of mining on the web it makes it more difficult and technically challenging for the reason that it requires natural language processing(NLP). Moreover, it is also very important for training[1].

We can notice from a decade that the internet has changed the pattern in which people express their judgements, feelings and thoughts. They have the flexibility to post on any of the byproducts they are using and can express how satisfied they are with that byproduct and if they need to change something in that product they can even suggest the changes too. I provide the flexibility of internet forms that save lots of paper, discussion groups where they can discuss anything in the world, blogs, Twitter, Instagram, Facebook, Foursquare etc[2].

The online oral messages show a new and accountable source of data/information with numerous applied applications. Social media mining became desired because of these types of features.

It is a process of drawing out, inspecting and constituting measurable samples from social media data/information. Social media mining launches an elementary concept and principal algorithm acceptable for scrutinizing large social media data; it considers conjecture and methodologies for various disciplines some of them are CS(computer science), ML(machine learning), DM(data mining), SNA(social network analysis), statistics, sociology etc. It surrounds the gadget according to protocol: produce, calculate, dummy and mine informative structure for huge social media data[1].

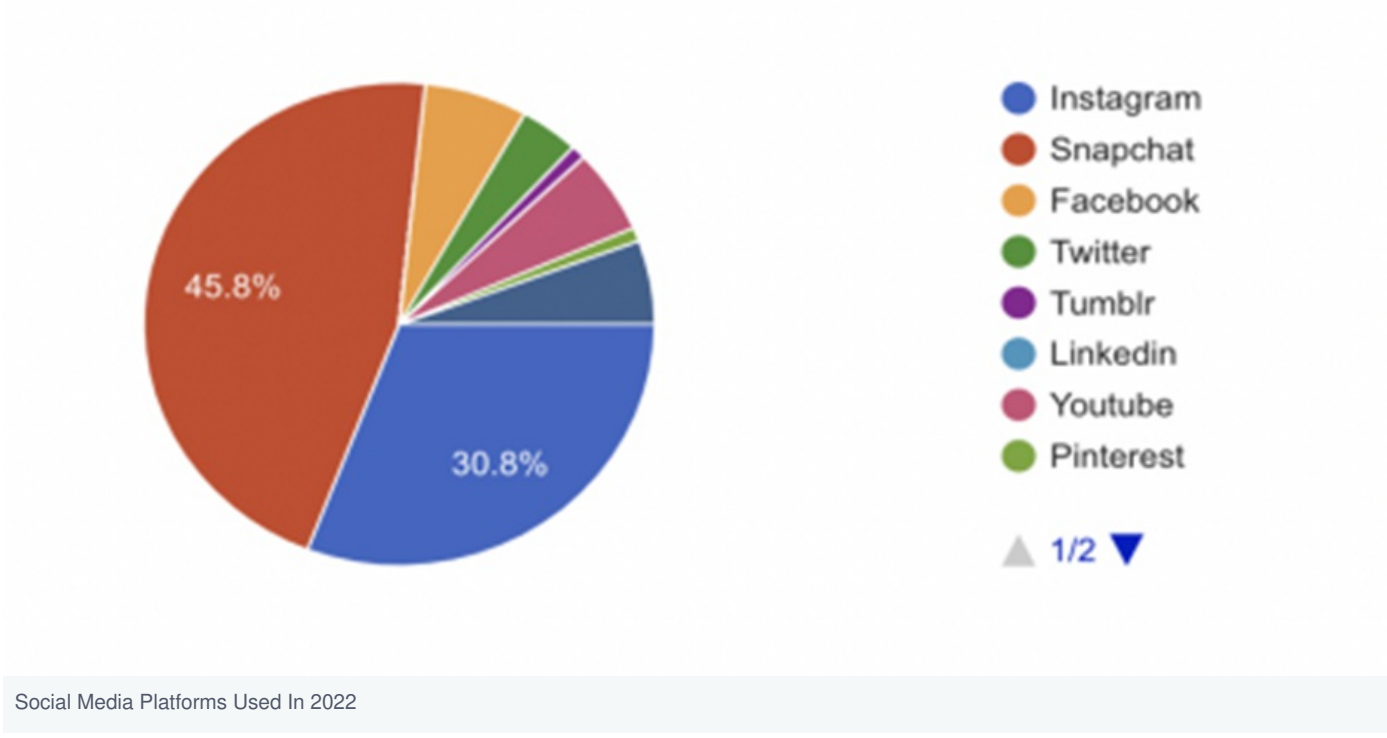
In 2014, Twitter and Facebook were the most popularly used applications but over time things have changed a lot in the

social or technological world. The new study shows that in 2022 the popular applications of 2014 are not that popular as compared to new social media applications like Instagram and Snapchat.

In this growing technology, or in this evolving world, everything is changing rapidly and people want to update versions of everything. Twitter and Facebook were new for users in 2010 or even in 2014 but the new generation finds them old and less expressive, so they chose Instagram and Snapchat over them[3].

The year 2020 is the year when the pandemic was at its peak, and thus it could have had an impact on the thoughts and opinions of the people. There has been an increase in the negative sentiment of tweets regarding both air quality and eco-friendly transport after the pandemic, whereas tweets with both positive and negative sentiments have increased significantly.

This might be because people are expressing concern and frustration regarding the degrading quality of air and its effects on health, and some are calling for action regarding the matter. This shows that post-pandemic (after 2020), there has been a growing awareness in regards to the quality of air among Twitter users and along with this, there has also been an increase in interest in eco-friendly transportation options.



Social Media Platforms Used In 2022

Data labeling: Visual and Auditory markers of arousal and mood.

Facial features of arousal and mood: Facial pronouncement of feeling were encrypted found on obvious facial motion.[4] Facial characteristics kept in touch with The online oral messages show a new and accountable source of data/information with numerous applied applications.[5] Social media mining became desired because of these types of

features. It is a method that applies Natural Language Processing(NLP) to abstract and transform the viewpoint from any means of social media information and then segregates it into positive, negative, or neutral sentiment.[6]

Speech content features of arousal and mood:

Speech attributes/features were extracted with natural language processing (NLP) using Receptiviti which uses the LIWC 2015 dictionary.[7] The latest type of information unfairness comes from Application Programming Interface(API) ambush that works on algorithms, information/data, and substantiate outcomes.[8] This type of work helps us to understand how different types of information, ideas and characteristics of social media information/data can be largely considered and authenticated and in what way can be in touch arbitration mechanisms can be drafted to control pessimistic consequences.

3.2. Machine Learning

Social Media is flooded with raw and up-to-date information and advanced machinery, especially ML and Artificial intelligence (AI), that allows data to be refined and turned into useful data that can be beneficial to the outside world.[9] The Paper will give a good understanding of the implementation of Sentiment Analysis over social media platforms by exploring related compositions between 2015 - 2022.

It is a method that applies Natural Language Processing(NLP) to abstract and transform the viewpoint from any means of social media information and then segregates it into positive, negative, or neutral sentiment. There have been many scholars working on and publishing papers and journals on sentiment analysis for more than 15 years and still growing promptly.

Sentiment analysis has been divided into three distinct levels, first is sentence-level analysis, second is document-level analysis and third is feature-level analysis. The chunk of data available on the internet helps people to understand the user's views and attitudes on most affairs by analyzing the sentiment of multimodal data.

Multimodal Sentiment analysis has special claim benefits over box office prediction, political elections or book reading in public, and many more. Sentiment analysis has two main methods, one is the Machine learning approach(MLA), and the other is the Lexicon-based approach. MLA includes algorithms to draw out and mark sentiment from information on the other hand Lexicon-based approach functions by calculating the constructive and destructive words in relation to the information.

3.3. Naive Bayes

Think about the project which has attributes, that you understand regarding the necessity of attributes and you are thinking about creating a machine learning classifier model that is shown to the outer world in a very tiny span of time.

How do you plan to do that? You consist of a huge amount of Dataset and on the other side very less number of attributes in a dataset. In the given circumstances you have been told to create a model. In that case, I must have used “Naive

Bayes", which is known for its very fast processing algorithm if we are looking for different class tasks.

Here I am explaining the process of the algorithm, how it works and how we can use this in our scenarios mostly for binary and multiclass classification.

Naive Bayes is one of the Machine Learning models(MLM) that provide the base to run on a huge amount of data set apart from the fact that it has millions of rows in it. This has been a very reliable source when it comes to NLP tasks such as sentiment analysis. It is a speedy and straightforward classification algorithm.

3.3.1. Bayes Theorem

It is used in conditional probability. Conditional Probability is the type of condition that comes up when something will happen by the provided circumstances it has already happened. It can give the probability of an event by understanding its given information.[2]

$P(X|Y) = P(Y|X).P(X)/P(Y)$[2]

Where,

P(X) = It is the probability of a given hypothesis(H).[2]
P(Y) = Probability of evidence.
P(X|Y) = Probability of evidence that the given hypothesis is true.
P(Y|X) = Probability of hypothesis that the given evidence is true.

Type of Naive Bayes Algorithm:

1. Gaussian Naïve Bayes: It says that when the distinctive benefits are regular in the given pattern then supposition has clarified that the point is connected with the given class.
2. Multinomial Naïve Bayes: It is said to be useful with data that consist of multinomial distribution. It is mostly used in text classification in NLP.
3. Bernoulli Naïve Bayes: When characteristic of the given data is in relevance to be capable of changing Bernoulli Distribution in that case Bernoulli Naïve Bayes is used.

4. Summary

In this paper, social media mining and sentiment analysis were pre-owned to investigate social media information/data in consideration of a few leading companies across the world. Different companies from various parts of the world are considered for this paper and analyze its data. Label details and storyboards are being used for describing the information regarding sentiment analysis on various social media applications across the world[1]. Pie charts have been used to show the distribution of social media usage by users across the continents and what are those applications that have been mostly used by the user or clients. Bringing out characteristics from social media messages has been demonstrated to be

a vigorous technique for many other brands.

In order to conduct a more comprehensive sentiment analysis about eco-friendly transportation and air quality, several other factors need to be considered., posts and comments from social media platforms will also be included in future research.

Overall, sentiment analysis of tweets data can help get valuable insights into the public's opinion on these topics, which allows businesses, organizations, and decision-makers to make informed decisions based on the sentiment, and the needs of their audience.

The organized literature review gives data on the analysis of sentiment studies on social media. The references give the three possible structures. More overly we have shown what methods we have been using in analyzing the sentiments of people on social media. It has given many techniques shown by the researchers but one of the popular methods is the Lexicon-based method with SentiWordnet and TF-IDF other than that machine learning is Naïve Bayes and SVM.

The best way of doing sentiment analysis on the given dataset is totally dependent on the type of dataset used. Because of this immense pursuit, the social media manifesto gives the leatest chance for research on human aspect, knowledge diffusion, and impact circulation at a level that is hard to imagine or can say impossible to gather. For more information, we have to work on developing sentiment analysis models that help to prove more hypotheses so that data can be understood well and other necessary majors can be taken by society.

References

1. [a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q] *T. Zhao, C. Li, M. Li, Q. Ding, and L. Li, "Social recommendation incorporating topic mining and social trust analysis," Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13. 2013. doi: 10.1145/2505515.2505592.*
2. [a, b, c, d, e, f, g] *R. Dwivedi, "What Is Naive Bayes Algorithm In Machine Learning?" https://www.analyticssteps.com/blogs/what-naive-bayes-algorithm-machine-learning (accessed Dec. 15, 2022).*
3. [a, b, c, d] *S. Wu, Y. Liu, Z. Zou, and T.-H. Weng, "S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis," Connection Science, vol. 34, no. 1. pp. 44–62, 2022. doi: 10.1080/09540091.2021.1940101.*
4. [^] *K. Schultebraucks, et al., "Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood," Psychol. Med., vol. 52, no. 5, pp. 957–967, Apr. 2022.*
5. [^] *A. Abbas et al., "Remote Digital Measurement of Facial and Vocal Markers of Major Depressive Disorder Severity and Treatment Response: A Pilot Study," Front Digit Health, vol. 3, p. 610006, Mar. 2021.*
6. [^] *I. Galatzer-Levy et al., "Validation of Visual and Auditory Digital Markers of Suicidality in Acutely Suicidal Psychiatric Inpatients: Proof-of-Concept Study," J. Med. Internet Res., vol. 23, no. 6, p. e25199, Jun. 2021.*

7. ^*A. Abbas, V. Yadav, M. M. Perez-Rodriguez, and I. Galatzer-Levy, "P.267 Using smartphone-recorded facial and verbal features to predict clinical functioning in individuals with neuropsychiatric disorders," European Neuropsychopharmacology, vol. 29. p. S199, 2019. doi: 10.1016/j.euroneuro.2019.09.301.*
8. ^*A. Abbas et al., "Facial and Vocal Markers of Schizophrenia Measured Using Remote Smartphone Assessments: Observational Study," JMIR Form Res, vol. 6, no. 1, p. e26276, Jan. 2022.*
9. ^*K. Schultebraucks, V. Yadav, and I. R. Galatzer-Levy, "Utilization of Machine Learning-Based Computer Vision and Voice Analysis to Derive Digital Biomarkers of Cognitive Functioning in Trauma Survivors," Digit Biomark, vol. 5, no. 1, pp. 16–23, Jan-Apr 2021.*