Machine Learning-Based Sentiment Analysis

T. Murali Krishna[1], Dr. D. William Albert[2], A. V. Rama Krishna Reddy[3]

[1,3] Asst. Professor, Ashoka Women's Engineering College; [2]Professor, Ashoka Women's Engineering College

Article Info
Page Number: 240 - 247
Publication Issue:
Vol 71 No. 1 (2022)

Article History
***Article Received**: 02 February 2022*
***Revised**: 10 March 2022*
***Accepted**: 25 March 2022*
***Publication**: 15 April 2022*

Abstract

Today, almost everyone uses some type of social media, most notably Facebook, Instagram, and Twitter. Through the use of this social media platform, everyone is able to express their own perspective, ideas, and emotions on any given issue. Tweets are the bite-sized chunks of text that users of the social networking service Twitter use to share and discuss their thoughts and opinions on a wide range of topics with their followers and the public at large. Twitter is one of the most well-known social media sites. Businesses can utilize this data to improve their services, increase customer satisfaction with their products, develop more effective strategies, and aid in the decision-making processes of their employees and clients. Tweets are a significant way for businesses to learn how their products are received by consumers, and the feedback they receive is invaluable in shaping future iterations. This study article focuses mostly on the development of sentiment analysis frameworks. Sentiment analysis allows us to read between the lines of someone's written words and understand what they're really thinking. The first stage in conducting sentiment analysis is collecting and organizing a large number of tweets or user thoughts from a social media platform. Using the Natural Language Processing Toolkit and a number of other approaches, we are classifying tweets as positive, negative, or neutral. Their sentiments guided the subsequent categorization of this tweet. Using Python's flask framework, the classified findings are presented in a variety of charts (including pie charts, bar charts, and line charts) and HTML pages.

Index Terms: Sentiment Analysis, Machine Learning(ML).

1. Introduction

Everyday life, including one's thoughts, feelings, opinions, and actions, is all shared on social media platforms these days. Users of social media sites like Facebook, Instagram, and Twitter can easily reach out to others in any part of the world. Conversely, people comment on and record their social activities as well as whatever else comes to mind. They can even affect politics and corporations with the words they put on paper. Businesses may reach out to their target audience more easily thanks to the medium provided by social media. For instance, companies can utilize social media to spread the word about their wares or to have direct conversations with clients to better understand their needs and wants. To gauge how customers feel about a company's offerings, Twitter is a popular social media platform used by nearly all businesses nowadays. Businesses can now quickly and easily gauge customer opinion of their products and services thanks to Twitter and other social media sites[1]. Organizations may take action at the right moment and craft effective strategies to ensure they provide customers with the highest quality goods, services, and experiences possible. We needed a large amount of data in the form of people's thoughts, which are easily accessible on Twitter, in order to undertake sentiment analysis. In order to do sentiment analysis, we need to collect data from Twitter. For instance, you can use natural language processing methods to get information out of tweets. These methods exist and can be discovered. There is no set style for tweets like these. Sentiment analysis requires a structured data set before it can be performed[2]. This can be done by applying several strategies to the data or tweets, such as transforming the information into a structured format and identifying the tweet's overall tone to determine whether it is positive, negative, or neutral. The NLP (Natural Language Processing) Toolkit, Python 3 (and its modules), and the Tweepy (Twitter Application Programming Interface) API are being used to make this a reality. We have used VADER's sentiment analysis. The name, which is an acronym for "Valence Aware Dictionary and Sentiment Reasoner," describes this analytical tool perfectly: it relies on rules and a vocabulary to draw conclusions[3].

Value-Aware Dictionary and Emotional Reasoner is the full name for this concept. The sentiment of tweets can be gleaned with the help of this instrument[4]. Different types of charts (pie charts, bar charts, line charts, etc.) are used to present the data analysis's findings. As an added bonus, we exclusively analyze individual tweets. We've included a section that can dissect a single tweet or line as well as the most recent news, a review of any product, or commentary on any topic provided in text form. The user only needs to type or paste the text he wants to look at into this field. In addition, the user can assess the tone of a massive amount of Twitter data simultaneously. Once everything has been processed, the user will be given the total number of tweets, the mood of each tweet, and the percentage of positive, negative, and neutral tweets[5].

2. Framework

1. **Natural Language Processing:** The underlying principles of NLP's methods can be traced back to the field of machine learning. Natural language processing, or NLP, is an area of computer science whose major goal is to teach computers to read, write, comprehend, and translate between human languages. To achieve this goal, one must adhere to a set of rules for education. A number of algorithms and lexical databases containing such things as word definitions, synonyms, and spelling corrections are used in natural language processing. Natural language processing (NLP) has made sentiment analysis more straightforward, facilitating the evaluation of the general tone of tweets and other human-provided input[6].

2. **Lexicon-based Approach:** Words that accurately convey a sentiment are the most reliable predictors of that sentiment. Common phrases used to communicate a positive or negative emotional state include the ones listed here. Sentimentally positive terms include "good," "beautiful," "great," and "lovely," whereas negatively charged terms include "awful," "mad," "sad," and "terrible." My definition of a term with no overarching good or negative meaning is that it is neutral. These kinds of expressions are collected into what is called a "sentiment lexicon," or a collection of terms that all mean roughly the same thing. Some techniques, called lexicon-based methods, utilize this set of keywords. A word's connotation and meaning are drawn from the dictionary and assigned to each entry on the list. The lexicon-based method compares the word's current sentiment with the desired sentiment[7].

3. **NLTK (Natural Language Processing Toolkit):** Different methods and tactics for NLP are being employed. When conducting sentiment analysis, we use a text categorization process driven by NLTK 2.0.4. A user's comment or tweet may be punctuated with emoticons, acronyms, or other symbols to indicate the user's mood. Due to Twitter's 140-character limit, tweets are often no more than a couple of lines long. Therefore, it is our goal to conduct some rudimentary analysis of the tweet in order to extract its meaning. The information gleaned from tweets and downloaded from Twitter was saved in JSON format. Our opinion is that tweets and retweets are a true reflection of the emotions of both individuals and institutions. We pay the most attention to details like Twitter handles, locations, tweet sentiment (positive, negative, or neutral), and posting timestamps. CSV files are used to save the readable data that has been extracted[8].

4. **Data Collection:** We mine Twitter for relevant tweets using the Python Tweepy API and collect thousands of them in one place. We can find and read tweets on everything, from KFC to a specific political party or a specific product. The most recent 4,000 tweets can be retrieved from Twitter. Information obtained via download is in JSON format. The readable tweets and user details

are then extracted from the JSON data. Sentiment analysis is done when data purification is complete. In conclusion, we offer some positive, equivocal, and pessimistic ideas. We are currently utilizing a dataset of "U.S. Airlines tweets" that contains 14,000 tweets, demonstrating the power of this method when combined with the use of datasets[9].

5. Application Programming Interface (API): Tweepy is a Python library that utilizes the Python Twitter API to fetch tweets. Either Tweepy or the Twitter API can be used to retrieve tweets from Twitter based on a wide variety of criteria, including topic, location, user, language, and more. Unfortunately, Twitter limits us to retrieving only 3,000 tweets at a time. There are now more APIs available that make it possible to extract tweets in bulk and conduct searches across more than 3000 tweets all at once.

6. Dataset: We're using a dataset culled from 3,000 scraped tweets. We employ a dataset dubbed the "U.S. airline dataset," which collects tweets about airlines based in the United States. There are a total of 14486 tweets in this collection. Tweets are in an unstructured format since they contain nonprintable characters like hashtags, urls, special characters, symbols, spaces, and so on. Initially, the dataset will be processed and cleaned before moving on to the sentiment analysis proper. Once the data has been analyzed, it is next subjected to a process called sentiment classification, which sorts the tweets into three groups: positive, neutral, and negative.

3. Implementation

a. **Pre-processing:** Twitter data is generally unstructured, so it must be cleaned and prepared before analysis can be performed. There is a lot of legwork involved in the data's pre-processing phase. In our case, we care solely about the text itself. Given that both the tweets we've acquired from Twitter and the data we're utilizing to determine their tone come to us in an unstructured format, Information like links, hashtags, and other symbols and characters that do not convey any sense and that machines are unable to grasp can be discovered in tweets or data. When applied to sentiment analysis, every word in a tweet is important. We were able to do our analysis of sentiment with only the tweet text and the data. This is done during pre-processing, which involves removing tweets' links, hashtags, special symbols, and characters before analyzing the remaining text for sentiment. Only humans can decipher the meaning of abbreviations like "Gm," which stands for "Good morning," or spelling errors like "goood" to "good," which aid in determining the tone of tweets. Some tweets may have been sent with the first letter of each word capitalized or may have had a misspelled term that was later rectified during pre-processing. Once everything has been processed, the forwarded tweets are evaluated to see how people feel about certain topics.

b. **Cleaning Tweets:** Due to its very unstructured nature, the language of tweets must be cleaned and sorted before analysis can be performed. The process of cleansing the data is lengthy. In our case, we care solely about the text itself. We removed URLs, stop words such as "the," "a," and "to," usernames and accounts, numbers and unnecessary spaces, punctuation, and emojis encoded in latin1 and converted them to ASCII. We then created a data frame using the tweets' extracted data. As soon as the text has been cleaned up and any unnecessary symbols have been deleted, the analysis can begin.

1. **Tweet Crawling:** The procedure is started by the user entering the topic in order to get some related tweets. The server received the topic, ran it through the Twitter API, and then returned a set of tweets that had been filtered for topical language.
2. **Tokenizing:** In this step, the tweets are parsed into their component words. Spaces in tweets have been eliminated, and all lowercase letters have been replaced with capitals.
3. **Slang removal:** The aim of this stage is to replace contemporary terms with more conventional ones that better conform to linguistic norms. This adjustment is made by first compiling a database collection of slang phrases and their synonyms. The next step is to check if any of the current words are slang by comparing them to the database's collection of such terms. If a match is found, it will be converted to the appropriate synonym. The word will be excluded from the results if it does not match.
4. **Stop Word removal:** Multiple words within a sentence give the appearance that the word's original meaning has been diluted. Words like "the," "an," and "the" are examples of stop words (a, the, an, etc.). These sorts of words are rarely used in tweets and are therefore removed. Now we compare each word to the list, and if it's there, we take it out of the tweet. A database of prohibited words is kept in the Twitter API.
5. **Stemming:** Currently, we're aiming to have the base word shared on Twitter. For this purpose, we utilize the Porter stemming tool from Python's standard library to discover the original word's meaning. This is done to guarantee that as little data as possible is used during the processing of tweets. Meaningful lexical building blocks, or root words, will emerge from this process. Take the word "running," for instance; we're going to get rid of the "ing" and just use the word "ran."
6. Numbers removal: Numbers in tweets are being removed since they are irrelevant and worthless to sentiment analysis. No further progress can be made in the process of erasing numbers from Twitter posts.

7. **Special character and symbols removal:** At this point, any non-alphanumeric characters (#, *, /, $, https://, www) have been removed from the data. This is done because we needed plain text for analysis because special symbols and letters add no meaning.

8. **Calculate emoticons sentiment score:** Emoticons are a type of graphic representation used to express human emotion without words. Emoticons are a fantastic tool that facilitate smooth communication. By using these graphical representations of human emotions, people can express themselves clearly and concisely online. To provide one example, the character personifies an upbeat demeanour. All of the scores for the emotions represented by the emoticons are now being tabulated. As a result, we've built a lexicon that incorporates not just the standard set of emoticons but also the full range of Emoji symbols, including :) ;): (. The positive emoticons were given the value (+), the negative ones the value (-), and the neutral ones the value (0). Here, the emoticons are checked against a database of emoticon definitions for a match. Then we take the sum of each emoticon's value and divide it by 10, arriving at our final emoticon score.

c. **Sentiment score:** Once everything has been filtered and cleaned up, what's left is the actual substance of the tweets. Words are retrieved one by one from the tweets and then compared to their matching entries in WordNet using the Vader lexicon tool from the NLTK library of Python. WordNet is a dictionary that provides definitions and etymologies, as well as moods and associations, for over ten million words. The Lexicon Vader tool obtained the word's value from WordNet and provided it. The emotion score of tweets is the numerical number that is derived from the weights assigned to each word. Once we have a sentiment score, we use a machine learning-based algorithm to put each tweet into one of three categories: positive, neutral, or negative.

4. Results

	+ve	**Neutral**	**-ve**	**Tot**
Total Tweets	6412	2820	5354	14586
% of Tweets	**45%**	**19%**	**36%**	**100%**

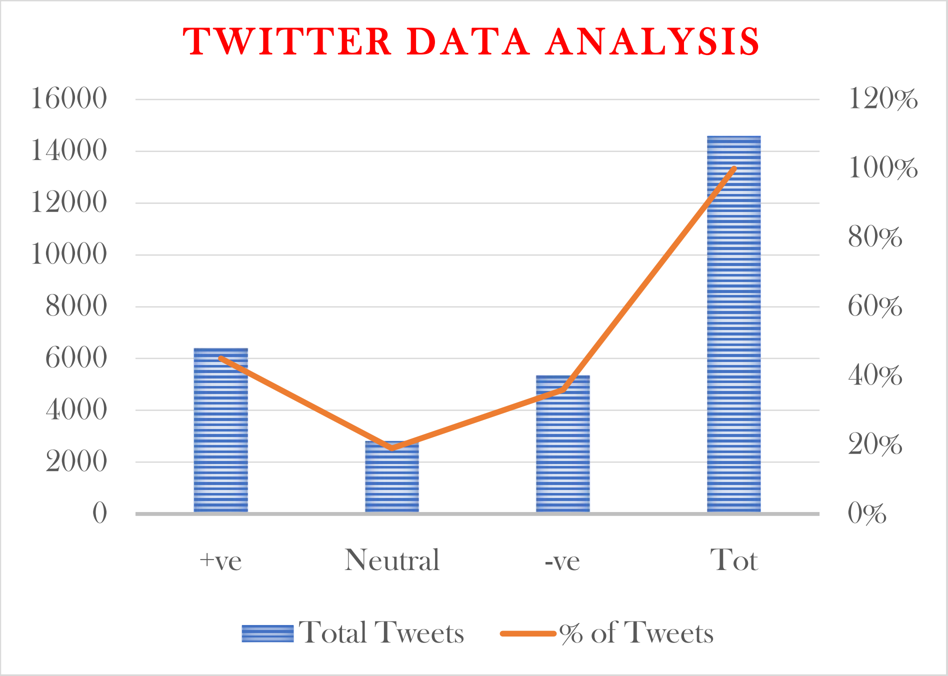

Fig.1 Twitter Data Analysis

5. Conclusion

Using Twitter to get insight into customer sentiment and inform business decisions is a primary motivation for the field of sentiment analysis. Sentiment analysis is a technique that may be used to get insight into how other people feel about a product and use that to inform one's own purchase choice. It helps businesses learn what their consumers think of the services and products they offer, and it helps them figure out how to best adapt to their needs and provide the best possible care. Using NLP, we are determining if the Twitter data expresses positivity, pessimism, or apathy and then categorizing the results accordingly. Following the determination of how Twitter users felt about a topic, the data was visualized using bar charts, line charts, and pie charts. Using the Flask framework in Python, we illustrated sentiment analysis in the portal's online environment.

References

[1] R. Aishwarya, C. Ashwatha, A. Deepthi, and J. Beschi Raja, "A Novel Adaptable Approach for Sentiment Analysis," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 5, no. 2, 2019.

[2] A. Blom and S. Thorsen, "Automatic Twitter replies with Python," in *International conference "Dialog*, 2012.

[3] B. N. Supriya, V. Kallimani, S. Prakash, and C. B. Akki, "Twitter sentiment analysis using binary classification technique," in *International Conference on Nature of Computation and*

Communication, 2016, pp. 391–396.

[4] S. Yoo, J. Song, and O. Jeong, “Social media contents based sentiment analysis and prediction system,” *Expert Syst. Appl.*, vol. 105, pp. 102–111, 2018.

[5] A. H. Huang, D. C. Yen, and X. Zhang, “Exploring the potential effects of emoticons,” *Inf. \& Manag.*, vol. 45, no. 7, pp. 466–473, 2008.

[6] R. Varaprasad and G. Mahalaxmi, “Applications and Techniques of Natural Language Processing: An Overview.,” *IUP J. Comput. Sci.*, vol. 16, no. 3, 2022.

[7] P. Lai, “Extracting strong sentiment trends from Twitter.” Stanford Digital Library, Stanford University, Stanford, California, United~…, 2010.

[8] M. S. Deelip, K. Govinda, S. Ramasubbareddy, E. Swetha, and A. S. T Srinivas, “Analysis of Twitter Data for Prediction of Iphone X Reviews,” *J. Comput. Theor. Nanosci.*, vol. 16, no. 5–6, pp. 2050–2054, 2019.

[9] A. S. T Srinivas, K. Govinda, S. Ramasubbareddy, and E. Swetha, “Sentimental Analysis of Demonetization Over Twitter Data Using Machine Learning,” *J. Comput. Theor. Nanosci.*, vol. 16, no. 5–6, pp. 2055–2058, 2019.