A Comprehensive Analysis of Approaches for Sentiment Analysis Using Twitter Data on COVID-19 Vaccines

Amrita Mishra1, Mohd. Saif Wajid2, Upasana Dugal3

1,2,3Department of Computer Science Engineering, BBD University, Lucknow, India
mishra17amrita@gmail.com1, mohdsaif06@gmail.com2, upasana_gupta31@bbdu.ac.in3

Abstract

Sentiment Analysis has paved routes for opinion analysis of masses over unrestricted territorial limits. With the advent and growth of social media like Twitter, Facebook, WhatsApp, Snapchat in today’s world, stakeholders and the public often takes to expressing their opinion on them and drawing conclusions. While these social media data are extremely informative and well connected, the major challenge lies in incorporating efficient Text Classification strategies which not only overcomes the unstructured and humongous nature of data but also generates correct polarity of opinions (i.e. positive, negative, and neutral). This paper is a thorough effort to provide a brief study about various approaches to SA including Machine Learning, Lexicon Based, and Automatic Approaches. The paper also highlights the comparison of positive, negative, and neutral tweets of the Sputnik V, Moderna, and Covaxin vaccines used for preventive and emergency use of COVID-19 disease.

Keywords

Sentiment Analysis (SA), Machine Learning (ML), Supervised Learning, Unsupervised learning, Twitter.

1. Introduction

Today the world is a Machine Dependency era. Well-formed systems for information exchange from peer to peer or B2B are established. The need of the hour is to ensure that besides navigating the data soil, customer’s sentiments are evaluated. The correct assessment of user sentiments proves to be highlighting feature in winning or losing the product’s name and growth in market. Earlier the information and feedback exchange systems were file and paper based which was accessible by
limited people. However, today social media like Twitter serves as major platforms where users freely express their opinions and it is accessible within remote areas. The developers can even analyze the tweets based upon selective geographic locations and form conclusions on regional basis.

Sentiment analysis (SA) is the area which deals with judgments, responses as well as feelings, which is generated from texts, being extensively used in fields like data mining, web mining, and social media analytics because sentiments are the most essential characteristics to judge the human behavior [29]. Customer sentiments can be found in tweets, comments and reviews. For example, Reviews delivered by customer on online sites after purchasing the product, property or visiting the hotels. Sentiment Analysis plays a vital role in Data Science as it brings forward a computational study about the diverse opinions. The calculative study provides a platform to derive the true meaning of customer opinion whether positive or negative or neutral.

What more that sometimes even if customer gives bias comments like, “The A.C. Compressor is working well but costly one.” can be examined by machine to draw accurate conclusion. Furthermore, dwelling into study of Sentiment Analysis involves descriptive study of SA via Machine Learning and Lexicon based approaches. Some basic types of SA are Fine grained, Emotion based, Aspect based and Multilingual based. Fine grained is used when polarity precision is important for a business. Example, Very positive = 5 stars and Very Negative = 1 star. Emotion based aims at detecting emotions, like happiness, frustration, anger, sadness, and so on while Aspect Based analyze sentiments of texts, let’s say product reviews to know which particular aspects has positive, neutral, or negative way. Multilingual Analysis involves a lot of preprocessing and resources e.g. translated corpora or noise detection algorithms are the There has been a lot of past research on different strategies to use the web innovation to expand the advantages of clients and in addition organizations in the commercial center.


Twitter, one of the most popular micro blogging social networking site where people tweet their opinions in a concise manner, typically in less than or equal to 140 words [14]. Twitter platform is widely used to deliver tweets relating with vaccines. Hence Twitter datasets of vaccines have been used. The paper is arranged as follows: Section 2 is the related work of the studies conducted by researchers using Twitter Data. Section 3 describes the methodologies in machine learning and lexicon-based approaches for sentiment analysis. Section 4 consists of Table of Tools for sentiment analysis. Section 5 approaches towards results and discussions of the positive, negative and neutral sentiments for the vaccines among Twitter users in Pandemic.

2. Literature Survey

Sentiment Analysis has been of avid interest to researchers lately. A lot of work has been put into it and there is a vast domain of its applications. Gaurav Bhatt [21] has performed Sentiment Analysis over Educational institutions Using Twitter Dataset of IIT, NIT and AIIMS Colleges in India with SVM, Naïve Baye’s and ANN algorithms and accuracy of 89.6%.

The area of Neural Networks has been investigated for performing sentiment analysis on benchmark dataset consisting of online product reviews. Bespalov et al. [2] carried out binary classification on Amazon and Trip Advisor dataset using Perceptron classifier and obtained one of the lowest error rates among their experiments of 7.59 and 7.37 on the two datasets respectively. Researchers have also been working upon prediction of accuracy of tested datasets using Machine Learning Algorithms. Kanakraj and Guddeti [3] used Natural Language Processing techniques for Sentiment Analysis and compared Machine Learning Methods and Ensemble Methods to improve on the accuracy of classification. Shahheidari, et al. [4] used a
Naïve Baye’s Classifier for classification and tested it for news, finance, job, movies and sportstaking into consideration Data Mining on basis of two emoticons (☺ and ☹).

Prediction of Election Results is another domain in which massive population expresses opinion over Social Networks. Rincy Jose and Varghese S Chooralil [7] have used Twitter Data with Classifier Ensemble Approaches with accuracy of 71.48% in predicting election results. Rincy, et al. [9] have also predicted election results with Word Sense Disambiguation with accuracy of 78.6%.

Mohd Saif Wajid et al. [25] have used Sentiment Analysis Based on A.I Over Big Data. They have introduced the methodology for creating user recommended data group (Big data) by elaborating a matrix for user recommended data group for big data which is then reduced by dimension reduction technique.

Neethu M.S and Rajasree R [5] used twitter post on electronic products, compared the accuracy between different Machine Learning Algorithm and further improved accuracy by replacing repeated character with two occurrences, including a slang dictionary and taking emoticons into consideration. Jotheeswaran and Koteeswaran(6) performed binary classification on the IMDB dataset by employing a Multi-Layer Perceptron Neural Network and using Decision Tree -Based Feature Ranking for feature extraction and a hybrid algorithm(based on Differential Evolution and Genetic Algorithm)for weight training, thereby obtaining a maximum classification accuracy of 83.25%.

Laszlo and Attila (20) have used fresh scraped data collections over the Recurrent Neural Networks to determine what emotional manifestations occurred in given time interval in COVID-19. The Sentiment Analysis helps in monitoring area based upon the opinion raised in different territories.

3. Methodologies

3.1. Twitter

Twitter is a micro-blogging site where the user’s posts comments and opinions related to services, products, activities, personalities in the form of tweets. Each user has a daily limit of 2,400 tweets and 140 characters per tweets [21]. Tweets of Sputnik V, Moderna and Covaxin are used as datasets. They are extracted using Twitter Developer account. The Credentials granted from Twitter Developer Account are connected with Python using Tweepy library. In this manner 2000 tweets for Sputnik V, Moderna and Covaxin have been extracted.

3.2. Machine Learning Approaches

They can be categorised in three fundament categories: Supervised, Unsupervised and Reinforcement Learning methods.

3.2.1. Supervised Learning

In this, with the input provided as labeled dataset, a model can learn from it. In labeled dataset the answer or solution to it is given as well. Major steps include, loading labeled input dataset, training model and testing. So, a labeled dataset of animal images would tell name of animal. It is further classified to Classification and Regression. The Classification algorithm predicts a discrete value that can identify the input data as a member of particular class or group. The Linear Classifier includes Support Vector Machine (SVM) and Neural Networks. Rule Based Classifier Predicts the result within well-defined set of rules. The Probabilistic Classifier are categorised into Bayesian Network, Maximum Entropy and Naïve based. Naïve Baye’s is based upon Baye’s Theorem and for handling Big Data Maximum entropy is applied. The Regression problems are responsible for continuous data for example, predicting the diabetes status of a patient given the blood pressure, sugar level, etc. Here, the input has to be sent to machine for predicting diabetes according to previous instances.
3.2.2. Unsupervised Learning

Here, no complete and clean labeled dataset is provided. It focuses on self-organized learning that helps find previously unknown pattern in dataset without pre-existing models. Different algorithms like K-means, Hierarchical, PCA, Spectral Clustering, DBSCAN clustering are used in unsupervised learning. For any input X and response variable Y, suppose f(X) = Y, in supervised learning there can be two goals 1. f(X) closely approximates Y, 2. Predict values of Y given X. In unsupervised learning there is no response variable Y. The clusters within dataset are identified based on similarity. It is more useful and dataset is less expensive.

3.2.3. Reinforcement Learning

An agent interacts with its environment by performing actions and learning from errors or rewards. It follows Trial and Error as there is no predefined data and supervision.

3.3. Automatic Approaches

The automatic approach shown in Figure 1, involves feeding a classifier with text as an input and obtaining the polarity category that is positive, negative, or neutral. It involves following two phases:

- **The Training Phase**: In this phase, the original text is divided into TAG and TEXT part. The tag is fed as whole to machine learning algorithm while Text is passed through Feature Extractor. The Feature Vectors for text are generated. The Tag and Feature vector of Text worked upon by machine learning algorithm produces Classifiers.
- **The Prediction Phase**: In this phase, input Text that has to be predicted is passed through Feature Extractor. Feature extractor generates the feature vectors. The Feature Vectors are fed into the Classifiers and suitable category which is positive, negative, neutral for Tag is obtained.

**Feature Extractor and Feature Vector**: A Tag is the predetermined classification or category that a Text fall into. Feature Extraction technique involves conversion of Text into numerical representation in vector form. In Feature extractor, ML uses Bag of words as dictionary, where a vector is obtained by comparison and transformation. For example, if we have defined our dictionary to following words { vaccine, accomplishing, for ,the ,beneficial} and we wanted to vectorize the text “The vaccine for Covid-19 is accomplishing” and “It is beneficial for masses” we would have following representations of text (1,1,0,0,1) and (0,0,1,1,0) as feature vectors. Multiple feature vectors are fed into classifiers.

![Automatic Approach Phases](image-url)
3.4. Lexicon-based approaches

Dictionary put together methodologies for the most part depend with respect to a feeling vocabulary, i.e., a gathering of known and precompiled supposition terms, states and even figures of speech, produced for customary types of correspondence, for example, the SentiWordNet dictionary be that as it may, considerably progressively complex structures like ontologies, or lexicons estimating the semantic introduction of words or expressions can be utilized for this reason. Two sub characterizations can be found here: Dictionary-based and Corpus based methodologies.

3.4.1. Dictionary-based strategies

This involves the utilization of an underlying arrangement of terms (seeds) that are typically gathered and explained physically. This set develops via looking through the equivalent words and antonyms of a lexicon. A case of that lexicon may be WordNet, which was utilized to build up a thesaurus called SentiWordNet. The principle downside of this sort of methodologies is the lack of ability to manage space and setting explicit introductions; all things being equal, it may be an intriguing arrangement relying upon the issue.

3.4.2. The Corpus-based strategies

This emerged with the target of giving word references identified with an explicit area. These lexicons are created from a lot of seed sentiment terms that becomes through the pursuit of related words by methods for the utilization of either measurable or semantic systems. Regular Language Processing and Information Retrieval in Sentiment Analysis According to Cambria, Sentiment Analysis can be considered as an extremely limited NLP issue, where it is just important to comprehend the positive or negative estimations concerning each sentence as well as the objective elements or themes. In any case, regardless of being a limited issue, all works in this field, and all works in Information Retrieval, dependably battle with NLPs uncertain issues (invalidation taking care of, named element acknowledgment, word-sense disambiguation, ) which are fundamental to recognize scholarly gadgets, for example, incongruity or mockery and thus, to discover and rate conclusions. The three dimensions of investigation that decides the distinctive undertakings of Sentiment Analysis are: (I) report level, (ii) sentence level and (iii) element/angle level. Report level thinks about that a record is an assessment on a substance or part of it. This dimension is related with the undertaking called report level opinion characterization. Notwithstanding, in the event that a report gives a few sentences managing distinctive viewpoints or elements, the sentence level is progressively appropriate. Sentence level is firmly identified with the assignment called subjectivity order, which recognizes sentences that express verifiable data from sentences that express emotional perspectives and sentiments Feature-based Opinion Mining and Opinion Summarization. A significant number of these papers pursue indistinguishable general procedures from other Information Retrieval works did previously, however supplanting a few factual or semantic factors for angles identified with assumptions. In this way, the principle distinction between these works is the element determination process. TextBlob, AFINN, VADER (Valence Aware Dictionary for Sentiment Reasoning) are used in Python for Lexicon Based Sentiment analysis.

4. Sentiment Analysis Tools

Sentiment Analysis tools are used in different fields such as politics, finance, business, etc. Sentiment analysis tools are given in Table2[10] and we also refer to Big Data analytics tools by considering a Comprehensive Survey on Big Data Analytics[11].
Table 1: SA Tools and Techniques

<table>
<thead>
<tr>
<th>Sentiment Analysis tools</th>
<th>Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>SentiWordNet</td>
<td>Lexical dictionary and scores obtained by semi-machine learning approaches</td>
</tr>
<tr>
<td>LIWC</td>
<td>Dictionary and sentiment classified categories</td>
</tr>
<tr>
<td>EMOTICONS</td>
<td>Emoticons contained in the text</td>
</tr>
<tr>
<td>SenticNet</td>
<td>Natural Language Processing for inferring the polarity at semantic level</td>
</tr>
<tr>
<td>Happiness Index</td>
<td>Affective Norms for English Words (ANEW) and scores</td>
</tr>
<tr>
<td>AFINN</td>
<td>Affective Norms for English Words (ANEW) but additional attention on language used in microblogging platforms</td>
</tr>
<tr>
<td>PANAS-t</td>
<td>Eleven-sentiment psychometric scale</td>
</tr>
<tr>
<td>Sentiment40</td>
<td>API that allows classifying tweets to classify classes positive, negative, neutral</td>
</tr>
<tr>
<td>NRC</td>
<td>Huge collection of human-provided words with their emotional tags.</td>
</tr>
<tr>
<td>EWGA</td>
<td>Entropy-weighed genetic algorithm</td>
</tr>
<tr>
<td>FRN</td>
<td>Feature relation network considering syntactic n-gram relations</td>
</tr>
<tr>
<td>Scikit-learn</td>
<td>Machine learning and has useful tools for text vectorization</td>
</tr>
<tr>
<td>NLTK</td>
<td>Natural Language Processing Library for Python</td>
</tr>
<tr>
<td>SpaCy</td>
<td>Provides a strong set of low-level functions for Natural language Processing and support for training Text Classifiers.</td>
</tr>
<tr>
<td>TensorFlow</td>
<td>Developed by Google, provides a low-level set of tools to build and train neural networks</td>
</tr>
<tr>
<td>Keras</td>
<td>Works with Neural Networks, RNN and CNN in python</td>
</tr>
<tr>
<td>PyTorch</td>
<td>Deep Learning Framework</td>
</tr>
</tbody>
</table>

5. Results and Discussions

Based upon the datasets containing 2000 tweets each for three vaccines Sputnik V, Moderna and Covaxin and implemented using TextBlob, Lexicon Based Approach following results have been drawn. The TextBlob script gets the tweets as input and returns the text’s polarity in terms of sentiment score. The sentiment score lies in the range of -1 to 1. Hence, the tweets are classified as ‘Negative’ if the score is less than 0, ‘Neutral’ if the score is equal to 0, ‘Positive’ if the score is greater than 0.

Table 2: Number of distribution of tweets

<table>
<thead>
<tr>
<th>Vaccine</th>
<th>Positive Tweets</th>
<th>Negative Tweets</th>
<th>Neutral Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sputnik V</td>
<td>47</td>
<td>42</td>
<td>1911</td>
</tr>
<tr>
<td>Moderna</td>
<td>226</td>
<td>238</td>
<td>1536</td>
</tr>
<tr>
<td>Covaxin</td>
<td>177</td>
<td>67</td>
<td>1756</td>
</tr>
</tbody>
</table>

Table 3: Percentage of distribution of tweets

<table>
<thead>
<tr>
<th>Vaccine</th>
<th>Positive (%)</th>
<th>Negative (%)</th>
<th>Neutral (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sputnik V</td>
<td>2.35</td>
<td>2.1</td>
<td>95.55</td>
</tr>
<tr>
<td>Moderna</td>
<td>11.3</td>
<td>11.9</td>
<td>76.8</td>
</tr>
<tr>
<td>Covaxin</td>
<td>8.85</td>
<td>3.35</td>
<td>87.8</td>
</tr>
</tbody>
</table>
From Table 2, a total of 47 tweets are regarded as positive, 42 as negative and 1911 as neutral for Sputnik V. 226 tweets are regarded as positive, 238 as negative and 1536 as neutral for Moderna. 177 tweets are regarded as positive, 67 as negative and 1756 as neutral for Covaxin.

From Table 3, Moderna has highest percentage of positive tweet distribution that is 11.3%. Covaxin’s positive tweets distribution of 8.85% which is higher than Sputnik V positive percentage of 2.35%. This translates to the observation that among three vaccines the positive tweets about Moderna are more positive in the magnitude of their sentiment and also indicates that it is most positively talked about other vaccines. Among the Russia based Sputnik V and India’s manufactured Covaxin more people are favoring Covaxin as its positive percentage distribution is more. The highest positive percentage of Moderna may give the inference that among all vaccines it has maximum satisfactory opinions among Twitter users. The highest negative percentage distribution for Moderna which is 11.9 % signifies that out of the total 2000 tweets, 238 negative tweets are found negative. So, its negative percentage is high. Hence Moderna is most positively and negatively discussed vaccine among twitter. These results are based upon observations made on February 2021.However they may vary for another time values [36-41].

The Visualization of observations is performed using Word Cloud. Word cloud contains representation of most occurring words in the tweets. Word clouds (also known as text clouds or tag clouds) work in a simple way: the more a specific word appears in a source of textual data (such as a speech, blog post, or database), the bigger and bolder it appears in the word cloud. The Word Cloud representations of Sputnik V, Moderna and Covaxin are shown in figure 2, figure 3, and figure 4.

**Figure 2**: Word Cloud of Sputnik V

**Figure 3**: Word Cloud of Moderna
6. Conclusion

The paper summarizes the techniques used for Sentiment Analysis. Different methodologies are explored which are applied in business, politics, government decisions, developing AI based products. Through this paper a deep insight to the approaches of SA is dealt with. From the Sentiment Analysis using Twitter Data of vaccines it is observed that U.S based Moderna is most promisingly discussed vaccine among twitter users. It has maximum positive opinions on Twitter than Covaxin and Sputnik V. Among the Russia based Sputnik V and India’s manufactured Covaxin, Covaxin is more positively favored than Sputnik V. The Word Cloud representations provides a simple way in which for the large databases maximum occuring words appears bigger and bolder for the Vaccines. It is helpful in data visualizations. The future scope includes sentiment analysis of videos and subtitles in it. Also, sentiment in different languages could be worked out using machine learning approaches.

7. Acknowledgement

Firstly, I wish to express my most sincere and profound gratitude to Mr. Mohd. Saif Wajid, Mrs. Upasana Dugal, Department Computer Science and Engineering, School of Engineering, Babu Banarasi Das University Lucknow, for giving inspiration and a chance to showcase my capabilities. I am also grateful to them for cooperation in providing all the required resources. I extend special thanks to friends and family members for their constant support.

REFERENCES


