Harnessing the power of ML and NLP for decision making in education sector from social media data

Harshita Murthy¹, Hemraj Lamkuche²*

¹,²Symbiosis Centre for Information Technology, Symbiosis International (Deemed University), Pune, Maharashtra, India

*Corresponding author:
hemraj@scit.edu

Abstract

The active social media users across the globe have passed the 3.8 billion mark. Platforms like Facebook, Reddit, Instagram, Twitter, and more are an ocean full of opinions and views. With over 500 million tweets being generated daily, this enormous volume of data can offer very prominent insights and allow organizations and businesses to make strategic decisions. The COVID-19 pandemic has changed the landscape of learning and education dramatically. With a sudden devi- ation from the classroom in many parts of the world, some wonder if the adoption of online learning will continue with the outbreak of the post-epidemic epidemic and how such a change could affect the global education market. The crux of the problem is how we need to analyze vast amounts of data efficiently. We chose to employ advanced ML and NLP techniques to analyze the sentiment of the masses on digital learning and important extracting demographic information from them. In this paper, we will make an effort to understand the orientation of the academicians towards the recent online education adoption. We will collect the data from the tweets using the trending tags of COVID-19 and Online classes.

Keywords

Social Media, Tweets, Pandemic, NLP, COVID-19, Sentiment Analysis, Education.

1 Introduction

The term ‘Social Media Analytics’ means the science of extracting important hidden information with large amounts of informal and poorly constructed social media data to enable informed and insightful decisions.

A study suggests that the current active social media users across the globe in 2020 is around 3 billion and is predicted to only grow in the coming years. These users are spread across social media platforms such as Twitter, Facebook, Reddit, etc. With the backing up of vast amounts of data at our disposal, it is only smart to design and engineer a tool scientifically to analyze the data containing people's views and opinions and use them to make decisions. In these extraordinary times of the COVID-19 pandemic rattling the world, it becomes extremely important to study and understand the sentiments and behavior of the people across the world engaged in various sectors [1].

One among the various sectors, which have been badly affected by the pandemic, is the Education sector. In this pandemic, how physical teaching has dramatically moved towards virtual learning is noticeable and has resulted in divided opinion among people voicing their views positively, negatively, and some being neutral. With the adoption of online learning, it is important to study and analyze how this major rehaul in education has impacted its people. With the recent development in social media usage, the generated data is being used in multiple areas like politics, sports, health, entertainment, shopping, education, and more. Social media is used by businesses to get insights into customers’ feedback, likings, and more; there are several benefits of Social Media: Communication, Information sharing, Research Media, Business Marketing, Discussion platforms, Reviewing, and more [2].

The Data Analysis applies satirical methods to evaluate data and draw significant insights. Some of the statistical methods used to conclude the data provided include multiple regressions, Logistics regression, ANOVA, etc. The architecture and the tools used for the analysis play a vital role in the analysis outcome [3]. Every type of method requires different forms of data to be fed. Hence, the pre-processing of the data is varied from method to method. Some of the popular transformation and extraction tools used in social media analytics are- IBM’s many eyes, Google trends,
2 Literature review

With increasing attention towards the sentiment analysis in social media, various researches on the users’ attitude are being analyzed using the hash tags such as #Coronavirus Outbreak, #COVID19 using the algorithms Naïve Bayesian Classifier and NLP library nltk for Python [5].

Social media is a great source of information and opinions for many users. It is a great spring of public perception and behavior analysis on social trends. The data can be derived from the APIs and further analyzed in Google Colab notebook using Python [6].

A huge crisis of coronavirus pandemic has invaded the world. People are bursting with panic and mass fear as time progresses, which call for discovering effective analytical methods to extract mass sentiments using machine-learning algorithms by using the abundant twitter data available [7].

With the increase in the ease of internet access, people wanting to learn while earning and in their free time, the scope of online education has widened. It is considered that as time progresses, the majority of the institutions will start providing a full-fledged online course for higher education [8]. Everyone who wants to learn is a potential student. The course assessment designed for an online course is relatively easier to pass. It could be a disservice to involved people. With the drastic increase in social media users, it is considered a very important source of data. Businesses analyze the social view of their product or service. Due to the COVID-19 pandemic, social media is even more active, and hash tag topic searching is efficient and an effective way to analyze data using Twitter API. Further, sentiment analysis can be carried out [9]. A study was conducted on several students, who were asked to attend the classroom teaching, and further tests were conducted. Further, a similar evaluation was on students who attended a fully online course and the key factors- All these factors may not directly result in higher grades in the exams but have a stable GFI-Goodness-of-fit index that makes it a good judgment criterion. It was concluded that classroom teaching was more effective than a fully online course. However, the quality of education of an online course could be improved by developing better note-making practices and a better learning environment [10].

The extraction and usage of geospatial data on Twitter can be used to produce valuable information. The extraction and analysis are done through several stages, including crawling, storing, analyzing, and visualization [11]. The study focuses on exploratory processes and the development and improvement of extraction techniques, steps, and analysis of the geospatial Twitter data. The open-source and the public nature of the data allow easy and effective trend analysis possible [12]. The trend analysis is very effective on the data generated by Facebook and helps in topic analysis. Based on the evaluations, the applied clustering algorithm to detect the trending topics was very greedy. Further usage of algorithms like LDA models, which perform LSA- latent semantic analysis and data tracking for a longer duration, results in a better result; Pablo experimented with multiple variations of the Naïve Bayes classifier to analyze the polarity of the tweets generated by Twitter. The most famous variations are Baseline and Binary. In baseline, the model classified tweets as positive, negative, and neutral. At the same time, in binary, the classification was done only as positive or negative [13].

The data from Twitter has also been used to adopt content recommendations by news firms to understand the topics and interesting news of Twitter users. The type of tweets read, people being followed, and tweets being retreated have been used to recommend the users [14]. High value, high performance, and highly opinionated data are considered important in many cases, which can be obtained by ranking the tweets. Topic recognition has been a very powerful technique in doing this. The two important Web 2.0 technologies - Twitter and RSS - Simple Syndication have been used to generate a good model of real-time news recommendation. Some experiments reveal that opinion retrieval performance is transformed when there are links, retweets, mentions, number of followers taken into the count of a Twitter user [15]. The strength of the opinion is biased while doing market research. The wide scope of social media and the huge volumes of data generated by social media has created the opportunity to use big data analytics. The data generated is a source of valuable information for several businesses. The very famous and advanced big data analytics tools for social media are - Keyhole, Edger, Google Analytics, Buzz Sumo, Brand watch, Agora Pulse, Apache Flume, and more. The most important aspect of social media analytics is doing timely
predictions or live predictions [16]. Being able to do live predictions is of great value to businesses like TV Operators, advertisers, and others. The real data can be used to understand the popularity of the VOD - Video-on-demand services for a broadcaster and content pricing strategy, and content delivery strategy can be formulated and used.

Support Vector Machine (SVM), Bayesian Networks (BN), and Decision Trees (DT) are the most used techniques in the area of social media analytics, according to the research papers and articles between 2003 and 2015. The major domains in which various data mining techniques are used for research and analysis are- Business and Management, Education, Finance, Medicine and Health, Government, and the Public sector [17]. The experiments carried out by M. S. Neethu show that the Naïve Bayes classifier performs best over Maximum Entropy in classification. Pak and Paroubek suggested it, using a model that categorizes tweets as positive, negative, and neutral using Twitter API, based on the Naïve Bayes model and using features like N-gram, similar to the practices of this paper. Emoticons or emesis add up as noise to the input of the Classification Model. It has a huge impact on an SVM classifier, but not much on a Naive Bayes classifier. Sentiment Analysis can be divided into the following tasks

- Subjectivity Classification
- Sentiment Classification

Subjective classification is further divided into Object Holder Extraction and Object/Feature Extraction.

Subjective classification deals with the classification of sentences into opinionated and indifferent/neutral sentences. In contrast, Sentiment Classification deals with the type of emotion- positive, negative, and neutral. Using Twitter data, Cross-lingual sentiment analysis can also be carried. Cross-lingual sentiment classification refers to acquiring the sentimental resources from a resource-abundant language to a language lacking resource. Text blob library available in Python is one of the easiest ways to do sentiment analysis on large data quality. Using the Tweedy library, tweets can be directly downloaded from Twitter.

The rapid propagation of opinions and negative reviews on a product can impact the company or service majorly. There have been multiple cases in which the companies have faced heavy losses in their businesses and services. They have again used social media to do damage control to understand the root of the problem and fix the crises by formalizing and analyzing big data generated by social media. Consumers can use sentiment analysis methods to find out about the products they want to buy reviews on the product’s website can be deceiving. However, Twitter’s analysis of the overall product will give a much-unbiased result. The challenges faced by social media analytics are as discussed below-

- Scrapping: Even though the data generated by the social media platforms is huge, informative, and rich, it is not easily and freely available to businesses, like they are available for research and academics.
- Data Cleaning: The data generated by social media is unstructured, and data cleaning is a tedious process that takes up most of the time.
- Data Analytics: The researchers face many issues while analyzing the data that has been procured, dealing with the foreign words, slangs used short forms of the words, as the nature and form of the language is ever-evolving.

Various ensemble techniques can be used for modeling text data. The ensemble can be done in two methods - Fixed rule ensemble and Meta classifier. In a Fixed rule, ensemble there is a common rule for a combination of results from individual models that include voting rule, addition rule, max rule, and product rule. In the Meta classifier, the outputs from the models, the class labels are observed as the new features being learned. Considering many models and evaluating them based on the sliding window Kappa Statistics, Stochastic Gradient Descent (SGD) based models perform the best when used with appropriate learning function.

Opinion mining on the huge data generated by Social media presents a great opportunity to target specific marketing like age, gender, locality, etc. In broad terms, Opinion Mining is the science of using textual analysis to understand the drivers of social emotions.

3 Methodologies

3.1 Data description and collection

The data was collected from the social media platform Twitter. It was done by using the API Key and Secret Keys provided by the Twitter Developers platform. The tweets are a humongous source of opinions. The tweets have been retrieved using the trending hash tags, which were online education-related and trending. The Twitter developer platform allows users
to openly access APIs, tools, and resources that enable the interested developers to harness the Twitter data. Standard APIs are available to developers whose accounts are accepted by Twitter when provided with a genuine reason for usage. With the standard APIs functions like tweet retrieval, post-retrieval, pull public account information, retrieve trends and follow, search and get users. Premium API is an advanced and paid platform with scalable and advanced functions. Other commercial APIs available are Enterprise API, Ads API, Twitter for websites, Twitter Developer Labs.

3.2 Data preparation and sentiment analysis

All of the Twitter data that was trending and included the hash tag provided was a desirable source. The tweets were downloaded using the Tweedy library in Python into a json file format. All the trending tweets containing the hardcoded hash tag were downloaded and passed on further to sentiment analysis.

The emotion determination behind the words is sentiment analysis; this is done by assigning strength to the words and assigning polarity to positive, negative, and neutral words, automatically carried out by many libraries like Text Blob, NLTK, spacey and more. Here, Text Blob was used. The following flowchart explains the overall steps involved in the sentiment analysis of the tweets as shown in Figure 1.

4 Results and discussion

Twitter is a very rich source of social opinions and thoughts. Daily, the data generated by tweets is very high. Understanding and interpreting the shift towards the likes and dislikes of the people towards online education resulting from the pandemic crises is analyzed using the sentiments displayed by the people as tweets. Hash tags like “#online education #pandemic,” “#online classes failed,” “study from home,” “elearning,” etc., when analyzed, show much opposition towards the now popular method of online education. It indicates that the people are not satisfied with the classes that are conducted online for students.

Digging a little deep into the individual tweets, we can say that people feel that student dedication, interaction, and seriousness are prominent in classroom teaching and is fairly absent in online education.

With the rise in COVID-19 cases and the increase in deaths, people's general viewpoint and conduct are to stay in their safety zones. They prefer staying at home and get done with their commissions, including education. So, we notice that under the hash tags like “Online Classes” and “Online Education,” we find much positive inclination towards online education.

Table 1 shows the compares the three most popular libraries in Natural Language Processing in Python- NLTK, Text Block, and SpaCy. NLTK (Natural Language Toolkit) is used for token making, lemmatization, stemming, parsing, POS marking, etc. This library contains tools for almost all NLP activities. It is one of the oldest and the bulkiest python libraries. NLTK does not analyze/split the sentences according

Fig. 1. Flowchart of various steps involved in sentiment analysis of tweets
to the semantics structure of the text. SpaCy is a great NLTK contestant. These two libraries can be used for the same purposes. It has a reserved number of languages available; it collects the most suitable and quickly usable text analytic tools. Text Blob is a Python library for processing text data. It provides a consistent API for accessing common language processing functions (NLP) such as tagging in the speech part, noun phrase removal, sentiment analysis, and more.

The word cloud was created by using the tweets retrieved under the positive and negative sentiment category also removing the stop words, as shown in Figures 2 and Figure 3. It can be said that people supporting online education say that the online mode of education is safe, easy, effective, flexible, cost-efficient, and comfortable. While others argue that the online education model was inefficient, lacked discipline, was prone to dropouts, difficult, doubts were not cleared much effectively, required immense commitment, interactions were low, and most importantly, a practical way of learning was missing. Not many users agreed on online education being as efficient as classroom teaching.

<table>
<thead>
<tr>
<th>Libraries</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLTK</td>
<td>slow</td>
</tr>
<tr>
<td>TextBlob</td>
<td>fast</td>
</tr>
<tr>
<td>SpaCy</td>
<td>fast</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparing few python libraries</td>
</tr>
</tbody>
</table>

5 Conclusions

Natural Language processing enriches the possibility of better decision-making in a sector. It is a rich source of behavior and sentiment analysis, enabling the decision-makers to understand the needs, demands, and orientation of the users and customers.

Multiple hash tags were used to understand the inclination of public interest towards the recent shift in education. The extracted tweets using the hash tags were categorized into positive, negative, and neutral. After inputting several trending hash tags regularly and noting the percentage record of the tweet types over some time, the negative response percent exceeded the negative and neutral sentiment percentage.

From this, it can be concluded that most people support classroom education over online education. When a few tweets were further analyzed, it was observed that the users found that classroom education was more binding than online education because of various distractions and lack of concentration. Many Twitter users argued that online education is the future in the present situation where social distancing is an important norm.

6 Future scopes

This paper’s scope was to understand people’s opinions to change the trend and method of education using Twitter analysis. The analysis method used can be improved by multiple folds. The algorithm used can be developed to detect sarcastic tweets also and categorize them as negative or positive. Further improvements can be done by making use of schematic analysis of the tweets. Often, the tweets do not include any adjectives, so they are categorized as neutral, even though they are not; by analyzing the schematics of the sentence, they can be properly categorized. Fur-
ther, the platforms like Facebook, Reddit, and Tumblr can be used for an even broader perspective.

Statement on ethical issues

Research involving people and/or animals is in full compliance with current national and international ethical standards.

Conflict of interest

None declared.

Author contributions

The authors read the ICMJE criteria for authorship and approved the final manuscript.

References

3. Chengang Zhu, Guang Cheng (senior member, IEEE) and Kun Wang (senior member, IEEE) Big data analytics for program popularity prediction in broadcast TV industries, (2017)