International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217269

Text and Sentimental Analysis on Big Data

Saifuzzafar Jaweed Ahmed, Prof. Vandana Navle
Department of Computer Engineering, Dhole Patil College of Engineering (DPES), Pune, Maharashtra, India

Article Info
Volume 7, Issue 2
Page Number: 324-334

Publication Issue :
March-April-2021

Article History
Accepted : 12 April 2021
Published : 17 April 2021

ABSTRACT

Big Data has become a very important part of all industries and organizations sectors nowadays. All sectors like energy, banking, retail, hardware, networking, etc all generate a huge amount of unstructured data which is processed and analyzed accurately in a structured form. Then the structured data can reveal very useful information for their business growth. Big Data helps in getting useful data from unstructured or heterogeneous data by analyzing them. Big data initially defined by the volume of a data set. Big data sets are generally huge, measuring tens of terabytes and sometimes crossing the sting of petabytes. Today, big data falls under three categories structured, unstructured, and semi-structured.

The size of big data is improving in a fast phase from Terabytes to Exabytes Of data. Also, Big data requires techniques that help to integrate a huge amount of heterogeneous data and to process them. Data Analysis which is a big data process has its applications in various areas such as business processing, disease prevention, cybersecurity, and so on. Big data has three major issues such as data storage, data management, and information retrieval. Big data processing requires a particular setup of hardware and virtual machines to derive results. The processing is completed simultaneously to realize results as quickly as possible. These days big data processing techniques include Text mining and sentimental analysis.

Text analytics is a very large field under which there are several techniques, models, methods for automatic and quantitative analysis of textual data. The purpose of this paper is to show how the text analysis and sentimental analysis process the unstructured data and how these techniques extract meaningful information and, thus make information available to the various data mining statistical and machine learning) algorithms.

Keywords: Big data, Text analysis, Sentimental Analysis

I. INTRODUCTION

Text analysis is an automatic and quantitative approach for collecting, processing, and interpreting text data. Getting new information from a text analytics process is very complex. It is a chain of operations with strong links to research goals, that can allow to have information in real-time, and help to make a decision in uncertain conditions.

Sentiment analysis of structured and unstructured sources of text, in particular, is evolving at an accelerated pace also text mining is evolving very fast. Disparate areas of knowledge such as pattern recognition, machine learning, as well as Natural Language Processing (NLP) are variously explored and brought together to process unstructured data. There are a couple of sectors of business life where sentiment analysis is finding more and more applications.

A. Big Data

Big data is an evolving phase which means large volumes of both structured, semi-structured, and unstructured data that pose a difficult task to be processed using traditional methods and databases. It is an approach for informed decision-making using analytical techniques to describe any data set that is large enough that requires the use of high-level programming skill and methodologies to make the data into a helpful asset for an organization then they can make the right decisions. There are three types of big data present mostly in text format i.e Structured, Semi-structured, and unstructured.

Such voluminous data can come from several different sources such as business transaction systems, customer databases, mobile applications, websites, machine-generated data, and real-time data sensors used in internet of things (IoT) environments. This comes with complexities commonly known as 8Vs i.e.

1) Volume: This is indicative of the huge data sets created at high-frequency rates.

2) Variety: This deals with the different data types, i.e. structured, semi-structured, unstructured, or all of these.

3) Velocity: This deals with the speed and frequency at which data may be generated by an application.

4) Veracity: This deals with the accuracy, truthfulness of the data, and if it's authentic.

5) Value: This deals with the worthiness of data extracted from various raw data available. Just having data abundance doesn't essentially imply being able to extract usefulness from it.

6)Visualization: Getting a meaningful result by processing big data is only half the work done. Unless it's represented or visualized in a meaningful way, there's no point in analyzing it.

7)Variability: In Big data analysis, data inconsistency is a common scenario that arises as the data is sourced from different sources. Besides, it contains different data types.

8)Validity: Validity has some similarities with veracity. As the meaning of the word suggests, the validity of massive data means how correct the info is for its purpose.

In a Big Data system, data holds all essential to all the knowledge and possibilities of its applications. In fact, Data Quality is most often the reason for any business' data and information problems. The key data dimensions are:

Completeness: Is data missing or not user-friendly?

Timeliness: Is data available for use in the time frame in which it is expected?

Conformity: Is the data conforming to the expected format?

Uniqueness: Is the data duplicated within the available data set?

Integrity: Ensure integrity of data and its relationships along with source or lineage of the data. Is the integrity ensured?

Consistency: Is there a single source of truth or are different versions for the same data entity available across multiple environments?

Accuracy: Is the data accurately representing the business data as expected?

B. Text Analysis

Text analysis is a machine learning technique that allows organizations to automatically understand and extract meaningful information from text data from any social platform, such as tweets, comments, emails, support tickets, product reviews, and survey responses. Text Analysis is defined as the process of

extracting implicit information from textual unstructured data. Because the implicit knowledge which is the output of text Analysis does not exist in the given storage, it should be distinguished from the information which is retrieved from the storage. The text classification, clustering, and association are the typical tasks of text Analysis, and they are covered in the subsequent chapters, in detail. Text Analysis is the special type of data Analysis, and other types such as relational data Analysis, web Analysis, and big data analysis. Therefore, this section is intended to explore the overview of text Analysis, before mentioning the tasks of text Analysis and the types of data Analysis.

Text is defined as the unstructured data which may consist of strings, special character, and numbers which are called words. Even if the collection of strings or words belongs to text in the broad view, it requires the meanings of individual strings and the combination of them by rules, called grammars, for making the text. Here, the scope of the text is restricted to the article which consists of paragraphs or a set of paragraphs and is written in a natural human language. We assume that a paragraph is referred to as an organized set of sentences and a text is an ordered group of paragraphs.

Nowadays, people express their opinions, feelings, and feedback through sharing images, tweeting, commenting on the social platform. The huge amount of user-generated content on any social platform gives opportunities for understanding social behavior and building socially intelligent systems to investigate and extract information with text analysis methods from social media data that may have any form of a tweet, image, or comment. Text mining aims to find meaningful information contents, topics, word relations, and patterns from the text data.

C. Sentimental Analysis

Sentiment analysis (SA) is the task of determining the emotion of a writer based on their written texts by considering the polarity of keywords utilized in the writing. The polarity of keywords might be positive, negative, or neural, for instance by using words like ‘happy’, ‘angry’, ‘sad’ or ‘indifferent’. SA is done by detecting the contextual polarity of documents using semantic orientation technique and machine learning. It can be limited to classifying positive and negative sentiments, or rate emotional levels. For instance, a case study of movie review was utilized in which reviews were rated on the appreciation of the movie from one to five.

Besides the categorisation of sentiment into positive, neutral and negative, it may further be considered as a range of emotions, thereby expanding the degree of sentiment into many factors (see below figure). Benjamin Graham, the father of “Value Investing”, describes the emotions of investors as a pendulum that swings between optimism and pessimism (shown by fear and greed in Figure 1.1). This is backed by research from behavioral finance that detects systematically biased trading behavior when emotional responses are triggered from new information.

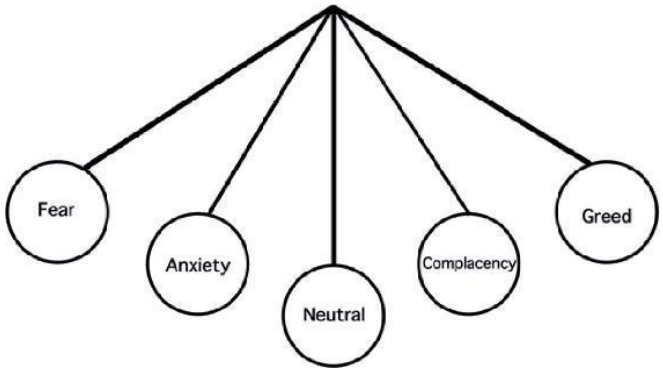

Figure 1. Sentimental Analysis

II. ANALYSIS PROCESS

A)Text Analysis

The steps of a text analysis process, albeit strongly linked to the objectives of the analysis, elaborate a process of chain analysis that's expressed in several macro- phases.

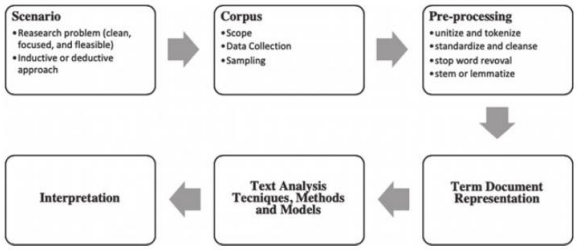

Figure 2. Steps of a text analysis process

The first step in the text analysis process is characterized by the definition of the research problem. The project, in fact, should be clear, well-focused, and versatile. In this phase, it is necessary to specify the population under study, defined deadlines and goals. If an inductive approach is applied, it is appropriate to examine any external influences such as backgrounds and theories. In the deductive approach, the existence of information is often used to test hypotheses, benchmark, and build models.

In the second step, the corpus is generated, collecting a set of documents which is also called as data collection. There- fore, many aspects related to the objectives and the sample collection will be clarified at this stage.

The third step is dedicated to pre-processing, which is extrinsic in a set of sub-phases dependent on the type of document and the aims of the analysis. A corpus that comes from a social survey with open questions, or a corpus that presents a collection of Tweets will get "the need" for a very different treatment.

In a corpus from a survey, pretreatment requires a reasonably traditional simple standardization, with the removal of punctuation, numbers, then it'll be chosen whether to carry out a morphological normalization like lemmatization. Moreover, any meta information from surveys can eventually support the building of a lexical table. In the case of a corpus coming from social media, there will be emojis and emoticons and many special characters that can be removed or incorporated in the analysis from the document or from data.

In the last step, the pre-processed text document is first transformed into an inverted index, which, within the fourth step, is transformed into a term-document matrix (TDM) or document-term matrix (DTM). In a TDM, the rows correspond to terms, and therefore the columns correspond to documents. Alternatively, in a DTM, the rows correspond to documents, and therefore the columns correspond to terms. Local, global, and combinatorial weighting are often applied to TDM or DTM.

In the fifth phase, methods, techniques, and models supported goals are applied. The foremost used techniques within the exploration phase are the latent semantic analysis (LSA) and therefore the correspondence analysis of lexical tables (LCA), and clustering techniques. For the research of the topics, a widespread probabilistic approach is that the application of the probabilistic topic models. In recent few years, machine learning classification techniques and sentiment analysis models have been developed and used to analyse unstructured data. The analysis phase is followed by the interpretative phase, which allows us to generate knowledge. This sixth or last phase of the supply chain can end the process or start another one if the results are not satisfactory.

B) Sentimental Analysis

The steps involved in sentiment analysis are often explained with the assistance of a flowchart, as shown below.

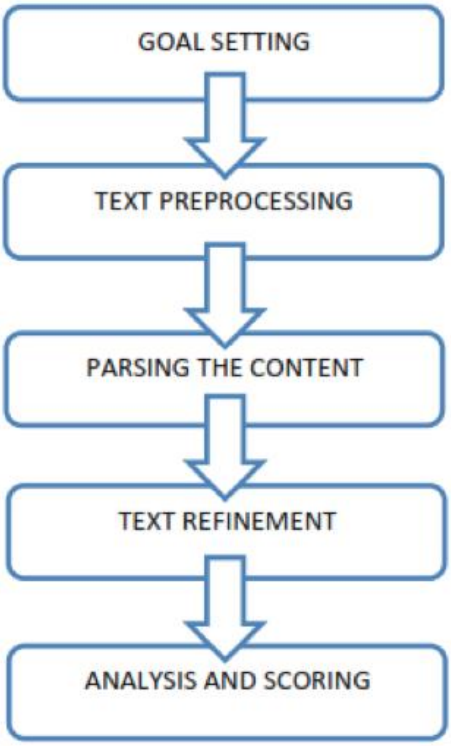

Figure 3. Steps of a sentimental analysis process

- Firstly, the goal of this process is to be set, which incorporates determining the sentiment analysis goal and therefore the scope for the text content.
- Secondly, text processing has to be done on input text which involves determining the source i.e. whether you're taking the data or files from the web, micro-blogging site, etc. The text then has to be loaded to the processing system (the system technique to be used for the analysis), this process deletes unwanted words from the text and organizes the emotional symbols that folks use in texts into words. Also, it observed that to precise strong sentiments, uppercase alphabets are used (such as OUTRAGEOUS!)
- Then, in the third step it parses the content which involves segmenting the words based on their polarity and tagging the parts of speech used (adjective, noun, etc.) after that identifying the terms.
- In the fourth step of the process to ensure the correct analysis text refinement should be, that is finding the stop words and synonyms, etc.
- The fifth and last step is analysis and scoring: this step involves finding the sentiments bearing phrases from the data and scoring them. Scoring is that the process during which the intensity of the sentiment is analyzed. An example for scoring is shown within the table below:

SCORE	TEXT
2	You're awesome and I love you
-5	I hate and hate and hate. So angry. Die!
4	Impressed and amazed: you are peerless in your achievement of unparalleled mediocrity.

Table 1. Sample scores with sentiments

III. METHODS AND TECHNIQUES

A) **Text Analysis**

There are some basic and more advanced text analysis techniques, each used for different purposes

1) Word Frequency

Word frequency is a text analysis technique or method that measures the most frequently occurring words or concepts in a given text using the numerical statistic TF-IDF(term frequency-inverse document frequency).

You might apply this method to research the words or expressions customers use most often in support conversations. For example, if the word 'delivery' appears most frequently during a set of negative support tickets, this might suggest customers are unhappy together with your delivery service.

2) Collocation

Collocation helps identify words that commonly co-occur. For example, in reviews of customers on a hotel booking website, the words 'air' and 'conditioning' are more likely to co-occur rather than

appear individually. Bigrams (two adjacent words e.g. 'air conditioning') and trigrams (three adjacent words e.g. 'out of home' or 'will come again') are the most common types of collocation you'll need to look out for.

Collocation analysis can be used to identify hidden semantic structures and improve the granularity of the insights by counting bigrams and trigrams as one word.

3) Concordance

The concordance method helps to identify the context and instances of words or a set of words. For example, below is the concordance of the word "simple" in a set of reviews of the app:

Reference	Preceding Context	Target	Following Context
Review 1, sentence 1	Hate the new update.	Simple	as that.
Review 2, sentence 4	It's quite good and	simple	to use.
Review 3, sentence 1	It's also slow and laggy. Takes a few seconds to send a	simple	message.
Review 4, sentence 2	Love it! Super	simple	and easy to use.

In this example, the concordance of the word "simple" can give us a fast grasp of how reviewers are using this word. It also can be wont to decode the anomaly of the human language to a particular extent, by watching how words are utilized in different contexts, as well as being able to analyze more complex phrases. Now that we've touched upon the essential techniques of text analysis, we'll introduce you to the more advanced methods: text classification and text extraction.

3) Text Classification

Text or document classification is the process of assigning text documents into one or more classes or categories, assuming that we have a predefined set of models to classify the document in its class. To make the process more efficient and faster, we can consider automating the task of text classification, which brings us to automated text classification. To automate text classification, we can make use of several ML techniques and concepts. There are mainly two ML techniques that are used to solving this problem:

- Supervised machine learning
- Unsupervised machine learning

Unsupervised learning is a specific machine learning technique or algorithm that does not require any pre-labeled training data samples to build a model or to train the model. We usually have a collection of data points, which could be any type like text or numeric, depending on the problem we are trying to solve. The process retrieves features from each of the data points using a process known as feature extraction and then feeds the feature set for each data point into our algorithm. This method tries to extract meaningful patterns from the data, such as trying to group similar data points using techniques like clustering or summarizing documents based on topic models. This is very useful in text document categorization and is also called document clustering, where we cluster documents into groups purely based on their features, similarity, and attributes, without training any model on previously labeled data.

Supervised learning is a specific machine learning technique or algorithms that are trained on pre-labeled data samples known as training data which will be used in training the machine learning model. In this method features or attributes are extracted from this data using feature extraction, and for each data point, we will have its own feature set and corresponding class/label. The algorithm learns various patterns for each type of class from the training data which is applied as input for training the model. Once this process is complete, we have a trained model by giving data as training data. This model can then be used to predict the class for future test data samples when we pass it to the model. Thus the machine or model has learned, based on previous training data samples.

There are two types of supervised learning algorithms:

- **Classification:** It is a process of supervised learning that is referred to as classification when the outcomes to be predicted are distinct categories, thus the outcome variable is a categorical variable in this case. Examples would be news categories or movie genres.
- **Regression:** Supervised learning algorithms are known as regression algorithms when the outcome we want to predict is a continuous numeric variable. Examples would be house prices or people's weights.

4) Text Extraction

Text extraction is another most used text analysis technique that extracts pieces of data that already exist within any given text. one can extract things like keywords, prices, company names, and product names and specifications, product reviews, and more.
You can auto-populate spreadsheets with this data or perform the extraction in concert with other text analysis techniques to categorize and extract data at the same time.

6) Clustering

Text clusters are a group of vast quantities of unstructured data. Clustering algorithms are very fast to implement because you don't need to tag examples to train models but it's less accurate than classification algorithms. That means these smart algorithms retrieve information and make predictions without the use of training data, otherwise known as unsupervised machine learning.
Google is a good example of how clustering works. When you search for a term on Google search engine, have you ever thought about how this search engine takes just seconds to pull up relevant results? The algorithm breaks down unstructured data from web pages and groups pages into clusters around a set of similar words or n-grams (all possible combinations of adjacent words or letters in a text). So, the pages from the cluster that contain a higher count of words or n-grams relevant to the search query will appear first within the results and the marching or relevant result also.

B) Sentimental Analysis

Sentiment analysis is also known as opinion mining. This analysis uses natural language processing (NLP) and machine learning algorithms, to automatically determine the emotional tone behind online conversations or to analyze a body of text for understanding the opinion expressed by it and other factors like mood and modality. There are different algorithms you can implement in sentiment analysis models, depending on the scenarios like how much data you need to analyze or how accurate you need your model to be.

1). Rule-based Approaches

A rule-based technique uses a set of manually created rules to help identify subjectivity, polarity, or the subject of an opinion.
This set of manually created rules may include various NLP techniques developed in linguistics, such as:

- *Stemming*, *tokenization*, *part-of-speech tagging*, and *parsing*.
- Lexicons (i.e. lists of words and expressions).

Below is some basic example of how a rule-based system works:

1. Defines two lists of polarized words (e.g. negative words like not good, worst, ugly, bad, etc and positive words like very good, best, beautiful, super, good, etc).
2. This method counts the number of positive and negative words that appear during a given text.

3. If the number of positive word appearances is bigger than the number of negative word appearances, the system returns a positive sentiment and the other way around. If the numbers are even, then this method will return a neutral sentiment.

Rule-based systems are very naive since they do not take under consideration how words are combined during a sequence. Of course, more advanced processing techniques are often used, and new rules are added to support new expressions and vocabulary. However, adding new rules may affect previous results, and therefore the whole system can get very complex. This method needs regular investments because rule-based systems often require fine-tuning and maintenance.

2) Automatic Approaches

Automatic methods, contrary to rule-based systems, These automatic methods don't rely on manually created rules, but on machine learning techniques. A sentiment analysis task is usually modeled as a classification problem, whereby a classifier model is fed a text as input and returns a category, e.g. positive, negative, or neutral.
Here are the steps how a machine learning classifier

can be implemented :

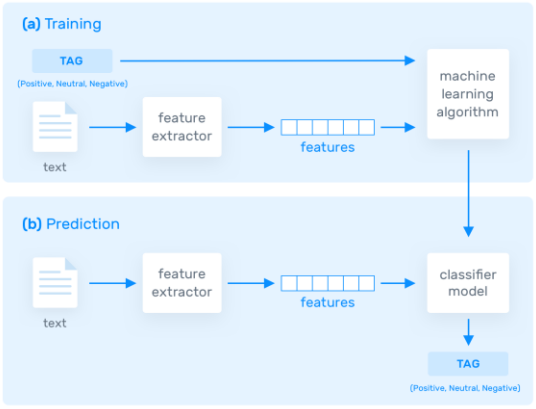

Figure 4. The Training and Prediction Processes

a)The training process: In the training process set of data points is used to train a model. The model learns from the sample input (i.e. a text) to give corresponding output (tag) based on the test samples used for training. The feature extractor does a very important task i.e to convert text input into a feature vector to represent numeric or symbolic characteristics. The machine learning model is generated by feeding Pairs of feature vectors and tags (e.g. *positive*, *negative*, or *neutral*) into the machine learning algorithm.

b)The prediction process: "Prediction" refers to the output of an algorithm after it has been trained from some old records or data set and applied to new data for forecasting. The sample test dataset or documents go through the same process of feature extraction and normalization. The test document features are passed to the trained Classification Model, which predicts the possible class for each of the documents based on previously learned patterns in the training process. You can evaluate the prediction performance of the model by comparing the true labels and the predicted labels using various metrics like accuracy if you have the true class labels for the documents that were manually labeled, . This would give an idea of how well the model performs its predictions for a new set of documents.

The feature extractor is used to transform unseen text inputs into feature vectors. These feature vectors are then fed into the model, which generates predicted tags (again, *positive*, *negative*, or *neutral*).

Text normalization:

Text normalization is defined as a process that includes a series of steps that should be followed to wrangle, clean, and standardize textual data into a form that would be consumed by other NLP and analytics systems and applications as input. Often tokenization itself is also a neighborhood of text normalization. Besides tokenization, There are various other techniques including cleaning text, correcting spellings, case conversion, removing stop words and other unnecessary terms, stemming, and lemmatization. Text normalization is additionally often called text cleansing or wrangling.

Feature Extraction:

We will use a generic function here to perform various types of feature extraction from text data. The types of features which we will be working with are as follows:

The first step in a machine learning text classifier is to transform the text extraction or text vectorization for the machine learning model, and the classical approach has been bag-of-words or bag-of-n-grams with their frequency. More recently, new feature extraction techniques have been applied based on word embeddings (also known as *word vectors*). This kind of representation makes it possible for words with similar meanings to have a similar representation, which can improve the performance of classifiers.

Following is the example of feature extraction from text documents.

```
from sklearn.feature_extraction.text import
CountVectorizer, TfidfVectorizer def
build_feature_matrix(documents,
feature_type='frequency'):
feature_type = feature_type.lower().strip()
if feature_type == 'binary':
vectorizer = CountVectorizer(binary=True,
min_df=1,
ngram_range=(1, 1))
elif feature_type == 'frequency':
vectorizer = CountVectorizer(binary=False,
min_df=1,
ngram_range=(1, 1))
elif feature_type == 'tfidf':
vectorizer = TfidfVectorizer(min_df=1,
ngram_range=(1, 1))
else:
raise Exception("Wrong feature type entered.
Possible values: 'binary', 'frequency', 'tfidf'")
feature_matrix =
vectorizer.fit_transform(documents).astype(float)
  return vectorizer, feature_matrix
```

IV. CLASSIFICATION ALGORITHMS

The classification algorithms usually involve a statistical model like Naïve Bayes, Logistic Regression, Support Vector Machines, or Neural Networks:

- Naïve Bayes: a group of probabilistic algorithms that uses Bayes's Theorem to predict the category or class of a text.
- Linear Regression: a very popular algorithm in statistics used to predict some value (Y) given a set of features (X).
- Support Vector Machines: a non-probabilistic model which uses a representation of text examples as points during a multidimensional space. For example, consider different categories (sentiments) are mapped to distinct regions within that space. Then, new texts are assigned a category based on similarities with existing texts and the regions they're mapped to.
- Deep Learning: a various set of algorithms that plan to mimic the human brain, by employing artificial neural networks to process data.

V. BENEFITS OF BIG DATA PROCESSING

The ability to process Big Data brings in various benefits for business, such as-

- Businesses can utilize outside intelligence while taking decisions

Access to social data from search engines and sites like Facebook, Twitter is enabling organizations to figure out their business strategies by analysing the data.

- Improved customer service

Traditional customer feedback systems are getting replaced with Big Data technologies and methods. In these new big data technologies, Big Data and natural language processing technologies are being used to read and evaluate consumer responses and feedback of the service or product because the user response is very important for business growth.

- Identification of risk to the product/services on the early stage, if any
- Better operational efficiency

Big Data technologies can be used for creating a staging area or landing zone i.e the storage area of a company for new data before identifying what data should be moved to the data warehouse. In addition, such integration of massive Data technologies and data warehouses helps a corporation to dump infrequently accessed data.

VI. APPLICATIONS

A) Text Analysis

- Search access of unstructured data.
- Email Spam filter
- Automatic ads placement
- Social media monitoring

B) Sentimental analysis

- Social Media Monitoring
- Brand Monitoring
- Voice of customer (VoC)
- Customer Service
- Market Research

VII. CONCLUSION

We discussed processes and techniques to extract meaningful information from the unstructured text through text analysis and Sentimental analysis thus, make the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. Text analysis techniques provide you the insight of information from the large size of data, Which may be helpful for business growth. Sentimental Analysis provides the ability to analyze the opinions of people for a particular product or for a company. Prediction of the stock market is really a hard nut to crack and requires a lot of effort. The market data if analyzed in a proper way can be very effective in predicting a company's future. We have mined data and trained a machine learning model to predict the answers of the input text.

VIII. REFERENCES

[1]. Atharva Patil, Nishita S. Upadhyay, Karan Bheda,Rupali Sawant “Restaurant’s Feedback Analysis System using Sentimental Analysis and Data Mining Techniques”, Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India

[2]. Manasee Godsay “The Process of Sentiment Analysis: A Study“ ,International Journal of Computer Applications (0975 – 8887) Volume 126 – No.7, September 2015

[3]. Ashish Juneja, Nripendra Narayan Das ”Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application”,2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019

[4]. Surya Prabha PM, Subbulakshmi B “Sentimental Analysis using Naive Bayes Classifier”, 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)

[5]. Sunil Kumar Khatri, Ayush Srivastava “Using Sentimental Analysis in Prediction of Stock Market Investment”, 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Sep. 7-9, 2016,AIIT, Amity University Uttar Pradesh, Noida, India

[6]. Dipanjan Sarkar, “Text Analytics with Python : A Practical Real-World Approach to Gaining Actionable Insights from your Data”,ISBN-13 (electronic): 978-1-4842-2388-8

[7]. Gautam Mitra, Xiang Yu “HANDBOOK OF SENTIMENT ANALYSIS IN FINANCE” ,ISBN-1910571571, 9781910571576

Cite this article as :

Saifuzzafar Jaweed Ahmed, Prof. Vandana Navle, "Text and Sentimental Analysis on Big Data", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 2, pp. 324-334, March-April 2021. Available at

Doi : https://doi.org/10.32628/CSEIT217269

Journal URL : https://ijsrcseit.com/CSEIT217269