Grenze International Journal of Engineering and Technology, Jan Issue

A Review of Various Sentiment Analysis Techniques, Methodologies and their Applications

Romit Bhalla[1], Prof. Sourabh Kumar Jain[2] and Prof. Praveen Kaushik[3]
[1-2]Gyan Ganga Institute of Technology & Sciences, Jabalpur, India
Email: bhalla.romit6876@gmail.com, sourabhkumarjain@ggits.org
[3]M.A.N.I.T., Bhopal, India
Email: praveenkaushikmanit@gmail.com

***Abstract*—To be able to analyse and predict human emotions and sentiments with highest possible accuracy has been a hot topic both in the field of automated computing as well as psychology. The integrated process of analysing the emotions and sentiments of an individual through his/her speech, text, expressions, actions etc. is termed as Sentiment Analysis. With #MENTALHEALTHAWARENESS and #LOCKDOWN trending worldwide, Sentiment Analysis has gained a lot of significance in recent time. Technically, Sentiment Analysis is the synonym for Opinion Mining. This paper provides an overview of the sequential process of Opinion Mining as well as a review of various Opinion Mining techniques, methodologies and their applications considering 'text' as the mode of expression for an individual.**

***Index Terms*— automated computing, psychology, Sentiment Analysis, Opinion Mining.**

I. INTRODUCTION

The 2020 COVID19 pandemic, world-wide lockdown, abrupt increase in the number of employees working from home and the nation-wide ban on a bunch of apps by the government organizations of several countries have brought about a drastic change in the statistics of social media. According to [1]., social media users are now spending an average of 2 hours and 24 minutes per day multi-networking across an average of 8 social-networking and messaging apps which is 43% more compared to pre-pandemic scenario. The same report stated that out of 4.57 billion active internet user, 3.96 billion are active social media users with FACEBOOK surpassing all the other social media platforms in terms of usage and popularity.

With around 350 million new internet users coming online within the last 12 months, the amount of data being generated from the internet as the primary source is no less than trillions of gigabytes per day. With such huge amount of data being generated every day, the biggest challenge laying around the corner is handling and management of this data. The process of collecting, managing, storing and visualizing this enormous amount of data being generated at a pace faster than that of normal data is termed as **BIG DATA ANALYTICS.**

Along with the challenges associated with the data mentioned above, there are few more questions that arise periodically while dealing with such enormous amount of online data generated (also known as BIG DATA). For example: Is the entire data being generated relevant and significant? If not, then what is amount of data that actually make sense or is significant? How to remove noise from the relevant data? Does the relevant data share the same characteristics? If not, how do we model the relevant data out of the BIG DATA? The questions mentioned above along with many others need to answered, tackled and solved.

Grenze ID: 01.GIJET.7.1.7

One recommended solution is to process this human-generated data in human-understandable form only and such kind of processing is termed as **NATURAL LANGUAGE PROCESSING.** Formally, [2]. defined NATURAL LANGUAGE PROCESSING(NLP) as *a branch of artificial intelligence* that deals with the reciprocity between computers and humans using natural language. He further added that the ultimate objective of NLP is to read, construe, realize and make sense of the human languages in a manner that is valuable by relying on **MACHINE LEARNING** to derive meaning out of human languages.
This paper deals with one of the major applications of NLP i.e. Sentiment analysis or Opinion Mining. Ref. [3]. defined **SENTIMENT ANALYSIS** as *the field of study that scrutinizes people's opinions, sentiments, evaluations, appraisals, attitude and emotions towards units such as products, services, organizations, individuals, issues, events, topics and their attributes*. He further added that Sentiment Analysis represents a large problem space. The umbrella of Sentiment Analysis incorporates within itself several other tasks like *opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining etc.*
The entire paper is divided into 4 sections: Section I gives a brief INTRODUCTION to the background of NLP and SENTIMENT ANALYSIS. Section II describes the entire process of Sentiment Analysis. SECTION III reviews the fields of applications of Sentiment Analysis. SECTION IV provides a valid and logical conclusion to the paper and the presented arguments.

II. Process of Sentiment Analysis

The stages of Sentiment Analysis process are shown in the Figure 1.

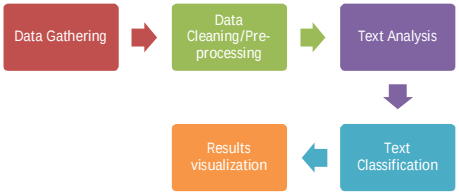

Figure 1. Process of Sentiment Analysis

A. Data Gathering

The process of gathering or fetching out data from the world wide web isn't as easy as it seems to be. One of the prominent reason behind this complexity is non-uniformity among the type of data. Statistics reveal that structured form of data form only 21% of the entire data generated by the internet. The rest 79% comprises of unstructured and semi-structured data which in turn are difficult to process and comprehend.
Hence, the primary goal of DATA GATHERING stage is to fetch data present in various forms and bring about the uniformity in that data required for its further processing and this entire process is technically termed as 'WEB SCRAPING'. Ref. [4] defined web-scraping/screen scraping/web data extraction/web harvesting as a technique used to extricate large amount of data from websites and store it to a local file in a computer or to a database in a table(spreadsheet) format.
Some of the most widely used web-scrapping tools/libraries/software are described below:
a. BeautifulSoup: Ref. [5] described BeautifulSoup as a Python Library for drawing data out of HTML and XML files. It works along with a parser to provide vernacular ways of steering, searching and modifying the parse tree saving programmers' efforts and time.
b. Tweepy: Ref. [6] defined Tweepy as an open source Python package that provides a very favourable way to access the Twitter API with Python. Tweepy incorporates set of classes and methods that represent Twitter's models and API endpoints to transparently mange various implementation details such as Data Encoding and Decoding, HTTP Requests, Results Pagination, OAuth authentication, Rate Limits and Streams dealing with low-level details like data serialization.

B. Data Cleaning/ Pre-Processing

Once the data-set is fetched out, the next crucial task to perform is to clean or pre-process the data. Data cleaning/pre-processing is necessary to be performed over a uniform data set in order to make data complete, unambiguous, normalized, channelized and relevant to the topic. Data cleaning itself is a multi-stage process. The various steps to clean/pre-process the gathered data are described below:

a. Tokenization: The process of splitting a large text/document, often called as *corpus,* into smaller constituent entities is called as tokenization. A corpus, most commonly, is either tokenized into sentences or words. Figure 2 and Figure 3 illustrate the working of a sentence tokenizer and a word tokenizer respectively.

```
import nltk

from nltk import sent_tokenize, word_tokenize

example_text= "Hello Mr. Davis, it is so good to see you. How are you?"
print(sent_tokenize(example_text))
```

```
['Hello Mr. Davis, it is so good to see you.', 'How are you?']
```

Figure 2. A Python code snippet illustrating Sentence Tokenizer

```
import nltk

from nltk import sent_tokenize, word_tokenize

example_text= "Hello Mr. Davis, it is so good to see you. How are you?"
print(word_tokenize(example_text))
```

```
['Hello', 'Mr.', 'Davis', ',', 'it', 'is', 'so', 'good', 'to', 'see',
'you', '.', 'How', 'are', 'you', '?']
```

Figure 3. A Python code snippet illustrating Word Tokenizer

In some rare applications, a corpus is often broken down into *grams* namely bi-grams[1], tri-grams[2] and N-gram[3].

b. Removal of Stop Words: Words that are generally used as fillers and do not add any meaning to the Sentiment Analysis process, in other words, they do not have any sentiments attached to it are called as stop words. Removal of stop words do not change the meaning of the text but instead filter out the irrelevant part from the text. Each natural language (for example: English, Spanish etc.) has its pre-defined set of stop words which keeps on updating periodically. An illustration of Stop Word Removal process is shown in Figure 4.

c. Stemming: The process of extracting out the 'stem' or the root word from one of its given form is called as stemming. The importance of stemming in the data cleaning process is based upon the fact that at a word affixed with prefix or suffix has the same contribution in terms of sentiments as that of its root word. For example: words like writes, writing, written, wrote etc. have the same contribution to sentiments as that of the word 'write'.

d. POS Tagging: The process of labelling each word of a text with its corresponding *Part of Speech* is termed as *Part of Speech Tagging* or *POS Tagging*. In *Python*- each part of speech is assigned an abbreviation some of which are listed in Figure 6.

e. Chunking: The process of grouping of words based upon subject similarity is called as *chunking.* In NLP, a complex sentence might refer to more than one subject or object. *Chunking* helps in grouping words that refer to the same subject or object resulting in simplified processing further. Ref. [7] defined chunk patterns as *normal regular expressions which are modified and designed according to particular sequences of part-of-speech tags.*

f. Chinking: The process of excluding a group of words or chunks matching a sequences of part-of-speech tags from a collection of chunks is termed as chinking. For example: it might be required to chink/exclude all nouns starting with the alphabet 'r' from a chunk containing all nouns

g. Named Entity Recognition: The process of labelling *named entities* in a piece of text with their respective entity tag is called as *named-entity recognition.* Few named entities recognized by *Python* are ORGANIZATION (Example: WHO), PERSON (Example: Narendra Modi), LOCATION (Example: Times Square), DATE (Example: 04-06-1996) etc.

h. Lemmatization: Lemmatization process is very much similar to that of stemming however, the end result of lemmatization may not always be the root word of the given form of word but can also be a synonym to it, adding the same value of sentiments to the text as the root word. For example: 'better' as an adjective can be lemmatized to 'good'. An example of Lemmatization is illustrated in Figure 5.

```
import nltk

from nltk import sent_tokenize, word_tokenize
from nltk.corpus import stopwords

example_text= "Hello Mr. Davis, it is so good to see you. How are you?"
stopwords=set(stopwords.words("english"))

set_example_text=set(word_tokenize(example_text))

print(set_example_text - stopwords)
```

```
{'How', 'Davis', '?', 'see', 'Mr.', 'good', 'Hello', '.', ','}
```

Figure 4. A Python code snippet illustrating Stop Words filtering

```
import nltk

from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

print(lemmatizer.lemmatize("children"))
print(lemmatizer.lemmatize("cacti"))
print(lemmatizer.lemmatize("mice"))
```

```
child
cactus
mouse
```

Figure 5. A Python code snippet illustrating the process of Lemmatization

[1] A *Bi-gram* is a pair of consecutive tokens
[2] A *Tri-gram* is a triplet of consecutive tokens.
[3] A N-gram is a collection of 'N' consecutive tokens

i. Handling Hashtags: Ref. [8] defined hashtag as a word or phrase preceded by the '#' sign that cast-off on social media to tag posts as a part of large conversation or topic. Hashtags play a vital role in terms of engagements and audience for the contents posted on social media. According to the statistics by [9], Twitter reports 125 million hashtags and 500 million GIFs every day, with only 1 hashtag in a tweet it is 69% more likely to be re-tweeted. Hence, we can easily conclude that hashtags do play a pivotal role in the process of Sentiment Analysis as most of the hashtags are considered to have an opinion or an emotion attached to it. Ref. [10] used hash-tagged dataset(HASH) to take a supervised learning approach using the existing hashtags in twitter data for building training data for new upcoming trending hashtags.

j. Handling Emoticons: Ref. [8] defined emojis or emoticons as an assortment of tiny graphics used in digital channels from text messages to social media. They are assembled using characters on the standard keyboard. Emoticons have now become an integral part of digital communication especially on social media. Each emoticon, as the name suggests, depicts an emotion, sentiment and sometimes even a complete message. Ref. [11] handled emoticons by creating a dictionary containing emojis, smileys and various other emoticons by converting them into text according to the emotional state they conveyed. Each of these states then substituted the emoticons appropriately. For example: :'(represented 'crying'. Ref. [12] utilized VADER[4], which is trained to perform sentence-level sentiment analysis on short, micro-blogging contexts, particularizing in emojis, online acronyms, and commonly used slangs on social media.

C. Text Analysis

The process of measuring out the polarity of a piece of text is termed as text analysis or sentiment analysis. Sentiment Analysis brings out the hidden emotions, sentiments and opinion polarity behind a given clean and pre-processed piece of text. The polarity of a text can be *positive, negative or neutral* and can convey emotions like *love, joy, surprise, anger, fear, sadness etc.*

Some of the most widely used libraries/modules along with their key features, used for Sentiment Analysis task have been described below:

a. AFINN: Ref. [13] defined AFINN *lexicon* **(**meaningful units in a language) as a list of English terms manually rated from integers between -5(negative) and +5(positive) by [14]. Ref. [15] proposed a method to calculate a score, named as '*strength score*' in which they first split the sentence by words and then searched the AFINN list. For each word in the sentence, if the word was also present in the AFINN list, they accumulated its score to a variable (initialized with '0') and if not, they considered its score to be 'zero'. If the final '*strength score*' of the sentence turned out to be *greater than zero*, it was said to possess '*positive polarity'*, while if the '*strength score*' of the sentence turned out to be *less than zero*, it was said to possess '*negative polarity*' and '*neutral polarity*' if the '*strength score*' summed up to '*zero*'.

b. SENTIWORDNET 3.0: an improvisation over SENTIWORDNET 1.0, designed by [16]., SENTIWORDNET 3.0 *is an augmented lexical resource explicitly devised for supporting sentiment classification and opinion mining applications*. Ref. [16] further define *SENTIWORDNET* as *a result of automatic elucidation of all the 'synsets*[5]*' of WORDNET according to the conviction of 'positivity', 'negativity' and 'neutrality'. Each synset 's' is kindered to 3 numerical scores Pos(s), Neg(s), and Obj(s) which indicate how positive, negative and objective(neutral) the terms contained in the set are. Each of these scores ranges in the interval [0.0,1.0], and their sum is 1.0 for each synset.* This aligns to the fact that each word may denote some level of positivity in a certain usage scenario, some level of negativity in some other usage scenario and the same applies for the objectivity of the word.

c. TEXTBLOB: Ref. [17] defined TextBlob as *a Python(2 and 3) library for processing textual data by providing a simple API for dealing with common natural language processing(NLP) tasks such as POS tagging, noun phrase extraction, sentiment analysis, classification, translation and more.* Ref. [17] further mentioned that *the sentiment property returns a named tuple of the form Sentiment (polarity, subjectivity). The polarity score is a float within a range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.* Also the value -1.0 of the polarity indicates '*extremely negative sentiment*', 0 indicates '*neutrality*' while +1.0 indicate '*extremely positive sentiment*'.

d. SENTISTRENGTH: This algorithm by [18]. can *analyse sentiments of 16000 words per second up to human level accuracy in English and many other languages*. SentiStrength estimates the strength of positive and negative sentiments in short texts and even works well for informal texts. The sentiments of a text are estimated in the scale of [-5, 5], where -1 denotes not negative, -5 denotes extremely negative, 1 denotes not positive and

[4] VADER is a tightly designed Rule-Based model for Sentiment Analysis of Social Media text designed by [38].

[5] Set of all words denoting the same concept or idea

+5 denotes extremely positive. SentiStrength can also report results in binary(positive/negative) and ternary(positive/negative/neutral).

e. SENTIMENT140: designed and developed by [19]. Sentiment140 *allows us to locate the sentiment of a brand, product or topic on Twitter* and is highly recommended for use in case of brand management, polling, planning a purchase etc. Sentiment140 *uses machine learning algorithms for the sake of classification* unlike various other simple keyword-based approach, thus providing high precision and recall.

f. SENTIMENT PACKAGE in R: Ref. [20] made use of 2 modules namely **classify_emotion()** and **classify_polarity()** from SENTIMENT PACKAGE in R. Ref. [20] further defined them as follows:

classify_emotion(_): This function is responsible for categorizing or analysing textual data of any type, into different emotions: joy, surprise, fear, disgust, anger and sadness. This function performs classification with the help of Naïve Bayes algorithm which is trained on the dataset provided by Carlo Strapparava and Alessandro Valitutti's emotions lexicon.

classify_polarity(_): This function is used for extricating the polarity of any textual document. The most basic approach followed by this function is categorizing the piece of text on the basis of negative, positive and neutral polarity. Moreover, this function works on Naïve Bayes algorithm which is trained on the dataset provided by Janyce Wiebe's subjectivity lexicon.

g. EMOTION MINING TOOLKIT(EMTk): EMTk is designed and developed by [20]., under MIT open source license, and is written in JAVA, PYTHON, and R. EMTk is a collection of modules and datasets that offer comprehensive solution for text mining, emotion analysis and sentiment analysis. Broadly, EMTk is an integration 2 tools namely **EmoTxt** by [21] and **Senti4SD** by [22]. EmoTxt helps in recognizing 6 major emotions out of text namely, *joy, love, surprise, anger, sadness and fear* while Senti4SD is a polarity classifier trained using supervised learning methods, following lexicon approach.

h. NRC Word-Emotion Association Lexicon(EmoLex): EmoLex is designed and developed by [23]., as a list roughly containing 14000 English words. Each word in the list is associated with either of the 2 sentiments namely *positive and negative*, and 8 basic human emotions (*anger, fear, anticipation, trust, surprise, sadness, joy and disgust*).

D. Text Classification

Ref. [24] defines TEXT CLASSIFICATION as *a task of designating a set of pre-defined categories to free-text.* There are many approaches for automatic text classification but they can be broadly classified into 3 major systems described in Table I, Table II and Table III respectively.

TABLE I. RULE BASED SYSTEMS

RULE BASED SYSTEMS	According to [24]. Rule Based Systems *classify text into ordered groups by using a set of customized linguistic rules.* These systems make use of *Semantic* similarities among the relevant elements of the text to classify them into relevant classes. A rule however, is a combination of a pattern and a predicted class.
ADVANTAGES	The rules are easy to comprehend and understand
DISADVANTAGES	The design process for the rules is time consuming, challenging, and requires a lot of analysis and testing related to the domain. The task of updating the rule set is also complex reducing the scalability of the system.

TABLE II. MACHINE LEARNING BASED SYSTEMS

MACHINE LEARNING BASED SYSTEMS	According to [24]. Machine Learning Systems *makes classifications based upon past observations*. These systems make use of pre-classified texts as examples and associations among the various elements of text to tag new/un-seen data with various pre-defined classes.
ADVANTAGES	Machine Learning Based Systems are more accurate than Rule Based Systems and works well with complex classification tasks.
DISADVANTAGES	The accuracy of the model depends on many factors like the quality of training data, the train: test ratio, cross-validation process etc. Also, the training phase is often more time consuming than prediction phase.

TABLE III. HYBRID SYSTEMS

HYBRID SYSTEMS	According to [24]. Hybrid Systems combine a base classifier trained with machine learning algorithm and a rule-based system, which is used to further improve the results.
ADVANTAGES	Hybrid Systems are comparatively fine-tuned which overcome the limitations of both Rule-Based Systems and Machine Learning Based Systems while combining their individual advantages, thereby increasing the accuracy.
DISADVANTAGES	The design complexity of Hybrid Systems is high.

Text Classification Algorithms: Some of the most widely used text classification algorithms are described in the Table IV.

TABLE IV. HIGHLIGHTS, DESCRIPTION AND LIMITATIONS OF VARIOUS CLASSIFIERS

Name	Highlights and Description	Limitations
Support Vector Machine Classifier	The basic working principle of the Support Vector Classifier(SVM) is to create a line or hyperplane which separates the data into classes (positive, negative and neutral). Hence, the output of a SVM is a hyperplane that segregates the input data into pre-defined classes. Ref. [25] defined *a hyperplane in an n-dimensional Euclidean space as a flat, n-1 dimensional subgroup of that space that divides the space into two disengaged parts.*	• Choosing an appropriate Kernel function is a tricky and complex task • SVM models are often very difficult to interpret because of their long training time • Requirement of humongous amount of memory
Naïve Bayes Classifier	Probably, the most widely used classification model in the field of machine learning is the Naïve Bayes Classifier. The basis of this classifier is the probabilistic Bayes theorem. Given below is the mathematical equation for the Bayes theorem: $$P(A|B) * P(B) = P(B|A) * P(A)$$ **Where P(A|B) is the probability of occurrence of event A given that event B has already occurred, P(B|A) is the probability of occurrence of event B given that event A has already occurred, P(B) is the probability of occurrence of event B and P(A) is the probability of occurrence of event A.** The 2 main rules governing the Naïve Bayes Classifier are: ▪ All the features of the data that help in finding out its sentiment class are all equally significant. ▪ All the features are independent of each other.	• Assumption of predictor features being independent of one another is impractical • Values not observed in the training dataset are assigned zero probability and hence the classifier fails to make any predictions regarding them
Deep Learning Algorithms	Deep Learning is a set of algorithms, techniques and models that can match the level of computation of a human mind. Artificial Neural Network(ANN) form the basis of Deep Learning. With recent researches and development in the field of ANN and AI, Text Classification using Deep Learning has reached all new level of accuracy. Some of the most widely used Deep Learning architectures for text classification are *Multi-Layer Feed Forward Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks, Recursive Neural Networks* etc. The prime advantage of using this method for text classification is that these models possess no saturation point for their accuracy i.e. more the examples fed to these networks in order to train them, more their accuracy improves.	• Inability of act intelligently • Failures encountered during multiple domain integrations • Inability to handle data beyond the training set
Decision Tree Classifier	In Decision Tree Classifier, a new data is classified by passing it through a series of tests that identify its class label. The tests are organized in a tree-like hierarchical structure called as decision-tree. The formation of a decision tree follows a top-to-bottom approach and the tests which maximize the information gain related to classification are performed first.	• Highly sensitivity as a small change in the dataset can lead to completely different Decision Tree • Long duration of training period • Difficulty in handling continuous values
K-Means Clustering	Ref. [26] defined clustering *as a type of unsupervised machine learning which aims to find co-relative sub-groups such that objects in the same group(clusters) are more similar to each other than the others.* K-means Clustering divides the dataset into 'K' clusters. The value of 'K' however, can be dictated to the algorithm which can be equal to or more than the number of sentiment classes.	• Decision regarding the value of 'K' is in the hands of user • Suitable only for numerical data • Deals well only with spherical clusters that too having fairly equal number of observations
K-Nearest Neighbours(KNN)	Ref. [27] described KNN algorithm *as a simple algorithm which uses the complete dataset in its training phase. Whenever a prediction is required for unseen data specimen, it searches through the entire training dataset for K-most similar instances and the data with the most similar instance is finally returned as prediction.* Some major applications of KNN are Recommendation Systems, Pattern Recognition, Intrusion detection, spam detection, and in systems where highly accurate and frequent predictions are required to be made.	• Does not work well with large datasets having variable dimensions • Sensitive to noise and missing values • Need for feature scaling methods like normalization etc.
Random Forest Classifier	Ref. [28] defined Random Forest Classifier as *a supervised learning algorithm. The 'forest' it builds, is a group of decision trees, usually trained with the 'bagging' method. The general idea of the bagging method is that a blend of learning models increases the overall accuracy.* In other words, it combines the results of predictions and accuracy of various decision trees in order to merge them into a more stable model, called as a 'forest'. Random Forest Classifier can be used both for classification as well as regression.	• Long training periods • Highly complex in designing and understanding

E. Results Visualization

In the last phase of the sentiment analysis process, the results of sentiment analysis and text classification, performed previously, are visualized using appropriate tools, libraries, modules etc. Along with visualization, certain other performance parameters for the classification model designed are calculated in this phase like *accuracy, recall* etc.

a. Visualization Tools: Some of the most widely used libraries and modules for result visualization are described below:

a.1. Matplotlib: Matplotlib was designed and developed by [29]. as an extensive library for creating static, animated and interactive figures in Python with over 700000 lines of code. Some widely used plots provided by Matplotlib are: Stacked Bar Graph, Horizontal Bar Graph, Histograms, Pie charts, Scatter Plots, Contours, Geographic plot etc. It is difficult to recognize any one of these plots as the most 'important' and 'detailed' one, as the use of each one of them depends on its area of application and domain.

a.2. Word Cloud: A Word Cloud (wordle, tag cloud, word collage etc.) is a visual representation of words that give greater prominence, emphasis and highlight to words that appear frequently than others. Such visualizations help the presenters to quickly depict data to the audience such as the most frequently used words, most frequently used words to convey a certain emotion, words highlighting the topics covered by a piece of text, trending hashtags in social media etc. An example of Word Cloud is depicted in Figure 10.

a.3. Data Visualization Using R: Nowadays, the language R is gaining immense popularity in the field of machine learning and data visualizations. The visuals created by R are both appealing and simple to code. This appeal is because of the diverse package provided by R for data visualization and gaining insights about data through charts, graphs etc. Some of the widely used R Packages for Data Visualization are depicted in Fig. 11.

a.4. Rapidminer: With the release of the version 5.2, RapidMiner is now equipped with new data visualization module known as *RapidMiner's Advanced Charts*. This module by far, has been treated as an alternative for the previous data visualization module *Plot View* but is planning to replace it complete with future releases. This new module provides a variety of charts that can work well with individual, grouped and aggregated data directly for creation of charts, hiding the complex data transformations required to be performed for the same.

b. Performance Parameters: After a classification model makes predictions about the classes of un-labelled texts, the results are compared with human or manually assigned labels to the same texts. This comparison leads us to certain performance parameters the values of which helps to evaluate the performance and correctness of the classification model. These parameters are described in Table V.

TABLE V: VARIOUS PERFORMANCE PARAMETERS FOR CLASSIFICATION MODELS

Parameter	Description
True Positives	The manually annotated sentiment class is 'positive' and the classification model also labelled the data with sentiment class 'positive'
False Positives	The manually annotated sentiment class is 'negative' and the classification model labelled the data with sentiment class 'positive'
True Negatives	The manually annotated sentiment class is 'negative' and the classification model also labelled the data with sentiment class 'negative'
False Negatives	The manually annotated sentiment class is 'positive' and the classification model labelled the data with sentiment class 'negative'
Accuracy	The percentage of texts that were predicted and labelled with the correct sentiment class.
Precision	The percentage of texts the classifier labelled correctly out of the total number of texts that it predicted for the given label.
Recall	The percentage of texts the classifier predicted for a given class label out of the total number of texts it should have predicted for the given tag.
F-score	It is the harmonic mean of precision and recall.

III. RECOMMENDED AREAS OF APPLICATIONS

Text Classification has a wide range of applications from classification of short texts like tweets, news headlines, etc. to much larger texts like journals, articles, books etc. with sentiment analysis as the most common motive behind performing text classification. *Sentiment Analysis is the automated process of classifying a piece of text into 'positive', 'negative' and 'neutral' classes and sometimes into classes that depict emotions like 'joy', 'fear', 'surprise' etc.* Many companies nowadays, rely on Sentiment Analysis for *product analytics, brand monitoring, workforce analytics, marketing and advertising, customer support* etc.

Out of so many applications of Sentiment Analysis mentioned above, some are discussed below with specifications:

A. Ranking Systems: Ref. [30] used Sentiment Analysis process to rank fan pages in Facebook taking in account the polarity of comments and posts and not just number of likes, comments and user engagements by making use of the AFINN database.

B. Medical Sciences: Ref. [31] performed data and sentiment analysis on data collected from several diabetic and non-diabetic people about their awareness about diabetes, its symptoms, food items causing and controlling diabetes etc. using Hadoop Streaming Map-Reduce in Python and presented the results in the form of word clouds. Ref. [32] mined Facebook data collected from a number of pages which deal with people and their reactions and engagements on several rare diseases. The text mining is performed using R and the results were compared with the priorities of the Decalogue of the Spanish Federation of Rare Diseases(FEDER).

C. News Headlines and Public Information Systems: Ref. [33] performed opinion mining on News Headlines using SentiWordNet to classify news headlines as positive and negative classes based upon their impact.

D. Stocks, Shares and Finance: Ref. [34] proposed the use of data analytics to make predictions of future Stock Prices and also performed sentiment analysis of the stocks using recent tweets on Twitter. This helped the investors to invest their money on the right stock and also analysed their sentiments on various stocks through their words on Twitter using Python and R.

E. Government Policies, Reforms and Politics: Ref. [35] collected a huge amount of data from various social media sites like Facebook, Twitter etc. and analysed the opinions and insights about how people felt about this drastic reform made by the government as a step against corruption and black-money on 8th November 2016 by making use of Python and R. On the very similar note [36] analysed sentiments and opinions of the public on revocation of Article 370 which gave a special status to Jammu and Kashmir from the Indian Constitution. They analysed 2200 tweets concerning the same and tried to predict the impact of the reform on trade, terrorism etc.

F. Education: Ref. [37] investigated opinion mining by means of supervised learning algorithms to collect and evaluate students' feedback on various topics such as grades, conscription information, coaching, study patterns, examinations etc.

IV. Conclusions and Future Scope

In this paper, we had a glance of what sentiment analysis is, various techniques and methodologies to perform sentiment analysis and its areas of application. We also had an overview of various modules and libraries provided by different programming languages to perform sentiment analysis and also reviewed various text classification models to fit in and classify new un-seen data into various sentiment classes. It is an ambiguous task to choose the best model and algorithm to perform sentiment analysis and text classification as different models perform with different accuracies for different applications under different scenarios. Sentiment Analysis/Opinion Mining provides a wide area of research, job opportunities and is also the need of the hour with more people, spending more and more time online, generating abundance of data every minute which might have impact both on their individual lives as well as on the society.

In future, we would direct our research towards more accurate, precise and smart methodologies to perform sentiment analysis and explore more practical and regular applications of the same as the modes of expressing views keep on evolving from text to speech and from hashtags to emoticons, stickers and gifs.

Acknowledgements

For this work, we would like thank all the faculty members and staff members of Gyan Ganga Institute of Technology & Sciences, Jabalpur and Maulana Azad National Institute of Technology & Sciences (MANIT), Bhopal for their co-operation and knowledge providence.

References

[1] Chaffey, Dave. Global social media research summary July 2020. SMART INSIGHTS. [Online] 3 AUGUST 2020. [Cited: 17 AUGUST 2020.] https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/.

[2] Garbade, Dr. Michael J. A Simple Introduction to Natural Language Processing. [Online] 15 October 2018. [Cited: 17 August 2020.] https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32.

[3] Liu, Bing. Sentiment Analysis and Opinion Mining. https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf. [Online] 22 April 2012. [Cited: 17 August 2020.] https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf.

[4] SysNucleus. What is Web Scraping? [Online] 17 August 2020. [Cited: 18 August 2020.] www.webharvy.com.

[5] Richardson, Leonard. Beautiful Soup Documentation. www.crummy,com. [Online] 2020. [Cited: 18 August 2020.] www.crummy.com/software/BeautifulSoup/bs4/doc/.
[6] Garcia, Miguel. How to Make a Twitter Bot in Python With Tweepy. [Online] [Cited: 18 August 2020.] https://realpython.com/twitter-bot-python-tweepy/.
[7] Gupta, Mohit. NLP | Chunking and chinking with RegEx. [Online] 28 January 2019. [Cited: 19 August 2020.] https://www.geeksforgeeks.org/nlp-chunking-and-chinking-with-regex/.
[8] Worthy, Paige and Newberry, Christina. The Ultimate List of Social Media Definitions You Need to Know. [Online] 31 March 2020. [Cited: 20 August 2020.] https://blog.hootsuite.com/social-media-glossary-definitions/.
[9] Bagadiya, Jimit. 309 Social Media Statistics You Must Know In 2020. [Online] 2020. [Cited: 20 August 2020.] https://www.socialpilot.co/blog/social-media-statistics.
[10] Twitter Sentiment Analysis: The Good the Bad and the OMG! Kouloumis, Efthymios, Wilson, Theresa and Moore, Johanna. 2011. 5th International AAAI Conference on Weblogs and Social Media.
[11] Social Opinion Mining and Concise Rendition. Balaji, Harish, Govindasamy, V. and Akila, V. s.l. : IEEE, 2016. International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)
[12] Lessons Learned: A Case Study in Creating a Data. Tiezzi, Jason, Tyler, Rice and Sharma, Suchetha. s.l. : IEEE, 2020. Systems and Information Engineering Design Symposium (SIEDS).
[13] Perry., Patrick O. AFINN Sentiment Lexicon. [Online] [Cited: 21 August 2020.] http://corpustext.com/reference/sentiment_afinn.html.
[14] AFINN. Nielsen, Finn Årup. s.l. : Informatics and Mathematical Modelling, Technical University of Denmark, 2011.
[15] The Lexicon-based Sentiment Analysis for Fan Page Ranking in Facebook. Ngoc, Phan Trong and Yoo, Myungsik. s.l. : IEEE, 2014. The International Conference on Information Networking 2014 (ICOIN2014).
[16] SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Baccianella, Stefano, Esuli, Andrea and Sebastiani, Fabrizio.
[17] Loria, Steven. TextBlob. [Online] 2020. [Cited: 21 August 2020.] https://textblob.readthedocs.io/en/dev/.
[18] Thelwall, et al. SentiStrength. [Online] 2010. [Cited: 22 August 2020.] http://sentistrength.wlv.ac.uk/#About.
[19] Go, Alec, Bhayani, Richa and Huang, Lei. Sentiment140 - A Twitter Sentiment Analysis Tool. [Online] 2013. [Cited: 22 August 2020.] http://help.sentiment140.com/.
[20] Sentiment Analysis of Indian Currency Demonetization . Vaid, Ishan, et al. s.l. : IEEE, 2017. International Conference On Smart Technologies For Smart Nation (SmartTechCon).
[21] EMTk – The Emotion Mining Toolkit . Calefato, Fabio, et al. s.l. : IEEE, 2019. ACM 4th International Workshop on Emotion Awareness in Software Engineering (SEmotion).
[22] EmoTxt: A Toolkit for Emotion Recognition from Text. Calefato, Fabio, Lanubile, Filippo and Novielli, Nicole. 2017. 7th Affective Computing and Intelligent Interaction (ACII'17).
[23] Sentiment Polarity Detection for Software Development. Calefato, Fabio, Lanubile, Filippo and Novielli, Nicole. 2017.
[24] Mohammad, Saif and Turney, Peter. NRC Word-Emotion Association Lexicon (aka EmoLex). [Online] [Cited: 25 August 2020.] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.
[25] Text Classification. [Online] [Cited: 23 August 2020.] https://monkeylearn.com/text-classification/.
[26] Pupale, Rushikesh. Support Vector Machines(SVM) — An Overview. [Online] 16 June 2018. [Cited: 25 August 2020.] https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989.
[27] Khan, Mudassir. KMeans Clustering for Classification. [Online] 2 August 2017. [Cited: 26 August 2020.] https://towardsdatascience.com/kmeans-clustering-for-classification-74b992405d0a.
[28] Sarkar, Priyankur. What is K-Nearest Neighbor in Machine Learning: K-NN Algorithm. [Online] 19 September 2019. [Cited: 27 August 2020.] https://www.knowledgehut.com/blog/data-science/knn-for-machine-learning.
[29] Donges, Niklas. A COMPLETE GUIDE TO THE RANDOM FOREST ALGORITHM. [Online] 16 June 2019. [Cited: 27 August 2020.] https://builtin.com/data-science/random-forest-algorithm.
[30] Matplotlib: A 2D graphics environment. Hunter, John. s.l. : IEEE, 2007.
[31] The Lexicon-based Sentiment Analysis for Fan Page Ranking in Facebook. Ngoc, Phan Trong and Yoo, Myungsik. s.l. : IEEE, 2014. The International Conference on Information Networking 2014 (ICOIN2014).
[32] Data Analytic on Diabetic awareness with Hadoop Streaming using Map Reduce in Python. Ramsingh, J. and Bhuvaneswari, V. s.l. : IEEE, 2016. International Conference on Advances in Computer Applications (ICACA).
[33] Mining Facebook data of people with rare diseases. Reguera, Natalia, Subirats, Laia and Armayones, Manuel. s.l. : IEEE, 2017. 30th International Symposium on Computer-Based Medical Systems.
[34] Opinion Mining of News Headlines using SentiWordNet. Agarwal, Apoorv, et al. s.l. : IEEE, 2016. Symposium on Colossal Data Analysis and Networking (CDAN).
[35] Stock Price Prediction Using Data Analytics. Tiwari, Shashank, Bharadwaj, Akshay and Gupta, Dr. Sudha. s.l. : IEEE, 2017. International Conference on Advances in Computing, Communication and Control (ICAC3).
[36] Sentiment Analysis of Indian Currency Demonetization . Vaid, Ishan, et al. s.l. : IEEE, 2017. International Conference On Smart Technologies For Smart Nation (SmartTechCon).
[37] Twitter Data Visualization and Sentiment Analysis of Article 370 . Patil, Rupali, Gada, Nishant and Gala, Krisha. s.l. : IEEE, 2019. International Conference on Advances in Computing, Communication and Control (ICAC3).

[38] Performance Analysis on Student Feedback using Machine Learning Algorithms . Katragadda, Sharnitha, et al. s.l. : IEEE, 2020. 6th International Conference on Advanced Computing & Communication Systems (ICACCS)
[39] Hutto, C., Gilbert, & E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. 8th International AAAI Conference on Weblogs and Social Media.