A Tool for Analysing Social Media Feedack To Gain Profit For Business Organizations

Rushan Silva, Dimuthu Yasas, Hiruni Mirando, Kushali Withanage, Yashas Mallawarachchi
Department of Software Engineering
Sri Lanka Institute of Information Technology
Colombo, Sri Lanka
rushan.v.s@gmail.com, viduranga008@gmail.com, hiruninm95@gmail.com, kushaliwithanage@gmail.com, yashas.m@sliit.lk

***Abstract*—Contemplating the fact that business enterprises globally, benefit from social media to gain profits, this research paper discusses a tool incorporated with analyzing mechanisms of customer responses on social media to maximize the gained profit and take befitting future decisions. The tool analyses the feedback received on social media posts in the forms of reactions, comments and shares. Gathered data will be preprocessed and undergo several techniques to achieve normalization and duplication removal. Moreover, a spam detection technique based on feature vectorization and Bayes theorem is used to identify the relevancy of the comments. Cleansed data is passed through a trained vector space model which label the comments into seven emotion categories related to sentiment analysis. Simultaneously, a keyword extraction model implemented using Maui, which has the underlying technology of the machine learning toolkit Weka, utilize the data to extract suggestions conveyed through comments based on semantic analysis techniques. Subsequently, all the retrieved and processed data undergo a random forest model trained to predict the successfulness of the post using the accuracy score.**

Keywords—Social Media; Spam Detection; Semantic and Sentiment Analysis; Random Forest; Keyword Extraction; Pattern Matching

I. INTRODUCTION

With the evolution of technology, the World Wide Web is rapidly evolving day by day. This has paved the way for social media to become a key factor in everyone's life. Social media has a major impact on the thinking patterns of people and has delivered an efficient media to express their thoughts. Social media offer more than a considerable amount of user-generated content and unsurprisingly people rely heavily on them in decision making. Further, user-generated content can be used to analysis thinking patterns and offer more products and services adapted to customer needs. This has lead business enterprises to select social media as the exemplary platform for marketing which results in sharing their brand promotions around the world within seconds. Gathering of customer feedback through social media platforms assist to analyze the behavior patterns of the customers towards the marketing campaigns. These analytical data aid to inform and guide the organization's marketing strategy to decide the next best step to be taken in the future. Due to this, social media analysis has been identified as a useful research discipline.

Presently, with the high demand for social media, around 200 social media platforms can be found with over 3 billion active users as of April 2018 [1]. This research focuses on Facebook, which has emerged as the top social media platform having over 2 billion users and 80 million enterprise users as of July 2018 [2]. Further, more than one billion people visit Facebook business pages monthly [3].

Traversing through customer responses fail to provide a detailed and an in-depth analysis of the responses. The tool discussed in the paper is provided with the capabilities of placing the comments under major emotion labels, extracting suggestions from the comments and predicting the successfulness of the post.

In this research, the attention was given to the food domain, due to ease of finding training data. The data for the training process was obtained from the Graph API of Facebook in the JSON format. In order to perform modes of analyzations, the retrieved data foremostly undergo a cleansing process. This process includes few data mining techniques namely, sentence splitting, stemming and lemmatization, spell check, grammar checking, duplicate removal. Subsequently, a spam detection technique operates over cleaned data in order to identify data which precisely related to the post. ElasticSearch is used as the query engine to enable fast and effortless querying. Further, emotions behind comments are analyzed and placed them under seven main emotion categories namely joy, sadness, anger, shame, disgust, guilt and fear. Simultaneously, the suggestions expressed in the comments are extracted. This task is achieved with the help of Maui, a GPL licensed library written in Java for keyword extraction. The repetition amount of each suggestion is given as a percentage using StringSearch, a pattern matching algorithm. Finally, the tool predicts the successfulness of the post with the assist of the supervised learning technique random forest, proceeding with hyperparameter tuning and optimization, and performance metrics measurement for classification.

The structure of the rest of the paper is as follows: Section 2 indicates the prime illustration of the background and the discussion of the related work with literature survey. Section 3 presents the methodology of the mentioned features while Section 4 discusses the results. The research conclusion and future dimensions are mentioned in Section 5 respectively.

II. BACKGROUND

Social media has attracted the attention of many types of research in various fields, especially data mining, sentiment

analysis and semantic analysis. This remains nothing of a surprise as to how social media platform users have made it a media to express opinions about their day-to-day lives.

Many types of research can be found regarding different data preprocessing techniques that will be used to clean raw data and generate a structured data set. Most of the researches that have done using social media data use the same outdated techniques to preprocess data. Once investigated, it can be seen those researches haven't paid much attention to proper data preprocessing.

The primary focus of "A Natural Language Normalization Approach to Enhance Social Media Text Reasoning" research was to enhance text understanding quality by using more sophisticated data preprocessing techniques [4]. Main features including filtering, removal of special characters, removal of stopwords, lemmatization and stemming, spelling correction, Negative Contraction Transformation, Affirmative Subject transformation, affirmative sentence transformation, word normalization and typos correction were discussed. One main issue that could be identified in this approach is it being highly generalized which is prone to incorrect results. Another shortcoming that is noticeable is that there is no method implemented in removing non-related data of a specific post on social media.

Another work is of regarding feature selection and extraction in data mining which stated how particular features can be collected when there are many features available in the data pool by reducing both training time and over-fitting [5]. This has been targeted to achieve feature selection without losing the performance with the same outcome based on supervised learning.

Although the possibility of expressing themselves by reacting to a post is present, it can never be as expressive as a comment. Sentiment analysis has paved the way to analyze these comments and label them into several emotion categories for a better customer understanding. When labeling comments into emotion categories, we meet two significant emotion models, categorical and dimensional models. Dimensional model, which consists of two subparts, Valence and Arousal, shows an effect in a dimensional form. Valence measures positive or negative affectivity of emotion and arousal measures calmness or excitement of the information [6]. The categorical model assumes their discrete emotion categories as Ekman's six main emotion categories, namely, happy, sad, angry, disgust, fear and surprise [7].

A research has been conducted on supervised vector space model (VSM) for indexing the document, finding word count, information retrieval and information filtering [8]. The attention has dawned upon reducing VSM representation with techniques well known in information retrieval such as Latent Semantic Analysis, Probabilistic Latent Semantic Analysis (PLSA) and the Nonnegative Matrix Factorization representations.

Comments can be categorized in three, namely, Objective Comments, which are unbiased sentences or facts about entities or events, Opinions, which are people's sentiments and feelings towards entities or events, and Suggestions [9]. Suggestions alone can be interpreted into two categories, namely, Explicit and Implicit, where Explicit Suggestions are expressed as wishes or improvements and Implicit Suggestions must be drawn from a negative opinion.

Recently, opinion mining and sentiment analysis has sprang into interest in both academia and the industry, as a comment which expresses either an opinion or a suggestion, can directly impose on the business or organization. Due to this high interest, people globally tend to place their researches on these terms. A research was conducted to extract suggestions from students' comments which provided feedback on the course and the instructor at the end of a semester [9]. Existing text mining and data visualization techniques have been used to achieve the target of extracting and visualization of comments which consist of implicit suggestions. Use of four statistical classifiers, namely, Decision Trees C5.0, Generalized Linear Models (GLM), Support Vector Machine (SVM) and Conditional Interference Tree (CTREE) can be seen.

A research was conducted to detect suggestions and improvements comprehended in user comments [10]. Opinion mining and sentiment analysis concepts together with Natural Language Processing techniques have been used to achieve the task of extracting suggestions automatically. Further, feature-based sentiment mining is performed on the collected data. Another research based on extracting customer-to-customer suggestions from reviews have provided a three-fold contribution of problem definition, benchmark dataset and detect suggestion approach [11].

The success rate of a social media post depends on the various measurements called social key performance indicators (KPIs) where it generally states as likes, comments and shares. Statistical details would pave the road to discuss the relationships between metrics and producing the accuracy score and other related performance tests which can discuss the whole idea behind a post. A research based on analysis of Facebook reactions versus shares when a crisis related communication occurred in social media exhibits how each individual reacts to a Facebook post related to a crisis by using its new reaction mechanism which comprises of 'like', 'love', 'haha', 'wow', 'sad' and 'angry' [12]. Negative binomial regressions where calculations were carried out using the statistical software called package R and R package MASS was used to analyze the data. Results were displayed and generated according to negative binomial regression mentioned. Relationships between Facebook reactions were calculated by correlation coefficients for each reaction and similar correlation coefficients for reaction pairs such as likes and love or likes and sadness were identified separately.

III. METHODOLOGY

The conclusive output of the research is a social media data analyzing tool with capabilities of spam detection, emotion analysis, suggestion extraction and success prediction. In order to achieve this goal the research was conducted in four main parts: the first is collecting data from Facebook, preprocessing and constructing them into a usable

format, the second is labelling comment related data into seven emotion categories, the third is extraction of suggestions expressed in the comments and obtain the repetition amount for each suggestion, and lastly prediction of success rate for a post based on various feedback received. Figure 1 depicts the architecture used to make the social media analyzing tool a reality.

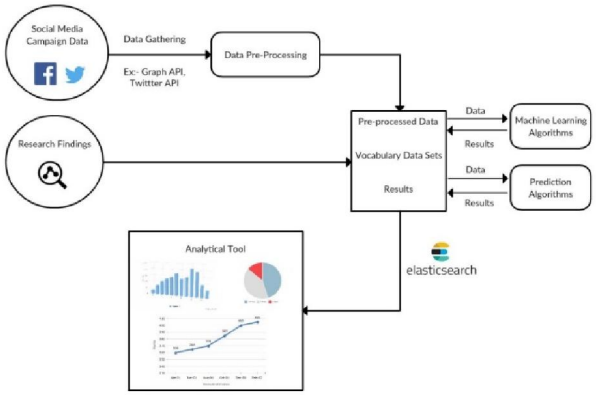


Fig. 1. Architecture of the social media analyzing tool

A. Data Preparation and preprocessing

The data is collected through the Graph API of Facebook. During the first phase of preprocessing the data undergo through sentence splitting, stemming and lemmatization, spell check, grammar checking and duplicate removal. Sentence splitting is the method of identifying where sentences begin and end. Stemming is the process of reducing inflected words to their word root form whereas lemmatization is the process of grouping together the inflected forms of a word, so they can be analyzed as a single item. The second phase of preprocessing, the data is checked for spam detection.

- *Spam Detection Technique*

First and foremost, a dictionary of most commonly used words in comments is prepared. Comments are considered as a bag of words in preparing the dictionary. For the training purpose of the model, it is required to prepare a dataset which is understandable to the computer. A technique known as feature vectorization is used to achieve this.

The method of feature vectorization is similar to the bag of words model. In machine learning, feature vectors are used to represent numeric or symbolic characteristics, called features, of an object in a mathematical, easily analyzable way [13]. A familiar example for feature vectorization is the RGB color description. A color is described according to the amount of red, green and blue present in it. In related to research, this model is used to represent the comments in a way our machine learning model would understand it.

Hi	Sam	How	Was	Your	Day	Account	Has	Been	Selected	For	A	Prize	Money
1	1	1	1	2	1	1	1	1	1	1	1	1	1

Fig. 2. Dictionary

Hi	Sam	How	Was	Your	Day	Account	Has	Been	Selected	For	A	Prize	Money
0	0	0	0	0	0	0	0	0	0	1	0	0	1

Fig. 3. Feature Vector for “Send me some money for groceries”

In the above representation, the sentence “Send me some money for groceries” contains “For” and “Money” which is in the dictionary. Hence, those words are given a value of 1. Other words are not in the dictionary. Therefore, they are neglected in this case. But usually most of the words are covered with the growth of the dictionary.

Naïve Bayes classifier is used to train the model. Bayes theorem is the inner computational part used in the Naïve Bayes classifier. The Bayes theorem is given by:

$$P(A|B) = P(B|A).\frac{P(A)}{P(B)}$$

where, A and B are events and P(B) is not equal to 0, P(A) and P(B) are the probabilities of observing A and B without regard to each event.

P(A | B), a conditional probability, is the probability of observing event A given that B is true. P(B | A) is the probability of observing event B given the A is true.

In associated to research, A is considered as the probability that the comment is spam and B as the contents of the comment. If P(A | B) > P(¬A | B), then comment is classified as spam [14].

B. Emotion Analysis

After passing the preprocessing phase, cleaned data is utilized to analyze comments in order to label them into seven emotion categories, namely, joy, sadness, guilt, anger, fear, disgust and shame, with the help of vector space model. Vector space model or term vector model is an algebraic model for representing text documents as vectors of identifiers. Vector space model can be fully understood by following:

- Term-Frequency

Term Frequency is used to represent each term in vector space. It is a measurement of number of times a word present in the vocabulary. It can be denoted as below [15]:

$$Tf(t,d) = \sum_{x \in d} fr(x,t)$$

where, $fr(x,t)$ is defined as:

$$fr(x,t) = \begin{cases} 1, & x = t \\ 0, & otherwise \end{cases}$$

This equation returns the number of times that term t present in document d.

Assume there are 10 documents and n times of terms in the vocabulary, below equation display the representation of documents and terms as vectors.

$$V\ (d3)\ =\ (tf(t1,d3), tf(t2,d3), \ldots., tf(tn,d3))$$

$$V\ (d3)\ =\ (8, 12, 6, 23)$$

As mentioned, the above example carries the term t1 which appears 8 times in document 3 and term t2 appears 12 times in document 3. For further calculation the final result can be turned into a matrix format.

- Inverse Document Frequency

Inverse Document Frequency (Idf) value can be calculated using the following equation:

$$idf(t) = \log \frac{|\ D\ |}{1 + |\ \{d: t \in d\}\ |}$$

D denotes the number of documents in the corpus and d: t denotes the number of documents where the term t appears. Adding 1 to the denominator means to avoid it becoming zero. Moreover, turning this output to matrix format would proceed the calculation of tf-idf value [8].

- TF-IDF

Applying the above equations tf and idf values can be generated. Both of these values are in matrix format. Therefore, the result can be obtained by multiplying each value.

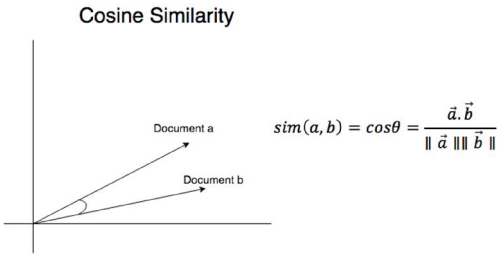
$$M\ (tf - idf)\ =\ M\ (tf)\ *\ M\ (idf)$$

Tf-idf value helps to map documents in vector space.

- Cosine Similarity

Cosine similarity is a measure that calculates the cosine of the angle between two vectors. Below equation demonstrate the cosine value [16].

$$\cos\Theta = ((a^{\rightarrow}.b^{\rightarrow})) / ((a^{\rightarrow}.b^{\rightarrow}))$$



Above cosine similarity formula describes the similarity between two vectors using angle between them. If cosine of the angle is near to 1, two vectors are in the same direction, and if angle between them is near to 0, that indicates two documents are similar to each other and if angle is near to -1, two vectors are in different directions which means that they are not similar. Mapping the emotion training and testing datasets in vector space proceeds the calculation of cosine similarity between them. Using cosine value, comments can be classified into emotion categories.

C. Suggestion Extraction

The suggestions or the ideas of improvement mentioned in the comments are extracted or separated and prioritized based on the repetition amount of each suggestion to make the task of business personnel easy and less tiring. To achieve this, keyword extraction along with pattern matching is used.

- *Keyword extraction*

Keyword extraction is the process of automatic identification of terms that best describe the subject of a document. But in this scenario, keyword extraction is used in reference to a controlled vocabulary to get the simplest form of the expressed suggestion. Maui, which stands for multipurpose automatic topic indexing, is a GPL licensed library written in Java [17]. The core of it is the machine learning toolkit Weka. It also can be referred to as a reincarnation of the keyword extraction algorithm KEA.

A vocabulary which contains suggestion related words for the food domain is constructed as the primary step. It can be either a terminology list, a thesaurus, a taxonomy or a lexicon. The purpose of this is to have consistency throughout the comments and to speed up the processing. Thesaurus and Wordnet, lexical databases for English, are used in the preparation of the vocabulary. The vocabulary should be in the Simple Knowledge Organization System (SKOS), a W3C recommendation designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary, in order to be compatible with Maui. Maui would tie the words in the vocabulary to unique IDs with the actual meaning of these phrases during the extraction process.

During the training process, to achieve better performance from the Maui, manually annotated comments are used. Once the probability threshold is increased, can get more important

words in the collection of comments which are not present in the formed vocabulary. If a comment is turned up without any of the words from vocabulary, it is categorized as a comment which doesn't contain a suggestion.

After a suggestion is extracted, it is compared with a list of already extracted suggestions to avoid the repetition of each suggestion. Instead of repeating each suggestion over and over again, a count of the repetition amount, i.e. the number of times each suggestion is repeated throughout the comments for a single post, is given. To accomplish this, pattern matching is used.

- *Pattern Matching*

Pattern matching is the act of checking a given sequence of tokens for the presence of the constituents of some pattern. In other words, matching would result in recognizing whether it is a perfectly a match or not. For this purpose, StringSearch, a high-performance pattern matching algorithm implemented in Java is used [18]. It provides the implementations of Boyer-Moore, a string-searching algorithm and Shift-Or (bit-parallel), a string-matching algorithm. These algorithms are five to ten times faster than implementing pattern matching through *java.lang.String*.

When the new suggestion comes, if it already exists in the list of suggestions, it would be a positive and if not, it would be negative, so it can be added to the list.

D. Success Prediction

The machine learning algorithm which will be discussed to perform the prediction is the random forest. The random forest (RF) methodology is known to be one of the most efficient classification methods widely used under supervised learning technique for prediction complications, where it addresses bootstrap aggregations commonly known as bagging and feature selection where it builds individual classification or regression trees for prediction [19][20][21].

However, according to the high dimensional feature space and data distribution, it can produce incorrect classification results and may contain bad trees classifiers. Moreover, a large proportion of bad trees in the random forest make the wrong decision since it gets the vote of all the trees to make an ensemble classification decision [22]. In this research, an ensemble algorithm called bagging is used for the decision trees which helps to avoid overfitting so that it improves the regression and classification prediction [23].

- *Measuring Performance Metrics for the Classification*

When considering the performance metrics, accuracy and f score metrics are mainly produced to predict the model. Since the accuracy is calculated using correct predictions by the total data points, it is known as the simplest and most commonly used performance metric. However, other than accuracy, recall, precision and f_1-score contribute to evaluate the better model [24].

Let TP be True Positive and FN be False Negative instances where it modeled as faulty modules where FT be False Positive and TN be True Negative instances under Non-Faulty modules. These instances are known as Confusion Matrix where it applies the measure of the performance for two class problem in the given dataset. Confusion matrix can be defined as follows using the aforementioned instances [24].

Correctly classified instances = True Positive (TP) + True Negative (TN)

Incorrectly classified instances = False Positive (FP) + False Negative (FN)

Total number of instances = Correctly classified instances + Incorrectly classified instances

- Recall

The recall is the ratio of TP to (TP + FN), which is the number of entire faulty modules.

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}$$

- Precision

Precision is the ratio of TP to (TP + FP) as known as entire module of fault-prone.

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}$$

- F_1 Score

The F_1 score is the combination of both recall and precision.

$$F_1 = \frac{2}{\frac{1}{recall}+\frac{1}{precision}} = 2.\frac{precision.recall}{precesion+recall}$$

For the positive real β, the formula is as follows,

$$F_\beta = (1+\beta^2).\frac{precision.recall}{(\beta^2.precision)+recall}$$

- Accuracy

Accuracy can be defined as the ratio of correctly classified instances to a total number of instances.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

- *Hyperparameter Tuning and Optimization*

The machine learning begins with some initial observation related to other science fields. In the scientific

method, finding a hypothesis which is commonly known as assumption can lead to execute the part of the experiments and finally analyze how approximately the earlier observations are reproduced.

While building the classifier, the hypothesis may produce incorrect assumptions. Therefore, tuning of the hypothesis by using hyperparameter which is a special type of configuration variable tend to produce the better hypothesis and figuring out the optimal values of these hyperparameters are known as hyperparameter tuning. In the process of evaluating the optimal values from the hyperparameters, the following parameters are used for the random forest.

- Number of decision trees in the forest
- Number of features to be considered when best split in a tree.
- The maximum depth of the tree

IV. Results And Discussions

A. Spam detection

Two sample comments were used to test the spam detection module where one comment is spam and the other one is not spam. "Wow, I love that place" and "I just got myself a free $500 gift card to Macy's, check out the URL below! http: //2url.org/?67562" are the spam and non-spam comments respectively. After testing these comments below results were obtained.

- Wow, I love that place – Not Spam
- I just got myself a free $500 gift card to Macy's, check out the URL below! http: //2url.org/?67562 – Spam

Above outputs were obtained with the use of Naïve Bayes Classifier, which was trained with sample comments. For every word in the dictionary, the input string is checked to see number of times a specific word is written. A features list is created with those counts and is passed to the trained model to classify the input string as Spam or Not Spam.

B. Emotion Analysis

While performing the emotion analysis of comments, two comments "When I won a trip to Greece in a competition" and "I felt disgusted when I realized that I had eaten something fatty", tested through the process. The following results were obtained for the mentioned comments. The first and second comments contained the emotion categories of 'Joy' and 'Disgust' respectively.

TABLE I. Output of the Emotion Analysis

Anger	Disgust	Fear	Guilt	Joy	Sadness	Shame
0.01	0.0	0.0	$8.38E^{-4}$	0.05	0.01	0.0
$5.77E^{-4}$	0.01	$2.13E^{-4}$	0.0	0.0	$5.05E^{-4}$	$2.01E^{-4}$

The output is based on the cosine similarity where the values near to -1 and 0 denote no effect and where the values near to 1 will describe which category that the comment will belong to. In the table, in the results of the first row, it can be seen that the highest value obtained is 0.05 which is for 'Joy', thus categorizing it as a joyful comment. In the second line of the table, it can be seen that the highest value is obtained for 'Disgust' making it a disgust comment.

C. Suggestion Extraction

After the training and validation processes, comments are passed through the model with the purpose of testing it. Two of the passed comments are 'Friendly staff. AC chambers. Foods are not delicious.', and 'Bad customer services. They don't provide chilli paste. They sell chilli paste in containers, kinda like their merchandise.'. The results obtained are as follows,

- Friendly staff. AC chambers. Foods are not delicious. – not delicious
- Bad customer services. They don't provide chilli paste. They sell chilli paste in containers, kinda like their merchandise. – bad customer service

When a post with 535 comments was tested through the suggestion extraction and prioritization method, the following result was obtained.

TABLE II. Output of the Repetition Amount

Comment	Repetition Amount
Not worth it	167
Not delicious	103
Bad Customer Service	77
Not clean	71
High service charge	19

98 comments were categorized as not containing any suggestion.

D. Success Prediction

As for the hyperparameters, random forest classifier worked with default values in the scikit-learn APIs. The successfulness of the post measured by performing accuracy rate and f_1-score for the unoptimized model and optimized model. Following figure display the outcome of the scores for unoptimized and optimized model respectively.

TABLE III. Output of the Random Forest

	Accuracy Score	F1 - Score
Unoptimized Model	0.9803	0.9803
Optimized Model	0.98333333	0.98333333

As seen above, the performance of the accuracy score is increased for the optimized model. The optimized model has 98% accuracy rate with the help of precision and recall. This optimized model has been used to predict the success rate of the post by providing the required values for the features in the random forest algorithm.

V. CONCLUSION

Presently, social media users grow at a rapid rate and would continue to do so in the future. Due to this growth, more and more businesses would shift their marketing to social media platforms. The tool discussed in this paper can be used to analyze social media feedback gained by an enterprise which can assist them in taking befitting future decisions towards success. The tool is implemented to perform spam detection, emotion analysis, suggestion extraction and success prediction on a Facebook post. First, all the unrelated comments are removed. Next, comments are labeled into seven emotion categories. Then, suggestions expressed in the comments are extracted with repetition amount for each suggestion. Lastly, the success rate of the post is predicted.

In the future, the plan is to refine and further improve the techniques in order to achieve an enhanced accuracy. Further, the domain is to be expanded and go beyond the current domain of food and to integrate more social media platforms to the tool.

REFERENCES

[1] "116 Amazing Social Media Statistics and Facts", Brandwatch, 2018. [Online]. Available: https://www.brandwatch.com/blog/96-amazing-social-media-statistics-and-facts/. [Accessed: 08- Aug- 2018].

[2] S. Reports, "By the Numbers: 800+ Amazing Facebook Statistics", DMR, 2018. [Online]. Available: https://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/. [Accessed: 08- Aug- 2018].

[3] [3]"Statistics of the top Facebook pages", Socialbakers.com, 2018. [Online].Available: https://www.socialbakers.com/statistics/facebook/pages/total/page-1-2/. [Accessed: 08- Aug- 2018].

[4] "A natural language normalization approach to enhance social media text reasoning – IEEE Conference Publication", Ieeexplore.ieee.org, 2018. [Online]. Available: http://ieeexplore.ieee.org/documents/8258148/. [Accessed: 01-Apr-2018].

[5] "Feature selection and extraction in data mining - IEEE Conference Publication", Ieeexplore.ieee.org, 2018. [Online]. Available: http://ieeexplore.ieee.org/document/7916845/. [Accessed: 02- Apr- 2018].

[6] Www2.bc.edu, 2018. [Online]. Available: https://www2.bc.edu/elizabeth-kensinger/Kensinger_RevNeurosci04.pdf. [Accessed: 02- Apr- 2018].

[7] Paulekman.com, 2018. [Online]. Available: https://www.paulekman.com/wp-content/uploads/2013/07/An-Argument-For-Basic-Emotions.pdf. [Accessed: 02- Apr- 2018].

[8] Minerazzi.com, 2018. [Online]. Available: http://www.minerazzi.com/tutorials/term-vector-3.pdf. [Accessed: 04-Apr- 2018].

[9] Ink.library.smu.edu.sg, 2018. [Online]. Available: http://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=4835&context=sis_research. [Accessed: 04- Apr- 2018].

[10] 2018. [Online]. Available: https://www.cicling.org/2013/rcs/Suggestion%20Mining_%20Detecting%20Suggestions%20for%20Improvement%20in%20Users_%20Comments.pdf. [Accessed: 04- Apr- 2018].

[11] Aclweb.org, 2018. [Online]. Available: https://aclweb.org/anthology/D/D15/D15-1258.pdf. [Accessed: 04-Apr- 2018].

[12] Scholarspace.manoa.hawaii.edu, 2018. [Online]. Available: https://scholarspace.manoa.hawaii.edu/bitstream/10125/50207/1/paper0320.pdf. [Accessed: 04- Apr- 2018].

[13] "Feature Vector | Brilliant Math & Science Wiki", Brilliant.org, 2018. [Online]. Available: https://brilliant.org/wiki/feature-vector/. [Accessed: 08- Aug- 2018].

[14] "How To Build a Simple Spam-Detecting Machine Learning Classifier", Hacker Noon, 2018. [Online]. Available: https://hackernoon.com/how-to-build-a-simple-spam-detecting-machine-learning-classifier-4471fe6b816e. [Accessed: 08- Aug- 2018].

[15] "Vector Space Model Pdf", Booktele.com, 2018. [Online]. Available: http://booktele.com/file/vector-space-model-pdf. [Accessed: 08- Aug-2018].

[16]"Implementing and Understanding Cosine Similarity", Masongallo.github.io, 2018. [Online]. Available: https://masongallo.github.io/machine/learning,/python/2016/07/29/cosine-similarity.html. [Accessed: 08- Aug- 2018].

[17] 2018. [Online]. Available: https://www.airpair.com/nlp/keyword-extraction-tutorial. [Accessed: 08- Aug- 2018].

[18] "StringSearch – high-performance pattern matching algorithms in Java", Johannburkard.de, 2018. [Online]. Available: https://johannburkard.de/software/stringsearch/. [Accessed: 08- Aug-2018].

[19] Breiman, L. 1996. Heuristics of instability and stabilization in model selection. The Annals of Statistics, Vol.24 Issue 6, pp. 2350–2383.

[20] Ho, T. 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.20 Issue 8, pp. 832–844.

[21] Bernard, S., Heutte, L., Adam, S. 2009. On the selection of decision trees in random forests. International Joint Conference on Neural Network , pp. 302–307.

[22] S. Bharathidason and C. Jothi Venkataeswaran, Ph.D, "Improving Classification Accuracy based on Random Forest Model with Uncorrelated High Performing Trees", Pdfs.semanticscholar.org, 2014. [Online].Available:https://pdfs.semanticscholar.org/7398/cb064c03cd6ca6bade115759ae5b0026f3cb.pdf. [Accessed: 06- Aug- 2018]

[23] "Ensemble methods: bagging and random forests", Nature.com, 2017. [Online].Available:https://www.nature.com/articles/nmeth.4438.pdf?origin=ppub. [Accessed: 06- Aug- 2018].

[24] D. Gupta, A. Malviya and S. Singh, "Performance Analysis of Classification Tree Learning Algorithms", Pdfs.semanticscholar.org, 2012.[Online].Available:https://pdfs.semanticscholar.org/c647/c68cf6ea691f80b03fe7dfd7cdb4fb3e44a4.pdf. [Accessed: 06- Aug- 2018].