# ANALYSYS OF SOCIAL MEDIA FEEDBACK TO GAIN PROFIT FOR BUSINESS ORGANIZATIONS

Project ID: 18-025

Project Proposal Report

S. A. R. V. Silva
K. D. Y. Viduranga
H. N. Mirando
K. V. N. Withanage

Bachelor of Science (Special Honors) in Information Technology
specializing in Software Engineering

Department of Software Engineering

Sri Lanka Institute of Information Technology
Sri Lanka

April 2018

# ANALYSIS OF SOCIAL MEDIA FEEDBACK TO GAIN PROFIT FOR BUSINESS ORGANIZATIONS

Project ID: 18-025

Project Proposal Report

(Proposal documentation submitted in partial fulfilment of the requirement for the Degree of Bachelor of Science Special (honors) In Information Technology)

Bachelor of Science (Special Honors) in Information Technology specializing in Software Engineering

Department of Information Technology

Sri Lanka Institute of Information Technology
Sri Lanka

April 2018

# DECLARATION

We declare that this is our own work and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

| Name | Student ID | Signature |
|---|---|---|
| S. A. R. V. Silva | IT15087584 | |
| K. D. Y. Viduranga | IT15035004 | |
| H. N. Mirando | IT15000354 | |
| K. V. N. Withanage | IT15003942 | |

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

……... ……

Signature of the supervisor Date
(Yashas Mallawarachchi)

# ABSTRACT

Contemplating the fact that business enterprises globally benefit from social media to gain profits, this proposal discusses the importance and impact of analyzing customer responses on social media to maximize the gained profit and take befitting future decisions. Main intention is to deliver a web application to cater business needs and goals by analyzing customer responses real-time to individual posts published on official social media accounts. Intended inputs for the machine learning algorithms would be the amount and the content of responses, where responses will be in the form of likes, comments, shares etc. By analyzing the mentioned inputs, the system would predict the successfulness of the post to assist the enterprise in their decision making. Further emotions behind comments would be analyzed and placed them under six main emotion categories which are happiness, sadness, anger, surprise, disgust, fear. Moreover, the mentioned suggestions in the comments would be extracted and the repeated amount of a suggestion will be shown as a percentage, aiding the enterprise to perceive customers timeless and effortlessly. In order to perform all the mentioned modes of analyzations, the data would be pre-processed, checked for proper spelling and grammar, and non-related data would be removed. Further, sentence splitting, normalization and duplication removal would be performed on the data before storage in database in a format which would enable fast and effortless querying. The proposed solution would display all the analyzed data in a user-friendly understandable manner and would provide the ability to compare current social media situation of the enterprise with their history.

# TABLE OF CONTENTS

## LIST OF FIGURES



## LIST OF TABLES

# 1. INTRODUCTION

## 1.1. Background & Literature Survey

Social media impact has turned out to be humans both personal and professional engagements with several sectors where entire society has dawn to capture the advantages through each part of it. In 21st century, social media phenomenon within the communication has widely covered the difficulties of gathering the information by assembling people together from making them online for a platform which provides every single detail around them. This is where the business enterprises perform exemplary behavior on marketing which leads to share their advertisements around the world within seconds. Therefore, customers can perform likes, dislikes, true human emotions intermittently, share the product details with one's friend list and comment the suggestions towards the products. Gathering of customer feedbacks through social media platforms assist to analyze the behavior of the consumers towards the advertisements or 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. Question here is how an organization would collect these customer feedbacks at no cost without thinking twice and gain the business decision up to next level. To achieve this, a platform that analyze the social media data and customer feedback related to published posts is in necessity.

The proposed solution is planned to launch under the name of 'Veracity' commercially making it available for all the enterprises which benefit from social media advertising. The product can be made available as several packages where each package would have a different set of features. Initially all these features can be given in a 30-day trial and afterwards the enterprise can select the necessary features and choose the required package to suit their requirements. These packages would have different prices ranging according to the amount and importance of the features they contain. If the enterprise has been a long-time client of Insight, some latest adding features can be delivered free of charge.

### 1.1.1. Data Preprocessing & non-related Review Removal Mechanism

Many researches 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 are done using social media data uses the same old school techniques to preprocess data. Many of those research papers don't even contain information regarding data preprocessing, which that those researchers haven't paid much attention on proper data preprocessing.

One research that could be found, which is about data preprocessing is "A Natural Language Normalization Approach to Enhance Social Media Text Reasoning" [1]. Primary focus of this research was to enhance text understanding quality by using more sophisticated data preprocessing techniques. 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 are the main features that are included in this research. One main issue that could be identified in this approach is that this is too generalized, which means that they haven't categorized their data when using these techniques. Therefore, when it comes to techniques such as Lemmatization and Stemming, Spelling Correction, Typos correction wrong decisions can be taken as this system is too generalized. Another shortcoming that is noticeable is that there is no method implemented in removing non-related data of a specific post on social media. This issue is also related with the fact that this system too generalized.

Another research regarding feature selection and extraction in data mining which stated how particular features can be collected when there is a large amount of features available in the data pool by reducing both reduction in training time and over-fitting. [2] Mainly they have mentioned how to do the feature selection as well as the feature extraction separately. This has been targeted to achieve feature selection without losing the performance with the same outcome based on supervised learning. According to the research, they have used feature extraction in two ways namely feature generation

and feature evaluation. Moreover, algorithms which are based on feature selection can be classified in to three categories called filters, wrappers and embedded techniques as well as feature extractions algorithms can be classified in to two main categories called linear and non-linear. Selected features will be truncated to nearest possible value. This indicates all the remaining features that stay closely get a general truncated value. Following truncation, evaluation of distance measure and logistic regression had been performed.

### 1.1.2. Text Analysis and Emotion Categorization

Although there is the possibility for people to express their emotions by reacting to a post, it can never be as expressive as a comment can be when expressing emotions. Through a comment, the ability to express what they truly feel is provided, which can't be achieved by the reacting to a post as it is limited to some emotions. Analyzing what has been commented and categorizing them into different emotions would paved the way for a better understanding of the customers.

When labeling comments into emotion categories, we meet two significant emotion models. One is categorical model and other is dimensional model. According to one research dimensional model shows affect in a dimensional form. [3] Further we can divide this model into two sub parts which are, valence or arousal. Valence measures positive or negative affectivity of emotion and arousal measures calmness or excitement of the information. In categorical model, it assumes their discrete emotion categories such as Ekman's six main emotion categories namely happy, sad, angry, disgust, fear, surprise. [4]

One research conducted is sentiment analysis on Twitter social media platform to label tweets into positive, negative and neutral categories. [5] They provide a supervised framework for expanding an opinion lexicon for tweet. To achieve this task, they use machine learning algorithms with word-level attributes based on POS(part-of-speech) tags and information calculated from streams of emoticon annotated tweets.

Another research is conducted to propose and investigate a paradigm to mine the sentiment from a popular real-time microblogging service and Twitter, where users

post real time reactions. [6] They use corpus based and dictionary-based methods to determine the semantic orientation of the opinion words in tweets. Further they use three main modules for their architecture namely Retrieval module, Preprocessing module and Scoring module. Retrieval module consists of data APIs. In their case they have used Twitter API. Data preprocessing module mainly concern techniques like removal of URLS, spell corrections, Emotion tagger, POS tagger operations. Lastly words are divided into verbs, adverbs and adjectives and a score was given to each word. Based on these scores they have separated into positive, negative, neutral categories.

A research has been conducted on supervised vector space model(VSM) for indexing the document, finding word count, information retrieval, information filtering. [7] They have also focused their attention on reduce 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.

### 1.1.3. Suggestion Extraction and Prioritization

From a mere like or a share, customers does not receive the ability to express themselves. To overcome this situation, commenting system was introduced, which soon became one of the most important aspects of understanding customers and gaining profit for an enterprise. A comment is a very powerful way of expression and a suggestion gives the degree of customer satisfaction which is a key pillar in business success. A customer can compose a comment to express their opinion about the organization and their related products or to give advices or warnings about the organization and their related products to other customers.

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. [8] Suggestions alone can be interpreted into two categories, namely, Explicit and Implicit, where Explicit Suggestions are expressed as wishes or improvements and Implicit Suggestions are similar to negative opinions and the suggestion must be drawn from

an opinion. In other words, a suggestion can be expressed either as a wishing the presence of a missing feature or regretting the absence of the same feature. [9]

Recently, opinion mining and sentiment analysis has sprang into interest 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. While on the hunt to knowledge about these trending research areas, below mentioned research works managed to capture the attention.

First and foremost, there was a research conducted to extract suggestions from students' comments which provided feedback on course and the instructor at the end of a semester. [8] This feedback was collected as quantitative rating using Likert scale and qualitative feedback as comments. They have used existing text mining and data visualization techniques to achieve their target of extracting and visualization of comments which consist of implicit suggestions. They have gone through the stages of pre-processing, implicit suggestion extraction and visualization to produce their output. In order to identify a comment as an implicit suggestion, four statistical classifiers, namely, Decision Trees C5.0, Generalized Linear Modals(GLM), Support Vector Machine(SVM) and Conditional Interference Tree(CTREE) have been used.

Secondly, a research was conducted with the intention of detecting suggestions and improvements comprehended in user comments. [9] They have applied opinion mining and sentiment analysis concepts together with Natural Language Processing techniques to address this new task of extracting suggestions automatically. They have defined a suggestion as either wishing the presence or regretting the absence of a missing feature. They have first identified whether the comment is positive, negative or neutral. Then feature based sentiment mining method is performed on the collected data. The conducted research contains four main components, namely, structured terminology of the topic, thesaurus of vocabulary, fine grained linguistic parser and extractor for suggestions.

Another research which has been conducted was to extract customer-to-customer suggestions from reviews. [10] This indicates that they have only identified the

suggestions which are given by customers to other customers relating to the particular organization and their products. Their desired final output was automatic detection of customer-to-customer suggestion expressing sentences in customer reviews. They have provided a three-fold contribution of problem definition, benchmark dataset and detect suggestion approach.

Lastly, a research has been conducted to extract and normalize entity-actions from user comments. [11] They have identified that a person has two intentions of writing a product review, which is to talk about the quality of the product and to help others in their decision making on product buying. They have also mentioned that an actionable comment as an expression with an entity such as a person or an organization and a suggestion that can be acted upon. They have separated the keywords in user comments and observed their frequency in order to identify an actionable comment.

### 1.1.4. Rating Measurement and Success Prediction

Most digital marketers or business enterprises willing to obtain an idea about the success rate of their posts after posted. This success rate depends on various measurements to give an instance if we consider a post on Facebook, we can gather count of likes, shares, comments as metrics measurements. Moreover retweets, likes and shares can be taken by analyzing Twitter. Human or post marketer can predict above analysis up to a common level yet increasing the number of measurements, finding relationships between attributes and calculating probability of measurements will guide this prediction goal beyond the human level.

When considering the literature survey for mentioned topic, we were unable to find comparatively numerous researches since this feature compromise of new methodologies and techniques.

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 crisis by using its new reaction mechanism which designate ‘like’, ‘love’, ‘haha’, ‘wow’, ‘sad’ and ‘angry’ and get the success of the post by analyzing the user behaviors. [12] The system analysis the data using negative

binomial regressions where calculations were carried out using the statistical software called package R and R package MASS. Results were displayed and generated according to negative binomial regression mentioned above. 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.

### 1. 2. Research Gap

During the literature review and information hunting, a range of researches and tools developed to achieve the task of the conducting research could be found. Although they may seem exactly the same at first glance, what the proposed system can achieve goes beyond what has already been achieved by the prevailing conducted researches and developed tools.

One feature which was common to all the researches but not to all the tools is the usage of single data source. When accessing the data from social media, most of the researches have turned their attention towards Twitter, whereas analytic tools provide the opportunity to use multiple social media. The proposed solution is based on multiple social media data sources, but mainly the attention would be focused on Twitter and Facebook as they are placed in top most five results of the most popular social networks. [13]

Another feature that came to attention was preprocessing of data. All the solutions proposed in conducted researches and tools, follow a series of preprocessing techniques before data storage. Although they preprocess direct data from APIs, none of them use feature extraction and removal of non-related data. In the proposed solution, the data would undergo a chain of preprocessing techniques and feature extraction along with a mechanism to remove non-related reviews. In feature extraction, all the related features which are necessary in analyzation will be extracted so that storage space won’t be wasted unnecessarily and to obtain higher processing speed. Further, with the intention of fast and time saving processing, non-related reviews to a particular post would be removed before storage.

In concern of semantic and sentiment analysis of comments, what can be seen is that, although there are many researches conducted based on them in related to comments published on social media, none of them is performed to enterprise domain to which the proposed solution would apply.

A huge amount of researches and tools can be found in related to sentiment analysis of comments published on social media. This basic sentiment analysis would categorize the comments as positive, negative and neutral. The solution proposed would move a step forward and take sentiment analysis deeper by categorizing comments into six main emotions, namely, happy, sad, angry, disgust, fear and surprise according to Paul Ekman. [4] We use vector space model for this classification and by using latent semantic analysis we try to increase efficiency of emotion classification algorithm.

All the previously conducted researches on sentiment analysis and opinion mining, just focuses in identifying and separating the comment which contains a suggestion. Again, the proposed solution would take semantic analysis deeper by giving the weightage of the suggestion considering amount of times that suggestion has been repeated throughout the comments. Further, for the convenience of the users of the proposed solution, the suggestions would be summarized as feature, polarity pairs where polarity would be either positive or negative.

Since every social media analyzing tool rely on basic measurement scores namely number of likes, shares and comments, proposed solution would build advanced relationships beyond the common patterns by manipulating relationships with human emotions which extracted from the comments. By calculating probabilities and measurements of all the possible social key performance indicators(KPIs) with real time, success rate of the post can be derived more accurately. Furthermore, the graphs and diagrams with statistically descripted charts can be produced to the user.

Following are the list of features to be examined or produced by proposed solution. Even though some features are available on other analytical tools mentioned below, the way they monitored and data engagement with social media metrics are much more simple and direct without digging into further analysis. Besides, getting number of

likes, comments and shares are most common feature of all the time and export them into a PDF or reporting mechanism is another common area found.

*Table 1: Comparison of Features*

| Features to be examined | Existing Social Analysis Tools | | | | | | | Proposed research tool |
|---|---|---|---|---|---|---|---|---|
| | Sprout Social | HowSociable | SumAll | Quintly | PeakFeed | Pulsar | discovertext | |
| Reporting solutions for the analyzing data to make impact of the business more reliable | Yes | No | No | Yes | Yes | No | Yes | Yes |
| Statistical details based on the days in a dashboard with graphs and diagrams | Yes | No | Yes | Yes | No | Yes | Yes | Yes |
| Comment analyzing and identifying the human emotions labeling | No | No | No | No | No | No | No | Yes |
| Displaying the success rate of the post by using advanced relationship with engagement | No | Yes | No | No | No | No | No | Yes |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| of the metrics | | | | | | | | |

### 1.3. Research Problem

Predominantly, the application intended to be developed is designed to maximize the profit received by an enterprise through social media. With the advancement of technology in information and communication sector, millennials are electronics-filled, increasingly online and socially networked, which drew businesses into digital marketing making social networks the best platform to communicate with customers. Even though the social media has helped to broaden the range of the customers reached by enterprises, business professionals have a difficult situation in deciphering customers. Due to this it is difficult for them to decide what precisely needs to be done and not in future which position them in a situation with an indecisive next move. This can be a very strenuous situation as customer understanding is the foundation to the success of a business.

Although what discussed in the above paragraph is the main problem faced by the enterprises which requires the aid of the application to be built, it can be specifically break down into following problems,

- **Lack of time to peruse customer feedback**

As a result of the competition prevailing presently, business professionals lack time to focus their attention on every single task that requires their attention. One such task is perusing customer feedback on their social media accounts which would make them understand customers' wishes. Enterprises' main goals include to understand what precisely customer needs and wants are extracted from their social media feedback.

- **No trend to analyze comments**

Social media provides a platform for both likes and comments and unfortunately enterprises only focus their attention on the amount of likes they gain for a post posted on social media. As comments are much more expressive than a like can ever be, a vast amount of data can be accumulated from analyzing them which would contribute immensely to the prosperity of the business.

- **No single platform providing all**

There prevails a numerous amount of social media analytics tools used globally. It is shameful that those analytics tools provide dissimilar features from each other. Enterprises must use more than one analytics tool to get requirements met which is cumbersome.

# 2. OBJECTIVES

## 2.1. Main Objective

- Improve achieving business goals of the company from social media product advertising by getting the idea of the post more clearly using advanced techniques.
- Assist to take future business decisions wisely by conveying suggestions for the next action that should be taken by the company.
- In excess of delivering manual human reactions towards the advertisement, prepare a better resolution of the post in less time consume by providing the success rate and other descriptive statistics of the post.
- Providing an effective customer service by delivering a descriptive report of the analysis other than managing only the campaign to the businesses via web marketing companies.
- To host a common platform to all business enterprises that use social media as their marketing strategy to make business decisions.
- To benefit the business by improving marketing in social media with the data gathered from posts, rating the success and delivering predictions beyond the human level.

## 2.2. Specific Objective

Application of key research areas in the project and application of techniques in the relevant key areas are discussed as follows.

### 2.2.1. The key objective aimed (1)

- Improve achieving business goals of the company through social media product advertising by getting the idea of a post more clearly using advanced techniques.

#### 2.2.1.1. The research areas/pillars used as optimal

- Data mining and data gathering through analyzing the social media posts which its objective is to rehabilitate the unstructured data and replace with the systematic pattern, analyzing the comments by identifying human emotions by

semantic analysis, qualitative methodologies and machine learning algorithms. Dividing the comments of a post into human emotion categories by manipulating sentiment analysis.

#### 2.2.1.2. Method of achieving the stated specific objective

- Comment of a post can be divided into three major categories which namely positive, negative and neutral. This status could be identified and categorized the expressions or opinions of customers by using sentiment analysis.

- In advanced, the expressions behind the comments can be again categorized into human emotions such as happiness, sadness, anger, disgust, fear and surprised by applying semantic analysis and some of the qualitative methodologies as content analysis.

- Preprocess data sets including words, terms; sentences which are related to human emotions categorized above can be identified using a machine learning algorithm.

### 2.2.2. The key objective aimed (2)

- Assist to take future business decisions wisely by conveying suggestions for the next action that should be taken by the company.

#### 2.2.2.1. The research areas/pillars used as optimal

- Training the machine learning algorithm with pre-processed data sets with guidance of the semantic analysis, collecting keywords and terms to map and build a relationship in between them which can be categorized after into suggestion domain. Using word tokenizing methods filter and extract the suggestions, classify the positive and negative points of the post by using a machine learning algorithm.

#### 2.2.2.2. Method of achieving the stated specific objective

- Data mining through social media by analyzing the comments of the post filter the suggestions by training the machine learning algorithm with pre-processed data sets under supervision of semantic analysis.

- After data analyzing, keywords and terms are categorized into suggestion domain by mapping relationships between each of them.

- Filtered keywords and terms can be now extract separately from the domain and store them accordingly.

- By tokenizing the keywords and terms, a machine learning algorithm will classify the positive and negative points of each post.

### 2.2.3. The key objective aimed (3)

- More than delivering manual human reactions towards the advertisement, prepare a better resolution of the post in less time consume by providing the success rate of the post.

#### 2.2.3.1. The research areas/pillars used as optimal

- Establishing advanced relationships and probability of measurements among the metrics of the post and providing a rating measurement as final outcome.

#### 2.2.3.2. Method of achieving the stated specific objective

- Obtaining countable engagement metrics or social key performance indicators(KPIs) such as number of likes, comments, shares in Facebook and number of likes, retweets and comments in Twitter as for existing examples of the popular social media by calling relevant APIs.

- To build more advanced relationships between metrics, identifying the intimacy of relationships or ties based on time which can train into a prediction algorithm.

- Furthermore, acquiring human emotion labeling in comments which can be engaged with the previously defined relationships or ties will cause a better judgment of the post.

- Success rate and other possible attribute variances can be manipulated via a descriptive statistical report comprise of convenient graphs and diagrams.

# 3. METHODOLOGY

This section describes actions which will be taken to investigate our research problem and the rational for the application of specific procedures or techniques that will be used to identify, select, process, and analyze information applied to understanding the problem. This section will elaborate how the proposed solution will be implemented by using most appropriate solutions and relevant technologies.

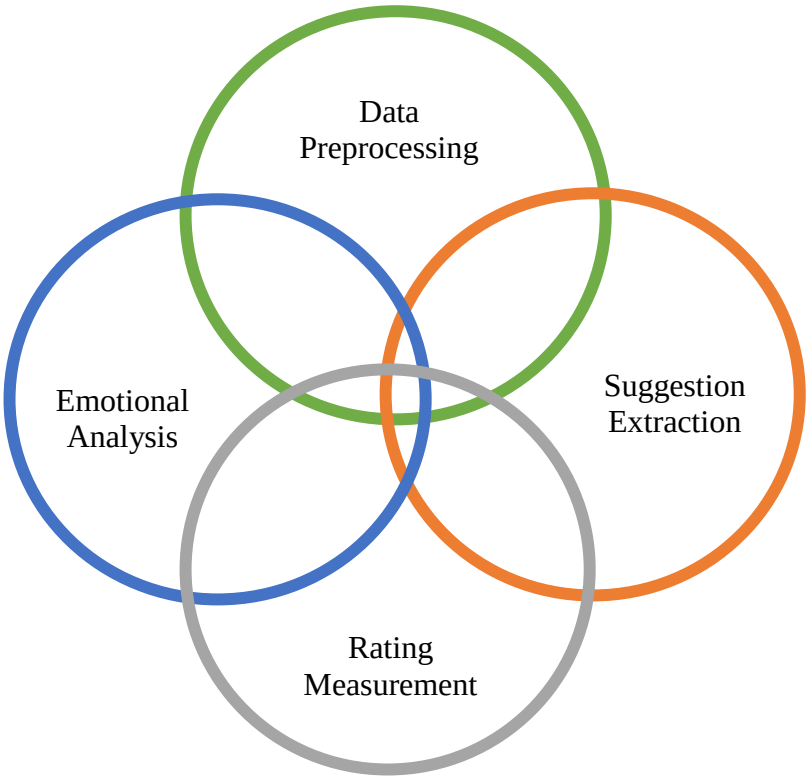


*Figure 1: Conceptual Diagram*

The above conceptual diagram shows the basic criteria in the proposed solution of the addressed problem in the outset. There are four main subparts which will be carried out by each individual proportionate to each criterion shown in the conceptual diagram. Flow of the processes that will be used to accomplish those individual subparts are discussed below.

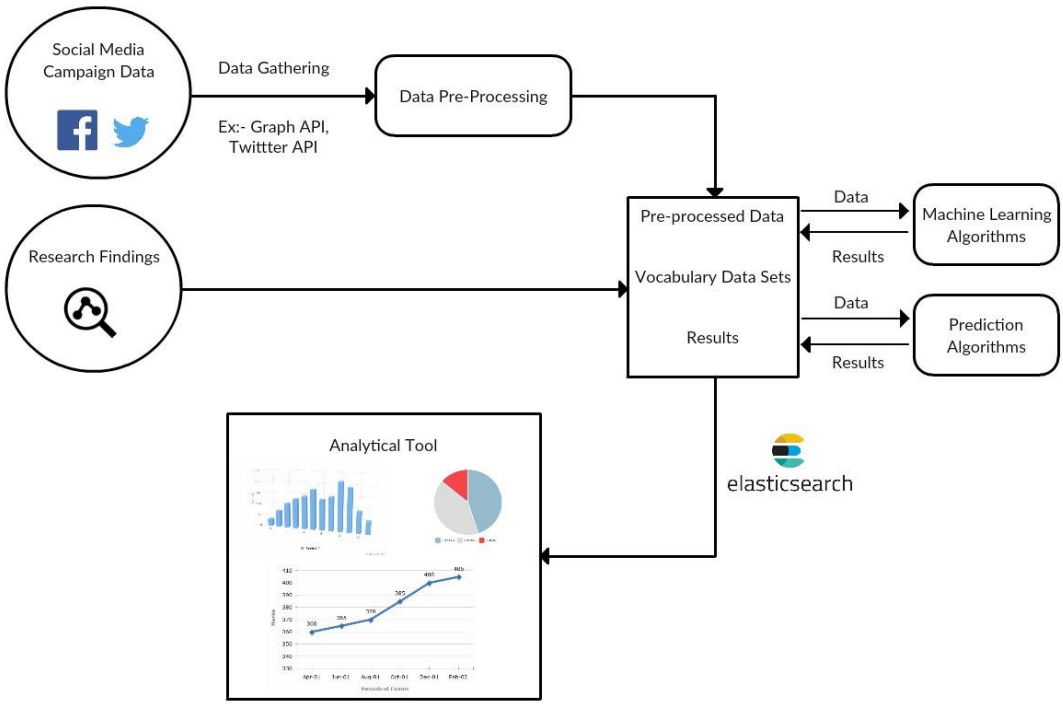


*Figure 2: Proposed High Level Architecture of the Solution*

### 3.1. Data Preprocessing and non-related Review Removal Mechanism

One of the most important parts in social media analytics is data preprocessing. Social Media data are texts performed by humans. Therefore, they are unstructured data which contains typos, non-standard acronyms and mutual meanings. It is very important properly clean these data before analyzing.

"The main objectives of data preparation are to process raw datasets, reduce time and space costs, enhance data quality with better interpretability and accuracy, and limit disclosure of sensitive information" [14]. There are four main steps in preprocessing data namely data collection, data cleaning, data reduction and data conversion.

Data collection is done from social network service providers. Social media websites such as Facebook and Twitter provide APIs for data crawling. As mentioned earlier there are many issues that needs to be fixed in these collected data. Few data mining techniques have been selected for data cleaning process, which will be suitable for the problem we are addressing, namely sentence splitting, stemming and lemmatization,

spell check, grammar checking, duplicate removal. Sentence splitting is the method of identifying where sentences begin and end. “The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form” [15]. An example of this can be demonstrated as follows:

***Am, are, is => be***

***Car, cars car’s, cars’ => car***

Result of this mapping will look like this:
***The boy’s cars are different colors => The boy car be differ color***

A special feature will be added to remove non-related data such as spam. This is a unique feature that will be implemented in the proposed system. In this feature implementation machine learning algorithms will be used to filter and remove non-related data. “In terms of spam precision, we can find that the **Naïve Bayes** method has the highest precision among the six algorithms while the **k-nearest neighbor** has the worst precision percentage and surprisingly the **Rough Sets** method has a very competitive percentage” [16]. Therefore, Naïve Bayes approach will be used for detecting whether some data are non-related or not.

In data reduction step feature extraction method will be used to extract desired information from the collected data. This method plays an important role in the machine learning and data mining. Feature extractions techniques are used for following reasons [17]:

- Simplifications of the data models to make easier interpretation.
- Reduction in the training time
- Reducing over-fitting.

With the use of this methodology we can overcome the issues such as higher computational cost that grows rapidly with the dimensionality increase of the data.

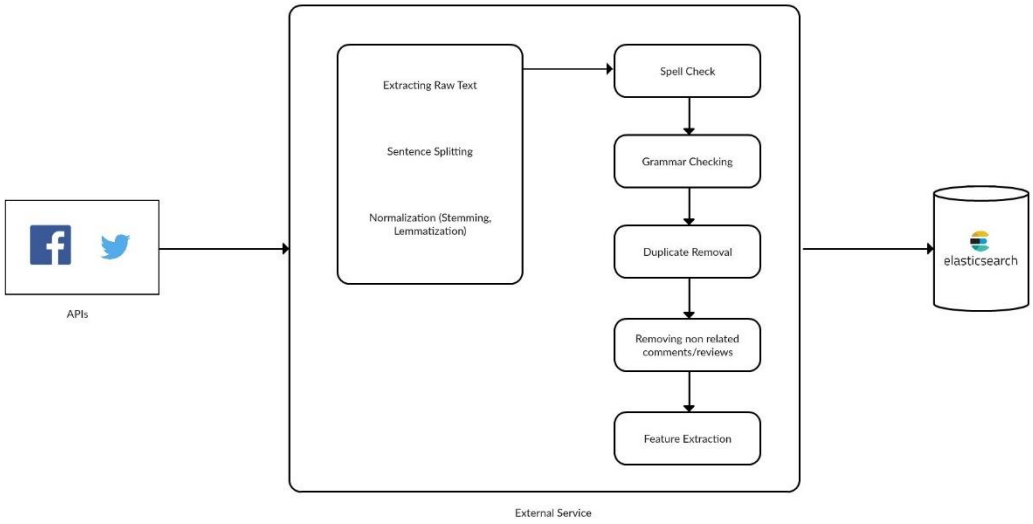


*Figure 3: Data Preprocessing Flow*

Finally, these extracted data will be stored in the Elasticsearch database. Elasticsearch database is selected as it is a real time distributed and analytics engine which will be very useful when implementing machine techniques upon those data. Moreover, Elasticsearch querying is really fast which is very important in analyzing data.

Data preprocessing is performed outside of Elasticsearch to make use of better flexibility and scalability. Elasticsearch's official Logstash can be used for different transformations of data such as pre-processing logs, converting them from an unstructured flat log-line into a structured better JSON document that is suitable for indexing [18].

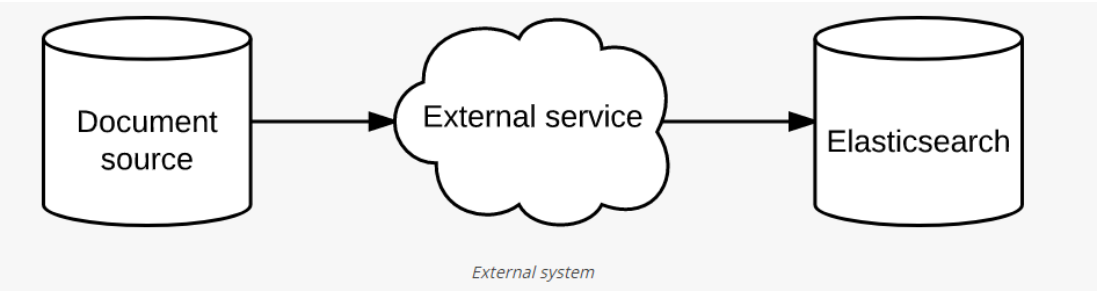


*Figure 4: Using data preprocessing as an external service to the Elasticsearch [18]*

## 3.2. Text Analysis and Emotion Categorization

Customers and consumers express their ideas by commenting on posts that are posted by business groups or advertisement marketing companies. Most of comments are grabbed with meaning and further we can narrow down human emotions from a comment. It is rarely we find comments with no meaning and no emotions. Human identify emotion from a text via its meaning and words related to various emotion categories. If we take a comment, we can divide it into three basic emotion categories such as positive, negative and neutral. We can do this kind of labelling using sentiment analysis. But if we provide only this analysis to our tool users, they get limited knowledge of human reactions and emotions on a particular post. This analysis does not touch actual human emotion categories.

Here our goal of emotion classification is to predict a single emotional label for given input sentence. Dimensional and categorical models are two significant emotion models when representing emotions. We conducted a research based on categorical model which assume there are discrete emotion categories. We mainly focus on Ekman's six main emotion categories namely happy, sad, angry, disgust, fear, surprise. In order to do this task, we use vector space model(VSM) for text classification. [19] In vector space model text is represented as word elements of a vector. According to frequency of words, each word is given a weight. Also, we can measure similarity between two vectors by calculating cosine value. We need a test set and a training set of data to create vectors. There are several organizations that provide datasets related to this domain. One of these organizations is ISEAR (International Survey on Emotion Antecedents and Reactions). Their dataset consists of 7666 text lines that are related to seven main emotion categories. We can split these training set and test set of data into words and represent as word elements of a vector. Then we can measure similarities between test vector and training vector to narrow down emotions.

Further we conducted our research on key technologies of Latent Semantic Analysis(LSA) algorithm. [20] By using this we can reduce the dimension of VSM. This technique is used to decompose term-document matrix into left and right singular matrix. This approach helps us to improve efficiency of text emotion classification.

Integrating emotion classifier to a social media analyzing tool gives capability to make wise business decisions based on human reaction and emotions.

### 3.3. Suggestion Extraction and Prioritization

Comments pave the way for the people to express anything they feel. These can either be just statements which would mean nothing in the perspective of the enterprise, opinions and suggestions. Opinions expressed in a certain way also can be considered as suggestions. Identifying customers' suggestions and being able to work on these helps to maximize profit gain of an enterprise.

Before filtering out the suggestions, it is necessary to understand what has been depicted in the comments. Therefore, as the first step, a vocabulary of the related terminology should be prepared. Keywords and terms that help to make suggestions can be collected relating to the enterprise domain. In other words, it is necessary to have a separate vocabulary comprised of words which gives out the meaning of a suggestion which can be used to identify a text as a suggestion.

As the next step, a thesaurus, which means words in groups of synonyms and related concepts, of the prepared vocabulary should be constructed. [9] This gives the advantage of identifying a suggestion in any of its forms and also map relationships between reviews and what is expressed by the customer and what truly needs to be said.

As the third step, it is necessary to understand whether the comment contains a suggestion or not. For this task, machine learning algorithms that detects suggestions with help of semantic analysis which is the process of relating syntactic structures to their language independent meaning, can be used. Pre-processed data sets which are related to the suggestion domain can be used to train machine learning algorithms. After identifying the suggestions, it can be filtered out. Those suggestions can be placed in a single line of comment or small paragraph and they can either be directly or indirectly pointed. It is better to re-filter it out for the purpose of providing a clear and easy understanding about it. Tokenization can be performed on the comments, to make the process of understanding effortless. Unlike traditional tokenization, in

semantic analysis, a sentence is broken into parts of speech which includes nouns, pronouns, verbs, adjectives, adverbs, conjunctions, prepositions, and interjections along with noun and verb phrases. [21, 22] These tokenized words can be mapped in a language-dependent manner to identify a language independent meaning. This identified meaning can be matched with previously constructed vocabulary to extract the suggestion. For this latent semantic analysis which means analyzing the relationships between a set of terms and documents by producing a related set of concepts, can be used.

Same suggestion can be suggested by different customers in different manners. Therefore, it would be useless for business personals to traverse through same type of comments which means the same thing wasting their time. So, the amount of repetition of a suggestion will be presented as a percentage. For this purpose, after the tokenization pattern matching would be performed on the newly tokenized comment with previously identified suggestions which would enable to recognize whether the suggestion has been mentioned previously. Here also latent sematic analysis can be used. A counter can be kept for each extracted suggestion and increasing the counter when the same suggestion is encountered when processing the comments for suggestions real time. This count can be taken as a percentage of whole amount of comments received.

Being able to understand suggestions from comments informs business groups to make future decisions based on customers' suggestions. This would walk them towards maximizing their profits as they would have the ability to provide exactly what the customer requires.

## 3.4. Rating Measurement and Success Prediction

The most frequently asked question of all the time by every business after posting a post is "what will happen next" and similar. Every business enterprise who uses social media as their marketing platform wants to know whether their post achieved the targeted audience or whether the customer is feeling awesome towards the post. Since this is all about the next level marketing, those customer feedbacks are way more valuable towards the profits which can be invested for the future. As stated above,

those feedbacks can be analyzed using social key performance indicators (KPIs) such as likes, comments and shares. According to Avinash Kaushik's four major social media metrics [23] we can measure those KPIs as follows.

1. Conversation rate – no of comments / replies per a post
2. Amplification rate – no of shares per post in Facebook and no of Retweets per Tweet in post
3. Applause rate – no of likes per post in Facebook and no of favorite clicks per post in Twitter
4. Economic value - no of sign up for site, no of reviews, no of created wish lists, no of calls which called to phone center of the company. In many instances this will state as value per visitor

Once we gather the data, we can perform the following score as follow [24]. Fundamentally it is a three-part formula which represents 1/3 of the total score.

Social media score = (Conversation rate + Amplification rate + Applause rate)/3

Moreover, if a business needs to know more about how applause rate has been come across of all the time, one's can change the formula where the particular rate is divided by 2/3 and change the other rates by dividing 1/6 each. Since the above calculation satisfies only the posts which performed best, the steps should be taken for the next best post remain still.

To achieve the above-mentioned issue, we build algorithms to calculate successfulness rate of a post based on human responses towards them. Without doubt we can directly take specific human responses or social key performance indicators (KPI) without any tough statistics. Nevertheless, in here, we are trying to research on more advanced measurements and relationships in between them. This indicates that we can discover relationships between defined time periods with count of responses and calculate success rate of the post. Further to achieve more, we gather the results of human emotions analysis as an illustration happy, sad, angry, disgust, fear, surprise from the comments. Afterwards we can build advance relationships between measurements as

particular post has 10 sad comments and 50 likes, 20 angry comments and 30 shares or 20 happy comments and 10 shares etc.

These relationships will be taken part of our final calculation and algorithms. Moreover, proposed analytical tool will provide user friendly graphs with statistical description about success rate and how attributes vary throughout the post campaign period to grasp the better understandability of human reaction. This may lead the next post to be the most successful post of all the time.

## 3.5. Gantt Chart

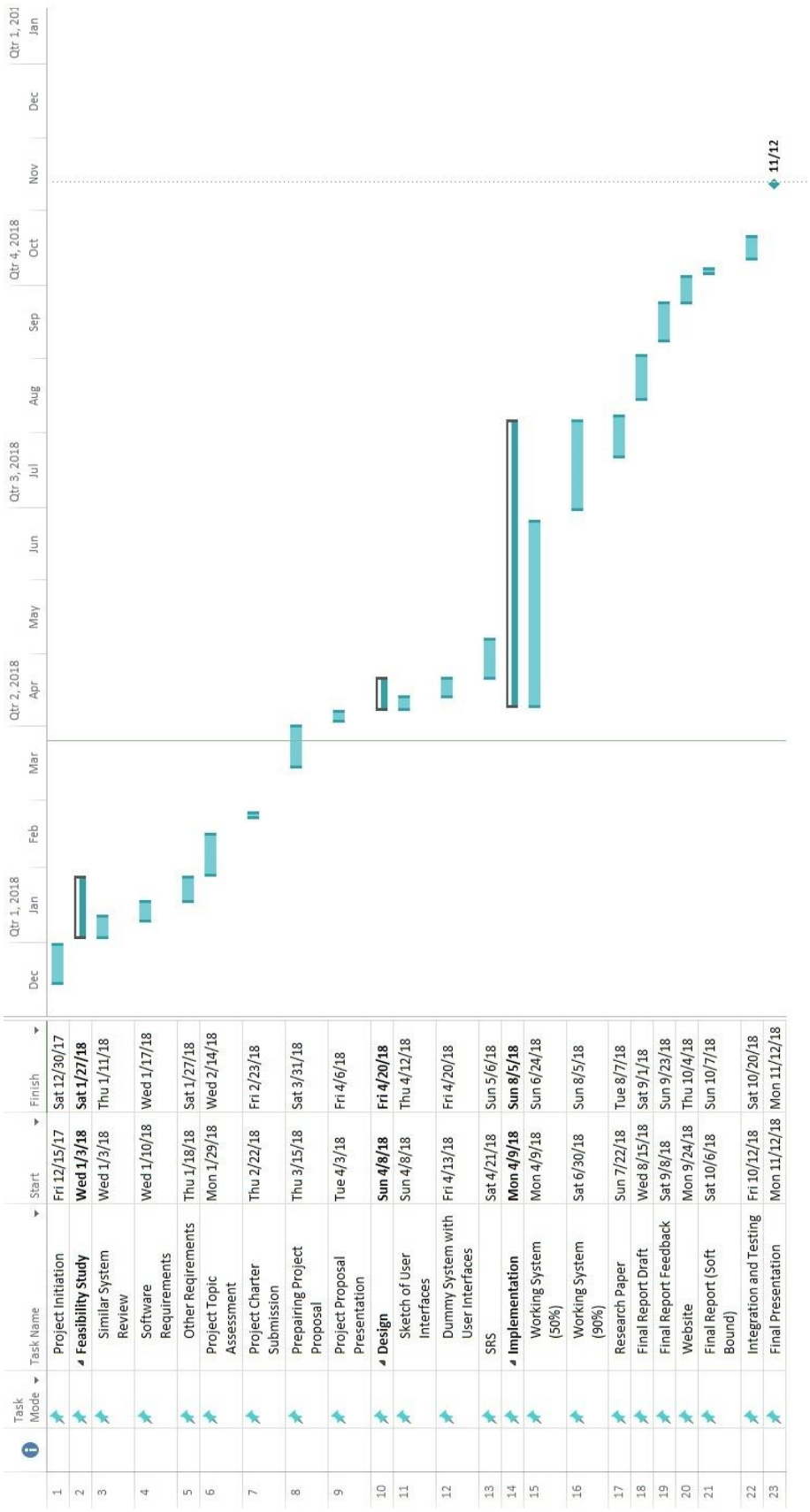


*Figure 5: Gantt Chart for the Project*

# 4. DESCRIPTION OF PERSONAL & FACILITIES

*Table 1: Personal & Facilities*

| | | |
|---|---|---|
| IT15087584<br>S.A.R.V. Silva | Data preprocessing and non-related Review Removal Mechanism | • Collection of relevant data from social media providers (APIs provided from Facebook and Twitter) for preprocessing them.<br>• Apply data cleaning methods on the collected data to detect and correct corrupt and inaccurate records.<br>• Use techniques such as feature extraction and spam detection to remove unnecessary and non-related data from the data set. |
| IT15035004<br>K. D. Y. Viduranga | Text Analysis and Emotion Categorization | • Building machine learning algorithms for identifying human emotion from comments.<br>• Labelling comments into emotion categories to get better understanding about customer responses.<br>• Display results in a graphical way in analytical tool. |
| IT15000354<br>H. N. Mirando | Suggestion Extraction and Prioritization | • Development of a vocabulary including words which helps to identify a suggestion.<br>• Building of an algorithm using machine learning algorithms to identify suggestions in a comment. |

| | | |
|---|---|---|
| | | • Extraction of suggestions from comments and storing them in an understandable manner.<br><br>• Get the percentage of suggestion repetition from entire comment set. |
| IT15003942<br>K. V. N. Withanage | Rating Measurement and Success Prediction | • Gathering social key performance indicators or metrics by calling relevant APIs.<br>• Identify and build advanced relationships and ties deeply among metrics.<br>• With the aid of machine learning algorithms, success rate and possible attribute variances are manipulated via graphs and reports. |

# 5. REFERENCE LIST


[1] “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].

[2] "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].

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

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

[5] Cs.waikato.ac.nz, 2018. [Online]. Available: https://www.cs.waikato.ac.nz/~eibe/pubs/ijcai15.pdf. [Accessed: 02- Apr- 2018].

[6] Ijcsi.org, 2018. [Online]. Available: https://www.ijcsi.org/papers/IJCSI-9-4-3-372-378.pdf. [Accessed: 02- Apr- 2018].

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

[8] 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].

[9] 2018. [Online]. Available: https://www.cicling.org/2013/rcs/Suggestion%20Mining_%20Detecting%20Suggestions%20for%20Improvement%20in%20Users_%20Comments.pdf. [Accessed: 04- Apr- 2018].

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

[11] Aclweb.org, 2018. [Online]. Available: http://www.aclweb.org/anthology/C12-2042. [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] Dreamgrow.com, 2018. [Online]. Available: https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/. [Accessed: 04- Apr- 2018].

[14] Ink.library.smu.edu.sg, 2018. [Online]. Available: https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=1100&context=etd_coll. [Accessed: 02- Apr- 2018].

[15] "Stemming and lemmatization", Nlp.stanford.edu, 2018. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html. [Accessed: 02- Apr- 2018].

[16] 2018. [Online]. Available: https://www.politesi.polimi.it/bitstream/10589/137564/1/Machine-Learning-Techniques-for-Social-Media-Analysis.pdf. [Accessed: 02- Apr- 2018].

[17] "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].

[18] "Document Processing and Elasticsearch | Elastic", Elastic Blog, 2018. [Online]. Available: https://www.elastic.co/blog/found-document-processing. [Accessed: 02- Apr- 2018].

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

[20] Lsa.colorado.edu, 2018. [Online]. Available: http://lsa.colorado.edu/papers/dp1.LSAintro.pdf. [Accessed: 04- Apr- 2018].

[21] Cs.tut.fi, 2018. [Online]. Available: https://www.cs.tut.fi/sgn/arg/klap/introduction-semantics.pdf. [Accessed: 05- Apr- 2018].

[22] "Grammar and Punctuation: The Parts of Speech - Aims Community College", Aims.edu, 2018. [Online]. Available: https://www.aims.edu/student/online-writing-lab/grammar/parts-of-speech.php. [Accessed: 05- Apr- 2018].

[23] 2018. [Online]. Available: https://www.kaushik.net/avinash/best-social-media-metrics-conversation-amplification-applause-economic-value/. [Accessed: 04- Apr- 2018].

[24] "How to Evaluate and Optimize Your Best Social Media Content", Social, 2018. [Online]. Available: https://blog.bufferapp.com/how-to-evaluate-and-optimize-social-media-content. [Accessed: 04- Apr- 2018].