S	
E1	
E2	
For Office Use Only	

Masters Project Final Report

(MCS)

2020

Project Title	Analysis of social media feedback to gain profit for business organizations using sentiment analysis techniques
Student Name	G.A.I.T. Wijewickrama
Registration No. & Index No.	2017/MCS/096 - 17440968
Supervisor's Name	Dr. M.G.N.A.S. Fernando

For Office Use ONLY

Analysis of social media feedback to gain profit for business organizations using sentiment analysis techniques

A dissertation submitted for the Degree of Master of Computer Science

G.A.I.T. Wijewickrama

University of Colombo School of Computing

2020

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: Gunarathna Arachchige Indusaranie Tharusha Wijewickrama

Registration Number: 2017/MCS/096

Index Number:17440968

_____

Signature: Date:

This is to certify that this thesis is based on the work of

~~Mr.~~/Ms. G.A.I.T. Wijewickrama

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr. M.G.N.A.S. Fernando

_____

Signature: Date:

Abstract

Contemplating the fact that business enterprises globally, benefit from social media to gain profits, this dissertation discusses the importance and impact of analyzing mechanisms of customer responses on social media to maximize the gained profit and take befitting future decisions. The main objective is to applying sentiment analysis techniques on the collected responses to analyze them and extract the customer feedback. This study focusses on finding out the best model for emotion categorization of the collected responses using the sentiment analysis techniques.

The collected responses are preprocessed and undergoes several techniques to achieve data normalization which enable fast and effortless querying. Cleansed data is analyzed in order to place the comments under six main emotion categories related to sentiment analysis. This analysis mainly consists with feature extraction and training the models. The two built-in methods available in scikit learn toolkit, namely; TfidfVectorizer() and CountVectorizer() are used for the feature extraction using both mono-gram and bi-gram features and those extracted features are used to train the selected four (04) classification models in order to find out the best performing model with the four (04) feature types extracted. Subsequently, this study will investigate and compare different features for the different classifier when categorizing emotions on social media. Then record the accuracy of each model with each feature type and select the best performing model with the accuracy of 68% with a 0.1 train-test accuracy difference.

Keywords: Social Media, Emotion Categorization, Sentiment Analysis, Feature Extraction

Preface

The basis for this research is the usage of social media for marketing purposes by various organizations. As the world moves in a digital era, new technologies and enhancements are invented day by day. In this digital age, each and every organization all over the world tends to use online approaches to do their marketing purposes.

But, are they using it effectively?

This research is to gain the profit of online marketing in an effective manner. To analyze customer responses in a more accurate manner and extract a future visionary feedback by using sentiment analysis techniques.

Acknowledgement

I would like to express my sincere gratitude to my supervisor Dr. M.G.N.A.S. Fernando for his guidance, motivation, and immense knowledge towards the successful completion of my research work. Moreover, his encouragement, insightful comments and guidance helped me in all the time of research and writing of this thesis.

Besides my supervisor, I my sincere gratefulness goes to our project coordinator Dr. Randil Pushpananda for his guidance given throughout the research work. Moreover, my thanks also delivered to the university staff and all the lecturers for the support given to successfully complete this research.

And finally, I would like to thank all the people who helped, supported and encouraged me throughout this research work.

Thank you for all your encouragement!

Table of Content

List of figures

List of Tables

List of acronyms

NLP	Natural Language Processing
NLTK	Natural Language Processing Tool Kit
TF	Term Frequency
IDF	Inverse Document Frequency
BoW	Bag of Words
POS	Part of Speech
SVM	Support Vector Machine
LRM	Logistic Regression Model
NBM	Naïve Bayes Model
RFC	Random Forest Classifier
TN	True Negatives
TP	True Positives
FP	False Positives
FN	False Negative

Chapter 1- Introduction

1.1. Introduction

With the advancement of technology in the information and communication sector many organizations use online marketing for their marketing purposes. Among them, social media marketing is used by most of the organizations. Under this situation, it is an important requirement to find the customer interaction and their opinion towards the organization using the feedback received by each post in order to maintain a good customer base around the organization. If this is done via a manual procedure, mainly it will be less efficient.

Also, most of the organizations are wasting their time in finding out the most appropriate advertisement for a certain project/ task due to no proper mechanism of grading the advertisements by using the received feedback. Based on identified two problems this study was to introduce a solution for the organizations to find out the best approach which help their business growth.

As well as manual procedures are less efficient, the text classification just to identify the negativity and positivity of the customer feedback are somewhat outdated. Now, the organizations are craving for more accurate and working solutions to build up their organizational reputation and dignity with the user of online marketing strategies. Therefore, this study is focused on text analysis for emotion categorization, something more than detecting the negativity and positivity. Hence, this study will 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 [9].

1.2. Problem Definition

Usually, every organization should have an appropriate method to launch their marketing strategies and measure the customer interaction for each and every advertisement/ post. It is an appropriate way to delight the customers and improve a good customer base around the organization. It is very

essential to determine the relationship of marketing approaches with the customer requirement and strengths since the customer reach is directly affect the organizational successfulness.

If consider Facebook as an example, organizations should be able to track/ view information of competitors, to identify the people who interact with the posts published in the page and to have a centralized analytic tool, which provides statistical data and reports covering a vast scope which are required for organizational development [1].

Even though social media has helped to broaden the range of the customers reach, business enterprises do not have an exact way to group the customers based on their choice. 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.

The usage of comments and opinions for a particular post has increased due to the enormous internet usage and ease of remote accessibility. Therefore, it is difficult to analyze the existing online information manually as well as the existing negative and positive level analysis techniques to make an accurate decision. Therefore, it is evident that it is hard to develop an efficient and effective sentiment analysis technique and that it should be able to summarize the sentiments of customer feedback automatically to enhance the decision-making process of organizations and to take correct decisions [2, 3].

If there are several comments/ feedbacks available on social media for a particular post, it is very difficult to take a correct decision based on available comments due to the non-availability of reliable numerically weighted quality factors since most of them provide only positive and negative quantifiers [4].

Further, it was revealed that there is a lack of researchers to address how to analyses the user feedback rather than depend on negative and positive quantifiers.

This research project is based on analyzing the comments for advertisements published in selected social media accounts of the organization (i.e.: Facebook, Twitter) by depending on the customer reach. Initially the framework will be able to follow the correct procedure to evaluate the customer interaction by analyzing the comments for each post. When evaluating the ranking of an advertisement, every company must focus on the customer interaction. Customer interaction is a basic measurement of the organization’s popularity and the ability to survive in the field. Therefore, evaluating the advertisements of an organization is important.

Due to the above problems and concerns, we are suggesting a better approach to evaluate the rankings of organizational marketing approaches using customer responses.

1.3. Motivation

As stated in above *Problem Definition* section, organizations do not have an effective customer interaction computation methodology and their point of view towards the organization by analyzing the comments/ customer feedback for a given advertisement/ post. Still there is not an exact way to extract customer emotions from the feedbacks. It is important to fill this gap with a possible modern approach. Therefore, the motivation is to take the initial step of reducing this gap by analyzing the customer feedback and categorize them relying on their choice.

1.4. Exact Computer Science Problem

A considerable amount of research has been carried out under text classification for social media analysis. These research solutions are widely used in day to day scenarios in order to provide best results to the consumer. But most proposed approaches take a considerable time to analyses and classify the feedback since the approaches used to analyses the comments, calculate word count and feed those values to the classification data mining or machine-learning algorithms for classification.

Thus, when the vector size (number of words) as well as the count of data points (number of comments for a selected post) get high, it has to be done an expensive calculation of distance measures. Therefore, it would become a time-consuming task as the classification algorithms and feature extraction processes get more complex. Hence as mentioned in the above *Problem Definition* section, organizations are still struggling to find a better approach to do their marketing and grow the customer base around the organization.

The computer science problem that is going to be addressed by this project is, find a good approach of text analysis in order to classify the comments of posts published in social media and analyzing them using data mining and machine learning techniques, which are lightweight, accurate, efficient and appropriate to the problem scope.

1.5. Research Questions

1. Is it possible to categorize human emotions in social media automatically by analyzing comments posted by users?
2. How to use a lexicon-based approach for emotion categorization?
3. How to use a machine learning approach for emotion categorization?
4. Can we find a combined solution by combining lexicon-based approaches and machine learning approaches?

1.6. Goal and Objectives

1.6.1. Goal

Let the businesses to to understand customer insight towards their products so that they can hold their products to gain maximum profits.

With traditional media of marketing, it was not easy to gather these important data to understand customer needs. Social media has made a huge and important impact on marketing with this fact. Businesses can use this information to analyze and understand customer needs and opinion towards their products.

1.6.2. Objectives

The objectives which need to cover in order to achieve the above-mentioned goal is listed below.

1. Do a review on existing text analysis methods and algorithms regarding the problem domain.

2. Analyze the existing feature extraction and classification algorithms in order to find the best performing methodology by comparing the effectiveness of different methods

3. Develop an effective machine learning algorithm for texts analysis in order to provide a better image on customer interest and interaction by analyzing comments

1.7. Research Methodology

Research methodology follows quantitative approach. In the first phase the literature review is done on different approaches used for the analyzing social media comments for success prediction. There consideration was mainly on the dataset preparation, experimental set up of the study and evaluation of the research work.

In the second phase of the research a model for text analysis will be built using machine learning algorithms used in literature. Newly created data set will be evaluated with the model.

1.8. Scope and Limitations

1.8.1. Scope

The scope of this document is to deliver an effective solution to cater business needs and goals by analyzing customer responses to individual posts published on official social media accounts. Intended inputs for the classification algorithms would be the amount and the content of feedback. The feedback will be the comments provided by customers for each post.

After analyzing the comments, the system would be able to predict the success of the post to help the enterprise to take their decisions wisely. The solution will be able to provide a complete analysis covering the 360-degrees as the customers interact with the posts.

1.8.2. Limitations

1. Even the proposed solution is based on multiple social media data sources mainly the attention will be focused on Twitter and Facebook as they are placed in the top five results of the most popular social networks.
2. The solution will be implemented only focusing on English language since the capacity of Sinhala responses are not enough to carry out the research.

1.9 Outline of the Dissertation

The thesis is structured as describe in this section.

The Chapter 2 includes the literature review explaining related work and the identified research gap. The Chapter 3 contain the methodology, explaining how we are going to handle this study. The implementation details following the points discussed in methodology section is included in Chapter 4 while the Chapter 5 presents the evaluation criteria with results. Finally, the Chapter 6 will conclude the study and will provide guidance to the future work.

Chapter 2 - Literature Review

This chapter provides a detailed description of the results obtained from the analysis of techniques used in previously conducted researches for analyzing social media comments for emotion categorization using sentiment analysis techniques. This chapter includes brief explanations on the results obtained by using different preprocessing techniques, feature extraction methodologies and classification algorithms. According to the number of researches emerging recently, it is clear that there is a significant attention towards analyzing social media comments for emotion categorization.

Mainly two main approaches were identified for this task as lexicon-based approaches and machine learning based approaches. And in most of the cases researchers have come up with a hybrid approach combining both lexicons based and machine learning approaches. Currently very few researchers have focused on deep learning approaches.

2.1 Lexical based/ Rule based Approaches

Lexical based approach is an important part of a text classification task which use to understand the lexical phrases. Different patterns of language, grammar and rules are fed to the machine in order to identify the lexical phrases. In this approach, the vocabulary is more important than grammar.

The lexical analysis has inbuilt lexicons which are widely used in data analytics. Each of those lexicons are assigned with a word and with a polarity rate which describe whether that word gives a positive, negative or neutral meaning. But, still there is no proper lexicon analysis method for social media feedback analysis. Already it contains a collection of words for feedback analysis, but it does not provide a polarity rate describing the amount of positivity or negativity.

Moreover, still there is no identified lexical analysis method for social media feedback analysis for emotion categorization, which is the domain we are discussing in this study.

2.2. Machine Learning Approaches

In machine learning algorithms, there is no need of defining rules by the programmer to do a specific task. The only thing needs to do is feed the relevant data to the machine learning algorithm and then, the algorithm will adjust itself to perform the task. This data driven approach is widely used in different domains of computer science, mainly in data analysis.

This has two (02) main stages namely, supervised learning and unsupervised learning. This depend on the labelling of the data set. If the input data set is labeled, then it is supervised learning. If the data are without labels, then it is unsupervised leaning.

In supervised learning, the data set is split into two as training data set and testing data set. Then the learning algorithm train the model using the data set and make the predictions using the training data. In the training phase, these steps are continuously repeating until the model is able to achieve a considerable level of performance. The next step is the testing phase. In this phase, the model creates the predictions using testing data set and calculate the performance.

Support vector machine, Naïve Bayes classifiers, Logistic regression models are few examples of supervised machine learning algorithms and k-means clustering and self-organizing maps are examples for unsupervised machine learning algorithms.

Moreover, social media feedback analysis is mostly targeted with supervised learning and unsupervised learning for this is relatively very law.

2.3 Hybrid Approaches

This is the approach used by most of the researches for the task of social media feedback analysis. Hybrid means, this approach uses the combination of machine learning approaches with lexical based approaches.

In some scenarios, lexical based approaches are used in first part of the study and then practices a learning-based approach. That means, researchers used lexical based approaches for data

preprocessing and normalization and then uses the machine leaning approaches to train the model, by feeding the normalized data into a machine learning model.

Moreover, in some other scenarios lexical based approaches are used to achieve both normalization and feature extraction and then uses the machine leaning approaches to train the model.

2.4. Related Work

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 the 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 sharing 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 feedback through social media platforms assist to analyses 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 the next level.

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. 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 [5].

Many researches can be found regarding different data pre-processing techniques that will be used to clean raw data and generate a structured data set.

One research that could be found, which is about data pre-processing is “A natural language normalization approach to enhance social media text reasoning” by Long Hoang Nguyen, Andrew Salopek, Liang Zhao and Fang Jin [6]. Primary focus of this research was to enhance text understanding quality by using more sophisticated data pre-processing techniques. Filtering, Removal of special characters, Removal of stop-words, 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 [6]. This issue is also related to the fact that this system is too generalized.

Another research regarding feature selection and extraction in data mining is the solution proposed by Aparna U.R. and Shaiju Paul, which stated how particular features can be collected when there is a large number of features available in the data pool by reducing both reduction in training time and over-fitting [7]. 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 into 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 nonlinear [7]. 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.

Although there is the possibility for people to express 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 reacting to a post as it is

limited to some emotions. Analyzing what has been commented and categorizing them into different emotions would pave the way for a better understanding of the customers.

When labelling comments into emotion categories, “Remembering emotional experiences: The contribution of valence and arousal” by Elizabeth A. Kensinger [8] provides a solution with 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 [8]. 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 the categorical model, it assumes their discrete emotion categories such as Ekman’s six main emotion categories [9] namely happy, sad, angry, disgust, fear and surprise.

The research “Positive, Negative, or Neutral: Learning an Expanded Opinion Lexicon from Emoticon-annotated Tweets” [10], which is conducted by Felipe Bravo-Marquez, Eibe Frank and Bernhard Pfahringer, about sentiment analysis on Twitter social media platform to label tweets into positive, negative and neutral categories. They provide a supervised framework for expanding an opinion lexicon for tweets. 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 [10].

Another research is conducted by Akshi Kumar and Teeja Mary Sebastian to propose and investigate a paradigm to mine the sentiment from a popular real-time micro-blogging service and Twitter, where users post real time reactions [11]. 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, pre-processing module and Scoring module. Retrieval module consists of data APIs. In their case they have used Twitter API. Data pre-processing modules mainly concern techniques like removal of URLS, spell corrections, Emotion tagger, POS tagger operations. Finally, 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 [11].

The research has been conducted on supervised vector space model (VSM) for indexing the document, finding word count, information retrieval, information filtering [12]. The attempt was to 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.

A customer can compose a comment to express opinion about the organization and their related products or to give advice or warnings about the organization and their related products to other customers.

According to "Extracting implicit suggestions from students' comments - A text analytics approach" [13] by Shankararaman Venky, Gottipati Swapna, Lin Jeff Rongsheng, and Gan Sandy, 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 [13]. 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, according to the solution provided by Caroline Brun and Caroline Hagege [14].

Recently, opinion mining and sentiment analysis has sprung 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 research on these terms. While on the hunt to gain knowledge about these trending research areas, below mentioned research works are managed to capture the attention.

First and foremost, the research conducted by to extract suggestions from students' comments which provided feedback on course and the instructor at the end of a semester [13]. 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. 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 Models (GLM), Support Vector Machine (SVM) and Conditional Inference Tree (CTREE) have been used [13].

Secondly, research was conducted with the intention of detecting suggestions and improvements comprehended in user comments [14]. 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.

"Towards the Extraction of Customer-to-Customer Suggestions from Reviews" is a solution provided by Sapna Negi and Paul Buitelaar in order to extract customer-to-customer suggestions from reviews [15]. 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.

Further, the research conducted by Swapna Gotipatti and Jing Jiang proposed a solution to extract and normalize entity-actions from user comments [16]. They have identified that a person has two intentions of writing a product review. They are; to talk about the quality of the product/service and to help other customers to decide about the product/service before purchase/use. 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.

2.5. Research Gap

During the literature review and information hunting, a range of research and tools developed to achieve the task of 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 research and developed tools.

One feature that came to attention was pre-processing of data. All the solutions proposed in conducted research and tools, follow a series of pre-processing techniques before data storage [6, 17].

In concern of semantic and sentiment analysis of comments, what can be seen is that, although there are many researches were 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 research 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 [9].

All the previously conducted research on sentiment analysis and opinion mining, just focuses on 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 the amount of times that suggestion has been repeated throughout the comments [8].

Since every social media analyzing tool relies on basic measurement scores namely number of likes, shares and comments, proposed solutions would build advanced relationships beyond the common patterns by manipulating relationships with human emotions which extracted from the comments.

2.6. Problem Statement

When analysing the Problem Definition described in Chapter 1 relying on the above described Research gap, the problem statement can be described as follows.

Now a days, it is very easy to find a huge number of tools to analyse social media comments and provide the feedback on them. Most of the approaches are combined with number of likes, shares and comments. Moreover, there are many analytic tools including text classification. But, most of them are either hate detection or sentiment analysis just to categorize text as positive, negative or neutral. It is very rare to find a sentiment analysis approach targeted for emotion categorization.

The main objective of this study is to find out an optimal solution for emotion categorization using sentiment analysis techniques to provide a working solution for the organizations who crave success through the efficiency and with the technology.

Chapter 3 - Methodology

This chapter describe how we are going to handle the study based on the knowledge gained from the literature review in above chapter 2. Based on the data gathered from the survey, many experiments and strategies have been taken into consideration to meet the objectives of the research. As mentioned in earlier chapters the main inspiration of this research is providing an efficient and scalable text processing approach for emotion categorization.

3.1. Problem Analysis

The aim of the study is to find a real-time and effective solution to analyze social media feedback using machine learning algorithms. The solution was carried out in two phases. First phase is preprocessing and normalizing the data set and extracting the features from the normalized data set using different feature extraction mechanisms. The second phase is selection and training a classifier to analyze social media feedback in order to categorize the emotions.

Social media feedback analysis is a key component to measure the existence and power of any organization within the industry. It helps the customers to find appropriate service providers to achieve the desired task and it can be helpful to grow the customer engagement and the successfulness of the organization. As well as for customers, it helps the management to decide the rankings of the organization and enhance the entire organizational framework in order to increase the customer growth. According to the literature review, most of the organizations following a manual process to analyze customer reviews or they are using a computerized approach which can be used to acquire only few criteria which are identified in the context of text analysis. This is not a user-friendly task for any kind of organization, since management must keep track of every past record manually.

3.2. Proposing Model/ Design

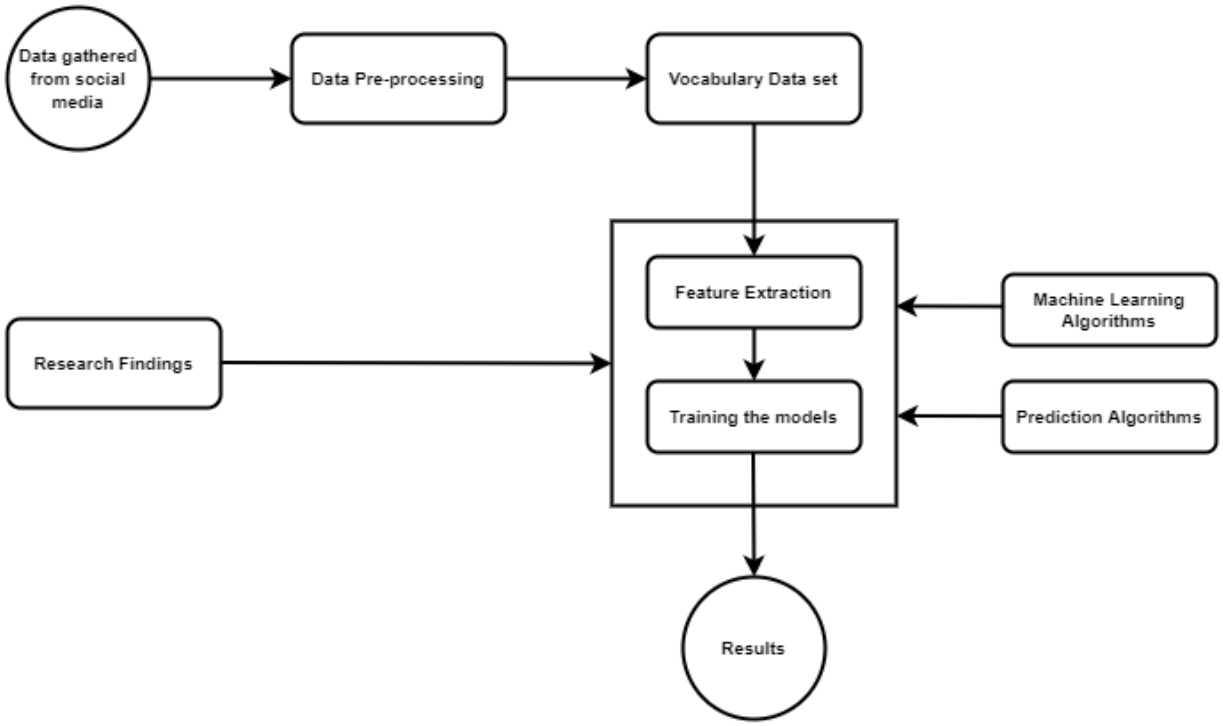

Figure 3.1 Flow Diagram of the research work

Above Figure 3.1 explains the work carried out in different phases of the research work. There are four main components in this research work, which are data gathering, data pre-processing, feature extraction and training the model for emotion classification. In the first phase, the focus was placed on Twitter and Facebook social media platforms and data gathering. Here, data can be specialized as comments extracted from social media related to skilled visa migration category.

As the second phase, those comments would be pre-processed and would be used in machine learning algorithms as both training and testing data sets.

Then the features will be extracted from this pre-processed data set by using the Tf-Idf vectorizer and Count vectorizer.

Finally, the extracted features will be used to train the selected models to extract human emotions out of the input text data. The models will be; Logistic regression model, Multinomial NB model, Support Vector machine model and Random Forest classifier model.

The foundation of this research work is categorizing user comments into seven emotion categories; namely; angry, happy, sad, disgust, neutral, surprise. This is based on supervised learning where there is an already labelled data set for each emotion category for the training purpose. User comments go through this service and classify into the emotion categories they belong.

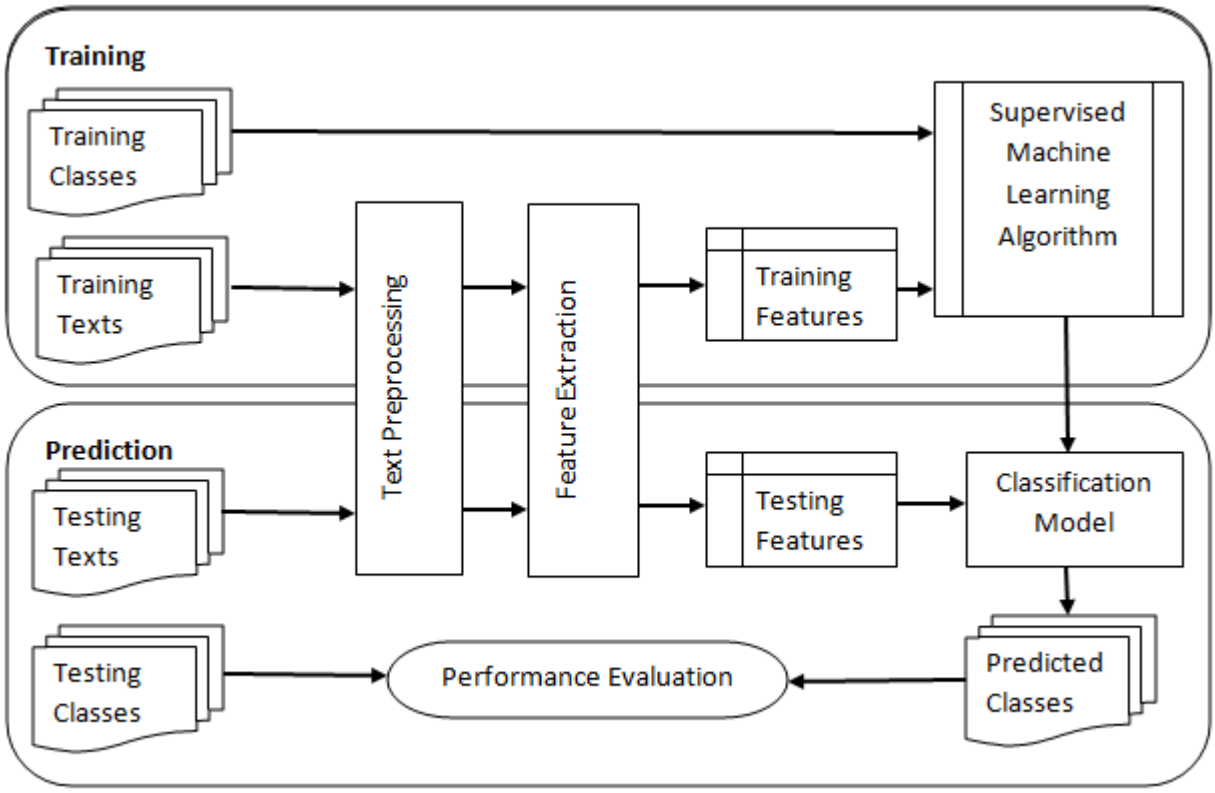

Figure 3.2 System Design for Supervised Models

The high-level design for the application of supervised machine learning algorithm is displayed in the Figure 3.2. Each step of the given process will be explained in detail in the latter part of the chapter.

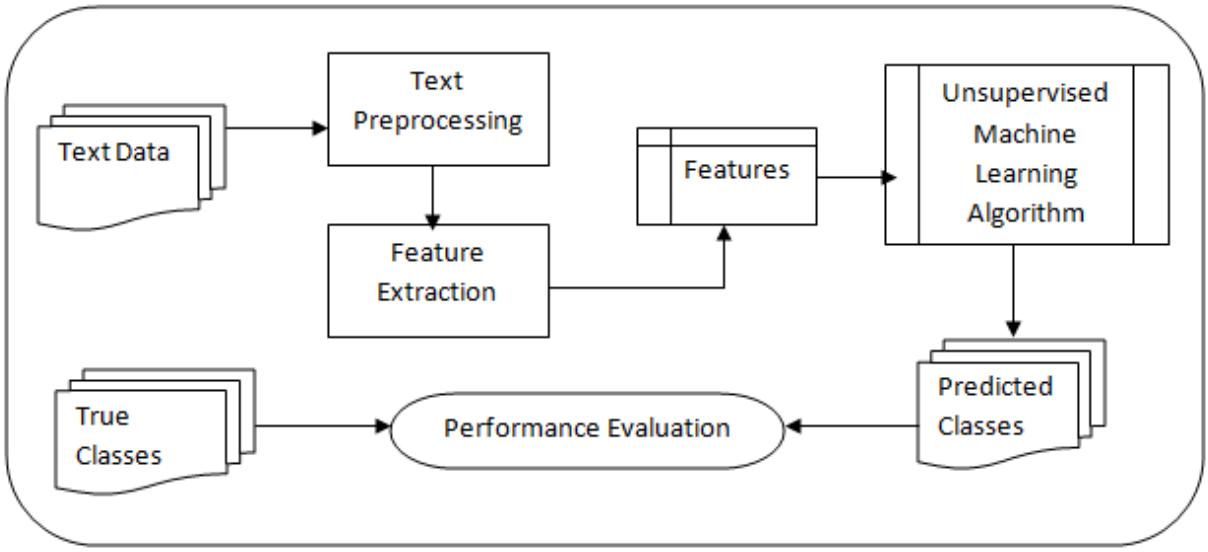

Figure 3.3 System Design for Unsupervised Models

The high-level design for the application of unsupervised machine learning algorithm is displayed in the Figure 3.3. In this approach, all the steps are same as the steps followed in design presented for supervised machine learning algorithm except the learning procedure.

3.3 Data Set Creation

3.3.1. Data Set Preparation

To perform a successful experiment on text analysis the availability of a rich corpus is important. Further, in order to deal with supervised leaning algorithms, a labeled corpus should be available.

The collected social media comments from Facebook and Twitter under student visa migration category are included in the data set.

The data set consist of 7000 comments and out of them 5600 (80%) as training data and other 1400 (20%) as testing data. The data set is manually annotated with the seven (7) emotion categories.

Training Data Set	5600 comments
Testing Data Set	1400 comments

Table 3.1 Categorizing the Data Set

3.3.2. Data Pre-processing

Comments are retrieved to analyze and get the required output. Comments in a post are texts performed by humans. Therefore, they are unstructured data which contains typos, non-standard acronyms and mutual meanings, which makes it crucial to perform proper data preprocessing. There are four main objectives in data preparation namely processing raw data sets, reduce time and space costs, enhance data quality with better interpretability and accuracy, and limit disclosure of sensitive information [6].

Further, four main steps of data preprocessing are called data collection, data cleaning, data reduction and data conversion.

After collecting these data, they undergo multiple data preprocessing techniques such as sentence splitting (tokenizing), stemming and lemmatization, spell check, grammar checking, stop words removal and special characters removal.

Sentence splitting is used to convert raw text into sentences to get more meaningful information out.
When splitting sentence,

Sentence boundary = period + space(s) + capital letters

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.
Am, are, is => be

Car, cars car's, cars' => car

The result of this mapping will look like this:

The boy's cars are different colors => The boy car be differ color

For the training purpose of the model, it is required to prepare a dataset which is understandable to the computer. Feature vectorization techniques are used to achieve this.

3.4. Feature Vectorization

In order to achieve the feature vectorization, can use the methods such as Bag of Word (BoW) method and TF-IDF methods. In machine learning, feature vectors are used to represent numeric or symbolic characteristics, called features, of an object in a mathematical, easily analyzable way [11]. A familiar example for feature vectorization is the RGB color description. A color is described according to the amount of red, green and blue in it. In our case, this model is used to represent the comments in a way our machine learning model would understand it.

3.4.1. Bag of Words

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

Table 3.2 Dictionary

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

Table 3.3 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. As well as, with the number of comments (data set) the size of the dictionary becomes bigger. It can be considered as a drawback of the BoW model.

3.4.2. TF-IDF

3.4.2.1. Term Frequency (TF)

Term Frequency is used to represent each term in vector space. It is a measurement of how many times a word present in the vocabulary. It can be denoted by the Equation 3.1 stated below [18]:

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

Where $fr(x,t)$ is define as:

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

Equation 3.1 Equation for Term Frequency

This equation returns the number of times that term t present in document d.
Assume there are 3 documents and n times of terms (words) in our vocabulary, below equation display the representation of document number 3 and terms as vectors.

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

$$V\ (d3) = (8,0,6,0)$$

As aforementioned the above example carries the term t1 which appears in 8 times in document 3 and term t2 appears in 0 times in document 3. Using vectors, we can show a list of documents in our dataset. For further calculation, the final result can be turned into a matrix form such as D *F shape. D is no. of documents in the space or our dataset. F is no. of features which are sizes of individual features that are appeared in the vocabulary. Normally this matrix contains a lot of zero values. We called it as ‘sparse’ matrix. We can show a sparse matrix and its shape as below.

$$M|D| * F = \begin{matrix} 8 & 0 & 6 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 1 & 0 & 2 \end{matrix}$$

Above feature matrix explains term frequencies of 4 terms (features) in 3 documents.

3.4.2.2. Inverse Document Frequency (IDF)

Inverse Document Frequency (Idf) value can be calculated using the Equation 3.2 stated below.

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

Equation 3.2 Equation for Inverse Document Frequency

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.
So, we have 4 features in our sparse matrix. We should calculate idf(t1), idf(t2), idf(t3) and idf(t4). Below examples denote how we can calculate 'idf' values of t1 and t3.

$$idf(t1) = \log \frac{|\ D\ |}{1 + |\ \{d : t \in d\}\ |} = \log \frac{3}{3}$$

$$idf(t3) = \log \frac{|\ D\ |}{1 + |\ \{d : t \in d\}\ |} = \log \frac{3}{2}$$

Total no. of documents appear in the corpus (D) is 3. Feature t1 appears in 2 documents and t2 appears in 1 document.
So, we can put these idf values into a vector format. Assume we get idf values of features as 0.65, -0.35, 0.14, 0 and we can show them like vector as below.

$$V\ (d3) = (0.65, -0.35, 0.14, 0)$$

For calculate tf-idf value we need to multiply feature matrix with idf matrix. So, we need to turn idf vector to matrix format. In order to do that, we convert idf vector to a 'square' matrix called as 'diagonal' matrix. This matrix's main diagonal has idf values and rest of the matrix's values are empty. Then we can calculate tf-idf value [19].

3.4.2.3. TF-IDF

So, we already know both tf and idf values in a matrix format. Therefore, the result can be obtained by multiplying each value.

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

Equation 3.3 Equation for Tf-Idf Matrix

This M (tf-idf) matrix stated in above Equation 3.3 is in non-normalize format. Also, this can cause to keyword spamming problem. So, we have to normalize the matrix. To do that, we have to calculate unit vector of each vector. Unit vector we can denote as below Equation 3.4.

$$V = \frac{v1}{||\ v2\ ||\ p}$$

Equation 3.4 Equation for calculating the unit vector

Here V is normalized vector and v1 is the vector that is going to be normalized. V2 is the length of the v1 vector. Let's look at an example which we going to apply this formula M (tf-idf) matrix's 1st row. We apply this in row-wise. Assume we get values 3, 4, 6, 12 as 1st row of M (tf-idf) matrix.

$$V = \frac{v1}{||\ v1\ ||\ p}$$

$$V = \frac{(3, 4,\ 6, 12)}{\sqrt{(3*3) + (4*4) + (6*6) + (12*12)}}$$

By normalizing each row of above matrix, we can get a normalized matrix. So finally, this calculated matrix can be used to map documents in vector space.

3.5. Classifiers

We use four (04) different classifiers in our experimental study. They are Support Vector Machine, Naïve Bayes Classifier, Random Forest Classifier and Logistic Regression model. All these classifiers are trained using the built-in methods and libraries in scikit-learn toolkit (Pedregosa et al., 2011). Further, this is considered as a multi-label classification, which contain 7 labels namely; anger, sad, disgust, happy, fear, surprise according to the Paul Ekam.

If we consider support vector machine, it consists with SVC() and SVM() methods. In our case, we use the SVC() method. SVM() taken in to use when the number of samples are smaller than the number of dimensions.

Naïve Bayes Classifiers are based on Bayes theorem with "naïve" assumption and it is a supervised learning algorithm. Particularly, this is very useful with large data sets. We used Multinomial Naïve Bayes for the experiment [17].

The Random Forest (RF)classifier is a supervised learning algorithm [21], where it addresses bootstrap aggregations commonly known as bagging [22] and feature selection [23] where it builds individual classification or regression trees for prediction.

Logistic Regression model is used with the categorical dependent variables. Since our task is a multi-label classification task we can use multi-nominal logistic model to estimate the probability of a binary response based on one or more or independent variables (features).

3.6. Evaluation

Since the research follows a quantitative approach, what practicing is a systematic investigation this can use statistical, mathematical techniques to accomplish the task.

The model building would be relying on the creation of a function or it model and train it is using the prepared data. The function refers to the algorithm that going to be used and the data would be used to train itself and recognize the patterns. The training should be done more carefully since it must generate a better classifier. The classifier is used to classification mechanism where it receives the input feature vector for classifying it with every tree in the forest and outputs the class label that received the majority of "votes" [26].

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 simplest and most commonly used performance metric. However, other than accuracy, recall, precision and f1-score contribute to evaluating the better model [27].

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 Metrix where it applies the measure of the performance for two class problem in given dataset. Confusion matrix can be defined as follows using the aforementioned instances [27].

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

Accuracy

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Equation 3.5 Calculating Accuracy

According to the above given Equation 3.5, the accuracy can be defined as the fraction of correct predictions. Accuracy is not a good measure of success when the data is unbalanced. Since accuracy gives the result considering the correctly classified data(class) even the other classes are misclassified.

Recall

$$Recall = TP / (TP + FN)$$

Equation 3.6 Calculating Recall

As shown in the above Equation 3.6, the recall is the ratio of TP to the (TP + FN) which is the number of entire faulty modules.

Precision

$$Precision = TP / (TP + FP)$$

Equation 3.7 Calculating Precision

The Equation 3.7 display the equation to calculate the precision. Precision is the ratio of TP to the (TP + FP) as known as an entire module of fault prone.

This does not consider about negatives. Therefore, this measure is used to measure the correctness of the positive predictions.

F1-Score

F-score = (2 x precision x recall) / (precision + recall)

Equation 3.8 Calculating the F-Score

As stated in Equation 3.8, the harmonic mean of precision and recall is calculated in F1-score. In our study, F1-score is considered as the main evaluation measure. Since, the F1-score is does not rely on either precision or recall.

Chapter 4 - Implementation

This chapter provides a detailed description from the beginning to end about the implementation approaches, methods and technologies used in the experiment of finding the best model for text classification using the sentiment analysis techniques. The main steps of the experimental process are as follows.

- Data collection and annotation
- Data preprocessing
- Feature extraction
- Training the model and evaluation

Since the first step, data collection and annotation, was discussed in the Methodology chapter, this chapter will continue the discussion from second step onwards.

4.1. Data Pre-processing

Steps used for preprocessing the corpus are presented in this section. Tools and packages used for the experiment will be presented. After considering different toolkits for Natural Language Processing decided to use Natural Language Processing Toolkit, NLTK library with python for the preprocessing task. NLTK can perform sentence splitting, tokenization, pos-tagging, lemmatization and many other preprocessing activities. Since, the language we used for all preprocessing and other experimental work is python; the preprocessed data are stored in the memory and this can be directly accessed for the next steps of the experiment.

The data set (the set of collected social media comments), containing both training and testing data are stored in a single excel file. Before the data set is subjected to the pre-processing the data set (comments) stored in the excel file are read into a data frame in python. Then all the pre-processing activities and other feature extraction and training activities were relying on this data frame.

As the first task the comments stored in the data frame are subjected to pre-processing and saved the pre-processed data to the same data frame.

Remove HTML tags

The *BeautifulSoup* python library used for this. This removes any available HTML or XML content from the data set.

```
for i in range(len(inputDf)):
    currentPhase= inputDf['Comments'].values[i]
    inputDf['Comments'].values[i] = BeautifulSoup(currentPhase,"html.parser").get_text()
```

Remove Null values

Used the NLTK built-in function, namely *noynull()* for this. This removes any available null values from the data set.

```
inputDf = inputDf[inputDf['Comments'].notnull()]
```

Removing Special Characters and Stop words

Punctuations and stop words in English were removed in this step. An array including all the non-alpha numeric characters used for the removal of special characters.

```
for i in range(len(inputDf)):
    currentValue = inputDf['Comments'].values[i]
    for ch in nonAlpha:
        if ch in currentValue:
            currentValue = currentValue.replace(ch,' ')
    inputDf['Comments'].values[i] = currentValue
```

The raw text	WOW… Good Job! Better than the year 2002.
The text after removing special characters	WOW Good Job Better than the year 2002

The stop word dictionary in NLTK is used for the stop word removal. Each comment in the data set is checked with the stop word dictionary and remove all the available stop words.

```
stop = stopwords.words('english')
   inputDf['Comments'] = inputDf['Comments'].apply(lambda x: " ".join(w for w in x.split() if not w.lower() in stop))
```

The text after removing special characters	WOW Good Job Better than the year 2002
The text after removing stop words	WOW Good Job Better year 2002

Convert all the text into Lowercase

Used the NLTK built-in function, namely *lower()* for this. All the letters in the data set are converted in to lowercase.

```
inputDf['Comments'] = inputDf['Comments'].apply(lambda x: " ".join(x.lower() for x in x.split()))
```

The text after removing stop words	WOW Good Job Better year 2002
The text after converting to lowercase	wow good job better year 2002

Tokenization

Splitting the sentences into words and punctuations is known as tokenization. An in-built function in NLTK toolkit, namely *word_tokenize( )* function, is used for the sentence tokenization.
Every comment in the data set is tokenized using this function.

```
for i in range(len(inputDf)):
    currentPhase= inputDf['Comments'].values[i]
    tokenizedList = []
    for curr in currentPhase.split():
      tokenizedList.append(curr)
      inputDf['Comments'].values[i] = tokenizedList
```

The text after converting to lowercase	wow good job better year 2002
Tokenized text	[wow, good, job, better, year, 2002]

Lemmatizing

The process of identifying the root or stem of a word is known as lemmatization. For lemmatization we used WordNet Lemmatizer. Before lemmatization POS tagging was used and words were tagged using WordNet. POS tagging was done since lemmatization was done base on the POS tags. After lowercasing all words, we simply used *nltk.pos_tag()* directly since this function loads the pre-trained tagger from a file where it was trained with Treebank corpus. Then using the POS tags, the Treebank tags are mapped to the WordNet POS names. For this tagging process we have only considered adjectives, verbs, nouns and adverbs of Treebank POS tagging. This was done since verbs, nouns, adjectives and adverbs play a major role in text analysis when compared with other types of words. Finally, the POS tag we obtained from WordNet was passed with the particular word for lemmatization. Following example will explain the process clearly.

```
word = going
wnl = WordNetLemmatizer( )
postag = nltk.pos_tag(word)
```

For the given word “going” this will return VBP as the POS tag. Since this is a verb then we map the POS tag from WordNet corresponding the tag verb.

```
    if pos_tag.startswith('V'):
        return wn.VERB
```

Then this returned POS tag will be passed for lemmatization with the corresponding word.

```
    wnl.lemmatize('going', wn.VERB)
```

As the output, we will get the word "go" as the lemmatized word.

```
lemmatizer = WordNetLemmatizer()
  for i in range(len(inputDf)):
    inputDf['Comments'].values[i] = [lemmatizer.lemmatize(word,pos='v') for word in
inputDf['Comments'].values[i]]
```

The text after converting to lowercase	wow good job better year 2002
Lemmatized text	['wow', 'good', 'job', 'better', 'year', '2002']

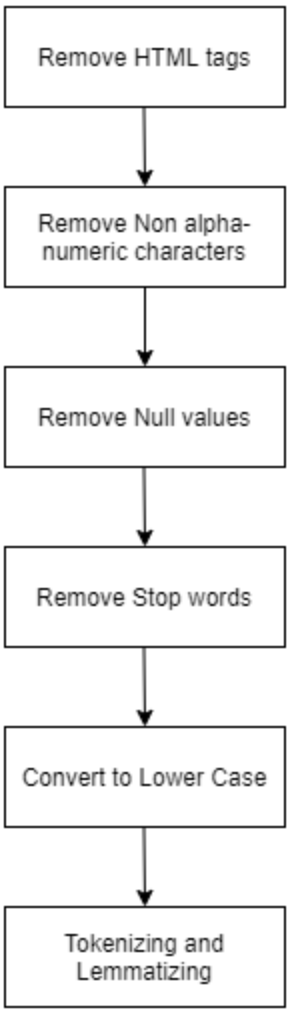

Figure 4.1 Steps of Data Pre-processing

After the preprocessing steps, the preprocessed data set was split into two as training and testing data by using the method available in scikit-learn, namely t*rain_test_split().* The splitting will be done by shuffling the dataset [28].

4.2. Feature Extraction

The next step of the experiment is feature extraction. Packages, libraries and methods in scikit learn toolkit is used for below-mentioned all feature extraction activities.

In this experiment, we have used Countvectorizer and Tfidf vectorizer, two of the widely using in-built vectorizers available in scikit learn toolkit. And in order to find the best performing model with different features, we have used both vectorizers with different ngram features.

4.2.1. Count Vectorizer – Bag of Word Features (BoW)

The *CountVectorizer()* allows to convert a text data set into a vector of token counts. This representation is done by converting the text into fixed length vectors relying on the frequency of each word.

```
vectorizer = CountVectorizer(ngram_range=(1,1), min_df=3)
```

The features are extracted relying on both uni-gram and bi-gram features of CountVectorizer().

```
vectorizer = CountVectorizer(ngram_range=(1,2), min_df=3)
```

4.2.2. Tf-idf Vectorizer – Term Frequency Features (TF-IDF)

The *TfidfVectorizer()* allows to convert a text data set into a feature index in the matrix. In this vectorization method, each unique token gets a feature index. Basically, this is used to measure the importance of a word and to measure the occurrences of a word. Then, the words with higher occurrences are identified as features.

```
vectorizer = TfidfVectorizer(ngram_range=(1,1), min_df=3)
```

The features are extracted relying on both uni-gram and bi-gram features of TfidfVectorizer().

```
vectorizer = TfidfVectorizer(ngram_range=(1,2), min_df=3)
```

4.3. Classification models and Evaluation

As mentioned in the Design chapter we have selected four (4) machine learning algorithms to build classifier models. All those models are trained using the features extracted in above Feature Extraction section.

classifier_lr = LogisticRegression()

classifier_nb = MultinomialNB()

classifier_svm = SVC()

classifier_rf = RandomForestClassifier()

4.3.1. Training the Logistic Regression Model

LogisticRegression() method in scikit learn toolkit is used for this. First build the logistic regression model using the above *LogisticRegression()* method.

classifier_lr = LogisticRegression()

After build the model, the next step was model prediction. *predict()* function offered in scikit learn toolkit is used for this.

Once the prediction is completed, we need to evaluate the performance by using the confusion matrix and using the other performance measurements such as; Accuracy, F1-Score, Re-call and Precision.

4.3.2. Training the Naïve Bayes Model

MultinomialNB() method in scikit learn toolkit is used for this. First build the logistic regression model using the above *MultinomialNB()* method.

```
classifier_nb = MultinomialNB()
```

After build the model, the next step was model prediction. *predict()* function offered in scikit learn toolkit is used for this.

Once the prediction is completed, we need to evaluate the performance by using the confusion matrix and using the other performance measurements such as; Accuracy, F1-Score, Re-call and Precision.

4.3.3. Training the Support Vector Machine Model

SVC() method in scikit learn toolkit is used for this. First build the logistic regression model using the above *SVC()* method.

```
classifier_svm = SVC()
```

After build the model, the next step was model prediction. *predict()* function offered in scikit learn toolkit is used for this.

Once the prediction is completed, we need to evaluate the performance by using the confusion matrix and using the other performance measurements such as; Accuracy, F1-Score, Re-call and Precision.

4.3.4. Training the Random Forest Classifier Model

RandomForestClassifier() method in scikit learn toolkit is used for this. First build the logistic regression model using the above *RandomForestClassifier()* method.

```
classifier_rf = RandomForestClassifier()
```

After build the model, the next step was model prediction. *predict()* function offered in scikit learn toolkit is used for this.

Once the prediction is completed, we need to evaluate the performance by using the confusion matrix and using the other performance measurements such as; Accuracy, F1-Score, Re-call and Precision.

Chapter 5 - Results and Evaluation

In this chapter we will discuss about the experiments carried out to find the best approach and the results of the experiments, following the experimental setups described in Chapter 4. Our main objective is to perform emotion categorization of the social media comments. In this study, different feature sets are trained with different classifiers comparing to accuracy, precision, recall and F-score measures in order to find the best model out from the selected models.

As mentioned under the Implementation chapter, all the comments were stored in an excel file and read into a data frame. The read data are pre-processed and then split into two with the 8:2 ratio for training and testing.

Training Data Set	5600 comments
Testing Data Set	1400 comments

Table 5.1 Training and Testing data

BoW Features

The countvectorizer() in Scikit-learn package is used to extract the bag of words (BoW) features. Then the extracted feature vector passed through four (04) different models and the training and testing performance of each model is evaluated.

	Accuracy	Precision	Recall	F1-Score
SVM	0.7	0.71	0.7	0.7
LRM	**0.71**	**0.71**	**0.71**	**0.71**
NBM	**0.68**	**0.68**	**0.68**	**0.68**
RFM	0.7	0.71	0.7	0.7

Table 5.2 Results with BoW Features

	SVM	LRM	NBM	RFM
Train Accuracy	0.99	**0.92**	**0.78**	0.99
Test Accuracy	0.7	**0.71**	**0.68**	0.7

Table 5.3 Comparing Train and Test Accuracy of the models with BoW features

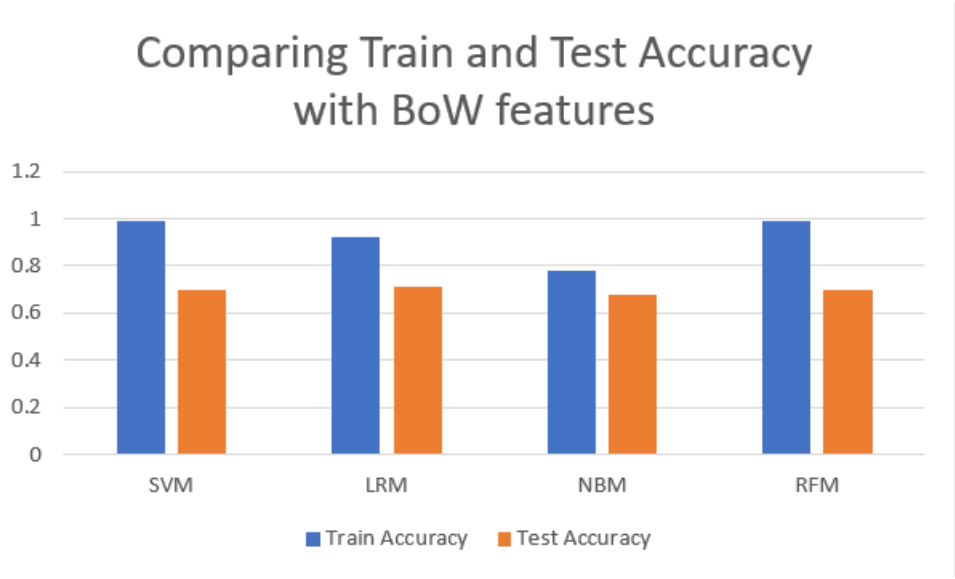

Figure 5.1 Comparing Train and Test Accuracy of the models with BoW features

According to the Table 5.2, the Logistic Regression Model has an accuracy of 0.71 with an F-score of 0.71. But, according to Table 5.3 the Naïve Bayes model has the least gap between train and test accuracies. Also, the train accuracies of SVM and RFM are nearly 1.0. That means, those models are overfit. Therefore, Naïve Bayes model has the best performance with BoW feature set.

The confusion matrix of predicted results of Naïve Bayes Model is as follows.

		Predicted Class						
True Class		0	1	2	3	4	5	6
	0	128	14	15	8	26	5	13
	1	6	154	12	11	13	5	5
	2	12	14	129	0	5	2	5
	3	10	10	6	160	12	2	0
	4	24	12	9	9	126	0	0
	5	8	3	6	5	10	120	37
	6	21	16	18	8	15	20	131

Table 5.4 Confusion Matrix of Naïve Bayes Model with BoW features

Bi-gram Features of BoW

The countvectorizer() in Scikit-learn package with ngram_range(1,2) parameter is used to extract the bi-gram features of BoW. Then the extracted feature vector passed through four (04) different models and the training and testing performance of each model is evaluated.

	Accuracy	Precision	Recall	F1-Score
SVM	0.7	0.7	0.7	0.7
LRM	**0.73**	**0.73**	**0.73**	**0.73**
NBM	**0.69**	**0.7**	**0.69**	**0.69**
RFM	0.72	0.72	0.72	0.72

Table 5.5 Results with Bi-gram Features of BoW

	SVM	LRM	NBM	RFM
Train Accuracy	0.99	**0.95**	**0.81**	0.99
Test Accuracy	0.7	**0.73**	**0.69**	0.72

Table 5.6 Comparing Train and Test Accuracy with Bi-gram features of BoW

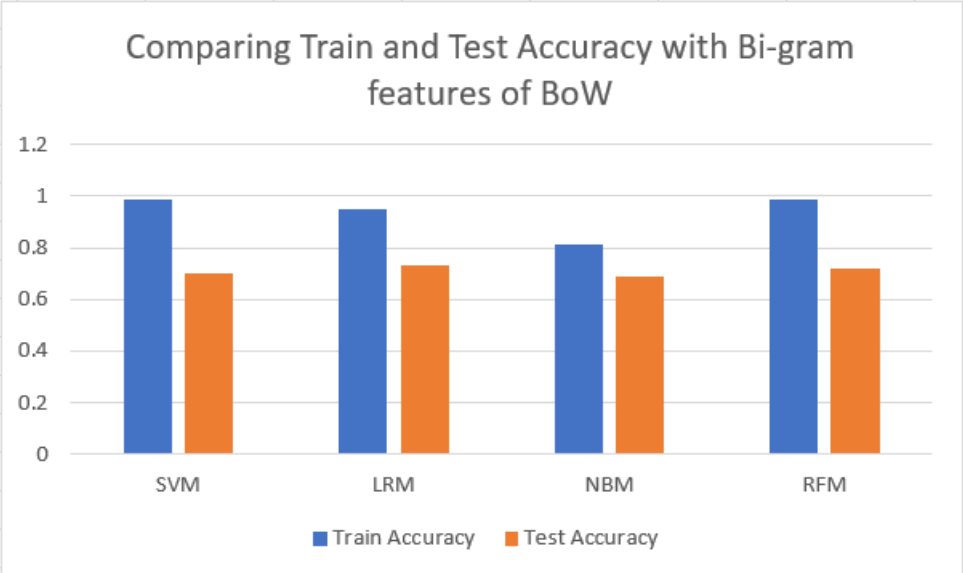

Figure 5.2 Comparing Train and Test Accuracy with Bi-gram features of BoW

According to the Table 5.5, the Logistic Regression Model has an accuracy of 0.73 with an F-score of 0.73. But, according to Table 5.6 the Naïve Bayes model has the least gap between train and test accuracies. Also, the train accuracies of SVM and RFM are nearly 1.0. That means, those

models are overfit. Therefore, Naïve Bayes model has the best performance with bi-gram features of BoW.

The confusion matrix of predicted results of Naïve Bayes Model is as follows.

		Predicted Class						
True Class		0	1	2	3	4	5	6
	0	136	11	12	9	25	5	11
	1	10	151	8	12	14	5	6
	2	16	14	125	19	6	2	5
	3	11	8	6	164	9	2	5
	4	21	10	4	9	135	1	0
	5	9	2	5	6	9	119	39
	6	22	11	15	8	15	20	138

Table 5.7 Confusion Matrix of Naïve Bayes Model with Bi-gram features of BoW

Tf-Idf Features

The Tfidfvectorizer() in Scikit-learn package is used to extract the Tf-Idf features. Then the extracted feature vector passed through four (04) different models and the training and testing performance of each model is evaluated.

	Accuracy	Precision	Recall	F-Score
SVM	**0.73**	**0.73**	**0.73**	**0.73**
LRM	0.71	0.71	0.71	0.71
NBM	**0.68**	**0.68**	**0.68**	**0.67**
RFM	0.71	0.71	0.71	0.71

Table 5.8 Results of Tf-idf Features

	SVM	LRM	NBM	RFM
Train Accuracy	**0.99**	0.87	**0.81**	0.99
Test Accuracy	**0.73**	0.71	**0.68**	0.71

Table 5.9 Comparing Train and Test Accuracy with Tf-Idf Features

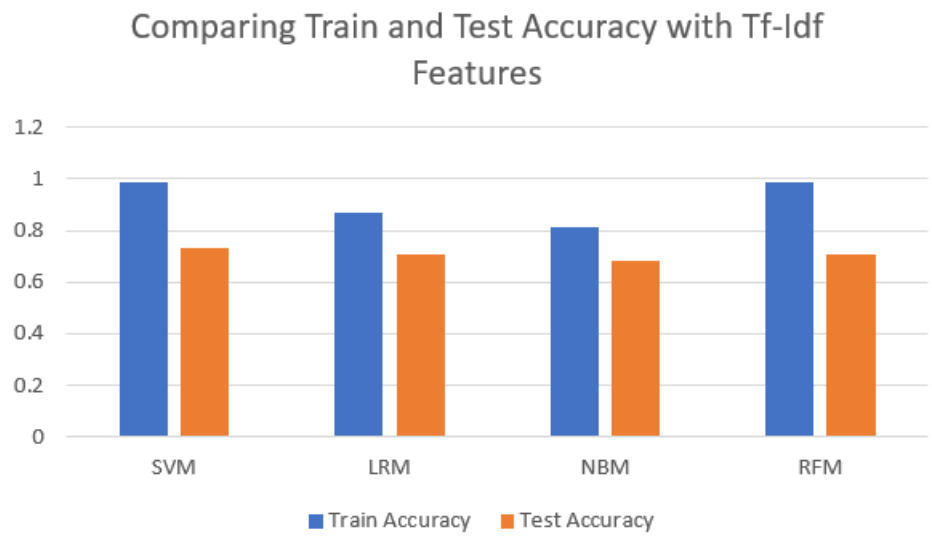

Figure 5.3 Comparing Train and Test Accuracy with Tf-Idf Features

According to the Table 5.8, the SVM has an accuracy of 0.73 with an F-score of 0.73. But, according to Table 5.9 the Naïve Bayes model has the least gap between train and test accuracies. Also, the train accuracies of SVM and RFM are nearly 1.0. That means, those models are overfit. Therefore, Naïve Bayes model has the best performance with Tf-Idf features.

The confusion matrix of predicted results of Naïve Bayes Model is as follows.

		Predicted Class						
True Class		0	1	2	3	4	5	6
	0	119	16	17	8	33	5	11
	1	6	160	12	9	13	3	3
	2	13	17	123	19	9	2	4
	3	9	13	4	160	11	2	1
	4	16	15	5	8	135	1	0
	5	6	3	7	5	9	121	38
	6	17	13	17	7	14	31	130

Table 5.10 Confusion Matrix of Naïve Bayes Model with TF-IDF

Bi-gram Features of Tf-Idf

The Tfidfvectorizer() in Scikit-learn package with ngram_range(1,2) parameter is used to extract the bi-gram features of Tf-Idf. Then the extracted feature vector passed through four (04) different models and the training and testing performance of each model is evaluated.

	Accuracy	Precision	Recall	F1-Score
SVM	**0.74**	**0.74**	**0.74**	**0.74**
LRM	0.73	0.73	0.73	0.73
NBM	**0.7**	**0.71**	**0.7**	**0.7**
RFM	0.71	0.72	0.71	0.71

Table 5.11 Results of Bi-gram Features of Tf-Idf

	SVM	LRM	NBM	RFM
Train Accuracy	**0.99**	0.91	**0.86**	0.99
Test Accuracy	**0.74**	0.73	**0.7**	0.71

Table 5.12 Comparing Train and Test Accuracy of Bi-gram Features of Tf-Idf

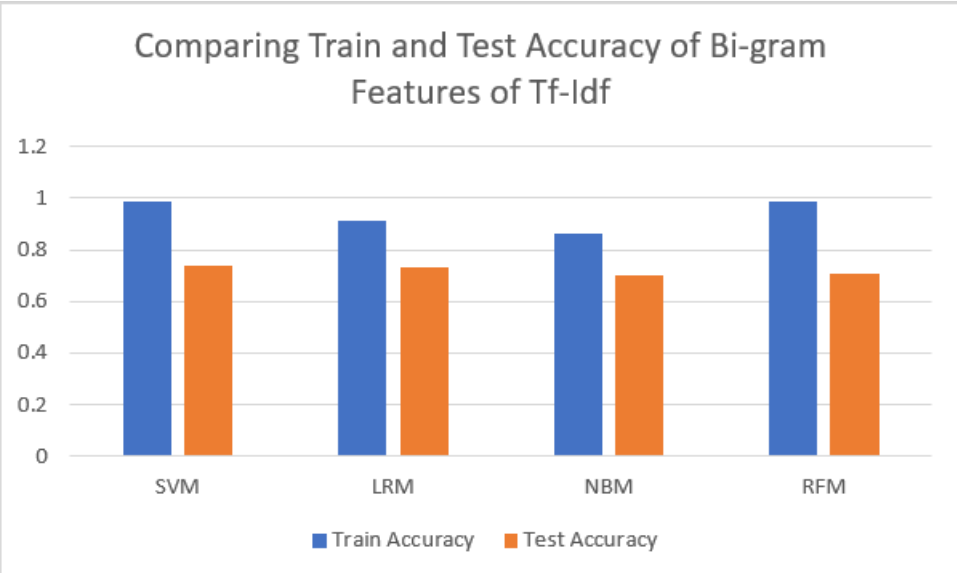

Figure 5.4 Comparing Train and Test Accuracy of Bi-gram Features of Tf-Idf

According to the Table 5.11, the SVM has an accuracy of 0.74 with an F-score of 0.74. But, according to Table 5.12 the Naïve Bayes model has the least gap between train and test accuracies.

Also, the train accuracies of SVM and RFM are nearly 1.0. That means, those models are overfit. Therefore, Naïve Bayes model has the best performance with bi-gram features of Tf-Idf.

The confusion matrix of predicted results of Naïve Bayes Model is as follows.

		Predicted Class						
True Class		0	1	2	3	4	5	6
	0	123	14	15	7	35	6	9
	1	5	158	12	9	12	5	5
	2	14	16	129	14	9	2	3
	3	9	10	4	160	14	2	1
	4	14	9	5	6	145	1	0
	5	5	2	2	6	10	124	40
	6	16	8	9	9	12	31	144

Table 5.13 Confusion Matrix of Naïve Bayes Model with Bi-gram features of Tf-Idf

Comparison of Different Features

Here, we list down all the identified best performing models with each feature type and select the best model out of them.

	Accuracy	Precision	Recall	F-Score
NBM-BoW	**0.68**	**0.68**	**0.68**	**0.68**
NBM-Bigram (BoW)	0.69	0.7	0.69	0.69
NBM-TfIdf	0.68	0.68	0.68	0.67
NBM-Bigram (TfIdf)	0.7	0.71	0.7	0.7

Table 5.14 Comparison of Models

	NBM-BoW	NBM-Bigram (BoW)	NBM-TfIdf	NBM-Bigram (TfIdf)
Train Accuracy	**0.78**	0.81	0.81	0.86
Test Accuracy	**0.68**	0.69	0.68	0.7

Table 5.15 Comparison of Train and Test Accuracy

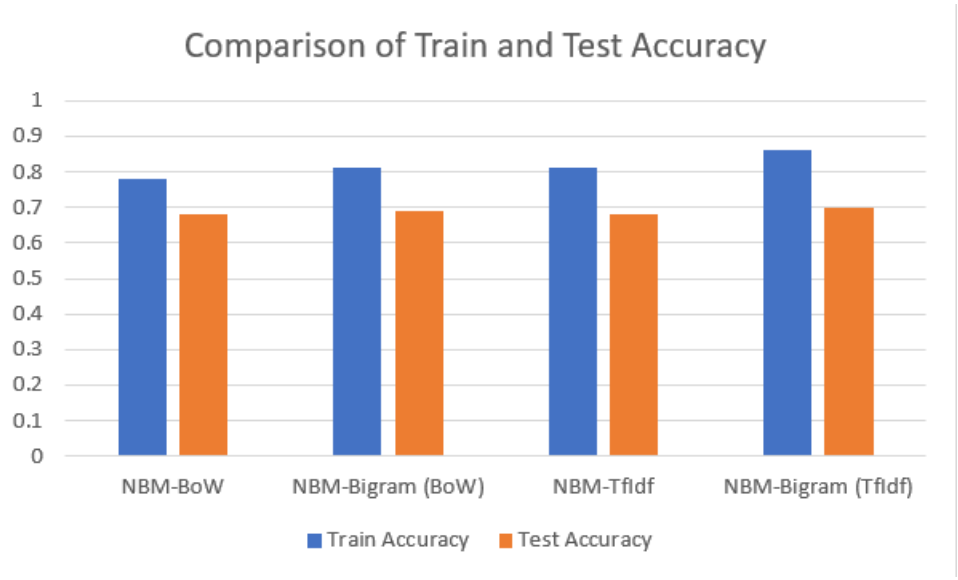

Figure 5.5 Comparison of Train and Test Accuracy

Out of the selected best performing models with each feature, all of the models have the accuracies which are approximately equal to each other. Among them, the model with the least gap between train and test accuracies is selected as the best performing model. Then, the Naïve Bayes model with BoW features is selected as the best performing model which has a 0.1 gap between the accuracies. This has an accuracy of 0.68 and F1-score of 0.68.

Applying the research outcome in order to gain profit for business organizations

According to the results obtained above, the Naïve Bayes model with BoW features is the best performing model with an accuracy of 0.68 and F1-score of 0.68.

This model can be used to implement the tools used by organizations in order to predict the successfulness of each post. Beyond providing a regular successfulness of the social media posts, this can be used to label each social media comment as an analysis of main human emotions (happy, sad, angry, disgust, fear, neutral and surprise) in order to provide a comprehensive visibility of post successfulness.

For an example, if a comment is identified as it contains the “happy” or “surprise” emotions then it will define that the customer has a positive attitude towards the post.

Likewise, this model can be used to analyze all the comments posted for the social media post and can have an overall feedback for the respective post.

As an example, if a post got 20 comments and 10 comments out of them express their gratitude towards the organization and 4 out of 20 are just neutral comments while the others disagree with the post what ever it contains; then the overall idea of the comment can be identified as “Fair” or “Positive” as the majority is agreed with the content or with the organization.

Likewise, this will be helpful to predict the successfulness of a selected post or posts quickly and it will help the organizations to take necessary actions to increase and maintain their reputation.

Chapter 6 – Conclusion and Future Work

This chapter includes the final comments, findings and future works to extend this study. In this study we find out and present the most efficient approach for text analysis using sentiment analysis techniques. The comprehensive literature review, survey of existing methods and algorithms to find out the efficient and more accurate approach for text classification and the application of the findings to evaluate and identify the best approach are the main three (03) steps in this study.

6.1. Conclusion

This social media phenomenon has turned out to be humans both personal and professional engagements with several sectors where an entire society has drawn to capture the advantages through each part of it. Due to this growth, more and more businesses would shift their marketing to social media platforms which lead to sharing 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 feedback 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.

In this document, we proposed sentiment analysis and machine learning techniques to classify the emotions using social media comments. The main objective is to analyze a large labeled dataset to find out the best suite approach for the classification. When comparing test accuracies, the SVM model with Tf-Idf parameters gives a better result than the other models used in the study. In that case, the Bi-gram features of the Tf-Idf parameters perform better than using the unigram features. But the Naïve Bayes model with BoW parameters is identified as the best performer among the other models, when comparing the both training and testing accuracies.

6.2. Future work

In the future, the plan is to refine and further improve the techniques in order to achieve an enhanced accuracy. This can be achieved by using the in-depth training techniques such as cross validations and tuning model parameters.

Moreover, this can be extended to get real-time output by analyzing the comments by integrating the solution with any social media platform. And the language is to be extended beyond English.

Further, this research can be expanded to predict the successfulness of each individual post by itself without depending on the external applications to do the success prediction and it will help to provide a comprehensive visibility of post successfulness.

References

[1]. Carver, B., "*5 Reasons to Invest in Dedicated Social Media Analytics Tools"*. [online]. Available: https://www.socialmediatoday.com/social-business/bobcarvertc/2015-10-09/5-reasons-invest-dedicated-social-media-analytics-tools

[2] Kim, C. and Yang, S., "Like, comment, and share on Facebook: How each behavior differs from the other". *Public Relations Review*, 43(2), pp.441-449.

[3] Pletikosa, Irena & Michahelles, Florian., "Monitoring Trends on Facebook". Proceedings - IEEE 9th International Conference on Dependable, Autonomic and Secure Computing, DASC 2011. 895-902. 10.1109/DASC.2011.150.

[4] Fernando, Noel., "Knowledge Based Approach for Concept Level Sentiment Analysis for Online Reviews". International Journal of Emerging Trends & Technology in Computer Science.

[5] Dreamgrow.com., [online] Available: https://www.dreamgrow.com/top-15-most-popular-social-networking-sites

[6] L. H. Nguyen, A. Salopek, L. Zhao and F. Jin., "A natural language normalization approach to enhance social media text reasoning," (*2017) IEEE International Conference on Big Data (Big Data)*, Boston, MA, 2017, pp. 2019-2026. doi: 10.1109/BigData.2017.8258148

[7] Aparna U.R. and S. Paul., "Feature selection and extraction in data mining," (*2016) Online International Conference on Green Engineering and Technologies (IC-GET)*, Coimbatore, 2016, pp. 1-3. doi: 10.1109/GET.2016.7916845

[8] Kensinger, E. (2004). *Remembering Emotional Experiences: The Contribution of Valence and Arousal*. [online] Pdfs.semanticscholar.org. Available: https://pdfs.semanticscholar.org/55ee/7d2ffa600eab05c8ce178a5531c4a8ec9439.pdf

[9] Ekman, P., 2018. [online] Paulekman.com. Available: http://www.paulekman.com/wp-content/uploads/2013/07/An-Argument-For-Basic-Emotions.pdf

[10] Bravo, F., Frank, E. and Pfahringer, B., [online] Cs.waikato.ac.nz. Available: https://www.cs.waikato.ac.nz/~eibe/pubs/ijcai15.pdf

[11] Kumar, A. and Sebastian, T., [online] Ijcsi.org. Available: https://www.ijcsi.org/papers/IJCSI-9-4-3-372-378.pdf

[12] E. Garcia., "An Introduction to Local Weight Models - Minerazzi". [Online]. Available: http://www.minerazzi.com/tutorials/term-vector-4.pdf

[13] SHANKARARAMAN, Venky; GOTTIPATI, Swapna; LIN, Jeff Rongsheng; and GAN, Sandy., "Extracting implicit suggestions from students' comments – A text analytics approach". *Proceedings of 25th International Conference on Computers in Education (ICCE 2017:, Christchurch, New Zealand, December 4-8*. 261-269. Research Collection School Of Information Systems

[14] C. Brun and Hag`egeC., "Suggestion Mining: Detecting Suggestions for Improvement in Users ...", [Online]. Available: https://pdfs.semanticscholar.org/d927/4cfd97362f221c5e8d0f8c07b4041a6b3898.pdf.

[15] S. Negi and P. Buitelaar, "Towards the Extraction of Customer-to-Customer Suggestions from ...", [Online]. Available: http://www.emnlp2015.org/proceedings/EMNLP/pdf/EMNLP258.pdf.

[16] S. Gottipati and J. Jiang, "Extracting and Normalizing Entity-Actions from Users", [Online]. Available: https://www.aclweb.org/anthology/C12-2042.

[17] Y. WANG, “Data Preparation for Social Network Mining and Analysis”, [Online]. Available: https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=1100&context=etd_coll.

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

[19] ink.library.smu.edu.sg, 2018. [Online]. Available: http://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=4835&context=sis_research.

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

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

[23] 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.

[24] 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.

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

[26] "Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis", *Irdindia.in*, 2018. [Online].
Available: http://www.irdindia.in/journal_ijaece/pdf/vol3_iss4/2.pdf.

[27] 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.

[28] Pedregosa, F., et al.,“Scikit-learn: Machine learning in Python. Journal of Machine Learning Research” 12:2825–2830, 2011.