TEXT-BASED EMOTION DETECTION

LEE BOON PING

SESSION 2017/2018

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY
MARCH 2018

TEXT-BASED EMOTION DETECTION

BY

LEE BOON PING

SESSION
2017/2018

THIS PROJECT REPORT IS
PREPARED FOR

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY
IN PARTIAL FULFILLMENT
FOR

BACHELOR OF COMPUTER SCIENCE
B.CS (HONS) DATA SCIENCE

FACULTY OF COMPUTING AND INFORMATICS

MULTIMEDIA UNIVERSITY

MARCH 2018

DECLARATION

I hereby declare that the work has been done by myself and no portion of the work contained in this thesis has been submitted in support of any application for any other degree or qualification of this or any other university or institute of learning.

Name of candidate: Lee Boon Ping

Student ID: 1142701239

Faculty of Computing & Informatics

Multimedia University

Date: 14 March 2018

Acknowledgement

I would like to take this opportunity to express my profound gratitude and deep regards to all who have helped me in this project. I am grateful to my supervisor, Dr. Soon Lay Ki, who is encouraging and supporting me to finish my report write-up. The support and guidance given by her time to time had to carry me a long way in this project which I am about to engage.

In addition, I very thankful to my friends Darryl Lim, Wayne Ng and Tazeek for their support. They really help much in this project on information and guidelines. A very thanks to Chun Wei and Zi Xiang for helping me to label the emotion based on the sentences.

Last but not least, I would like to acknowledge with much appreciation to my family members and friends for their support and encouragement. Without their support, it will be hard for me in this project.

Abstract

This project is focused on the emotion detection from tweets. The construction of dataset annotated for 13 emotions which are enthusiasm, fun, happiness, love, anger, boredom, hate, relief, sadness, worry, empty, neutral and surprise.

Several machine learning techniques have been used to build a classification model for emotion detection to test the real data which are streaming from Twitter API. This project demonstrates the three main supervised algorithms which are Support Vector Machine (SVM), Random Forest and Naïve Bayes. Besides that, training dataset goes through a series of data preprocessing before fitting into the model. New test datasets are generated when the dataset is preprocessed. Every different data preprocessing has generated different feature of datasets by using the same dataset.

Train-test split technique is applied to the dataset before fitting to classification model. Other than that, another technique which is cross-validation is performed to make the classification model fit the training data better. Due to the imbalance dataset, the classification model is analyzed through Area Under the Receiver Operating Characteristic Curve (ROC AUC) to identify the true positive rate (TPR) against false positive rate (FPR). From the experiment, Naïve Bayes Classifier has produced the best result in terms of ROC AUC with minimal of data pre-processing steps.

Table of Contents

List of Figures

List of Tables

Chapter 1 : Introduction

1.1 Overview

Emotion is intuitive or instinctive feeling as distinguished from reasoning or knowledge. Emotion is common and important to all aspect of human lives. It will affect human decision-making, form our daily behaviour, influence our social relationships, even outlasts human memories (Wang W. e., 2012). The emotion detection from the text is widely studied as it can be used for affective computing, opinion mining and others in computer linguistics (Strapparava, 2008).

With the rapid growth of microblog post in Twitter, there is an opportunity to create an automatic tool for identifying the emotions of Twitter users expressed in text (Wang W. e., 2012). Most of Twitter users share their opinion on different of topics, write about their life and discuss the trendy issues. The users of Twitter tend to shift to microblogging services as the services provided informality of posts and accessibility of the microblogging site (Pak A. &., 2010).

The emotions in a tweet are very important for the product-based company and service-based company. These companies can understand the feeling from the audience in order to improve the services or products. Other than that, some of the companies such as Intel, Twitter and IBM are using analytics to evaluate the satisfaction of employee toward the company. In this case, these companies have some program to help enhance the likelihood employees stay on the job (Brown, 2015).

However, identifying the expressed emotion in tweets is a difficult task for these studies. In September 2017, Twitter has increased the length of the tweet from 140 characters to 280 characters (Newton, Casey, 2017). But still, insufficient of context and free format of text make the emotion detection a far harder task.

1.2 Problem Statement

Nowadays, people tend to share their thinking or views in microblogging services such as Twitter. Unconsciously, people are sharing their emotion through the expressed text.

Meanwhile, it is a great chance to build a system which can analyze the emotion of a tweet. This system can be applied into different purposes such as suicide prevention, election, product review and others.

There are several pieces of research that have been explored this sector. They have used a lot of different type of techniques such as machine learning technique and lexicon-based approach to detect the emotion from the text. However, the comparison of the different type of machine learning algorithm towards the different type of pre-processed of datasets is not yet explored.

1.3 Proposed Solution

Machine learning techniques are classified into two categories, which is supervised and unsupervised. In this project, the supervised technique is chosen to implement in emotion detection. The machine learning algorithms used in this project are Support Vector Machine, Naïve Bayes and Random Forest.

There are seven types of data pre-processing steps that are needed to go through before data transformation. After each data pre-processing step, the dataset will be saved and used for model fitting. The eight datasets including the original dataset will be used for comparing the performance of data pre-processing.

After that, Term Frequency-Inverse Document Frequency (TF-IDF) is applied to transform the textual representation of information into a matrix of TF-IDF features. Then, train/test split and cross-validation are implemented before fitting into classification model.

1.4 Objectives

Objectives of this project are as follow:

1. To identify the appropriate data pre-processing steps for emotion detection from the short text.
2. To transform and extract relevant features for effective emotion detection from short text using selected classification algorithms.
3. To examine the performance of selected classification algorithms for emotion detection using annotated dataset.
4. To evaluate the effectiveness of built models on newly crawled data from Twitter.

1.5 Scope

The training dataset of this project is obtained from CrowdFlower. The dataset is a corpus of tweets with annotated of emotion. This training dataset contains 13 different types of emotion. This project will be coded in Python language as it provides a wide range of text processing packages.

Firstly, the corpus of tweets has proceeded to data cleaning. The irrelevant and redundant data is removed from the dataset. Next, the pre-processed data is converted from informative text into numerical features which can be supported by machine learning.

Then, the dataset will be split into train set and test set before fitting into the model. Other than train/test split, another way which is cross-validation is also implemented in model fitting.

After modelling, the performance of the model will be evaluated in order to find the best model. The best performed of the model will be taken as the predictive model and will be applied to the test dataset which is streamed from Twitter.

1.6 Report Organizations

The remainder of the project is sorted as follows. Chapter 2 presents the relevant study with emotion models, dimensionality reduction and approaches used to detect the emotion from the text. While Chapter 3 describes the method that has been taken in this project. Chapter 4 presents the discussion of the result. Chapter 5 concludes the work of the project.

Chapter 2 : Literature Review

2.1 Emotion Models

Emotion models are the base that presented the emotions in a different way. There is a number of approaches about how to represent the emotion according to studies in psychology. Figure 2.1 presents the techniques to implement the categorical and dimensional models (Calvo, 2013).

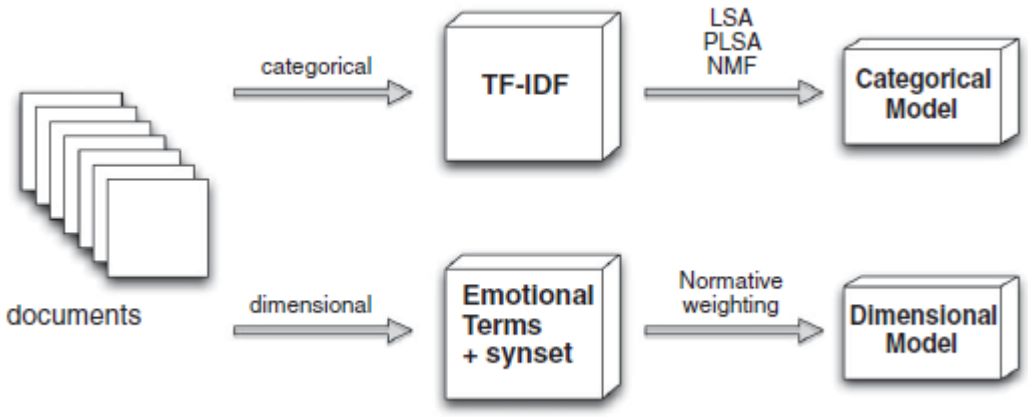


Figure 2.1: The implementation of techniques in categorical and dimensional models (Calvo, 2013)

2.1.1 Categorical Model

Categorical model is widely used in emotion model (Calvo, 2013). The categorical model is the classification task that converts the text into predefined categories based on a thesaurus. A hypothesis is made that human using the same language have identical understanding for distinct discrete emotion based on the thesaurus in this model (Calvo, 2013). This model is focused on unique emotion labels or classes. In this model, Canales assumed that there are discrete emotion categories (Canales, 2014). The six basic emotions which are sadness, anger, joy, disgust, surprise and fear are found in Ekman's basic emotion model which also included in the categorical model (Canales, 2014). Roberts added the love emotion to Ekman's basic emotion model due to informal text in Twitter which is commonly found (Roberts, 2012). Figure 2.2 shows the emotion ontology for the six Ekman's emotion including one more emotion, which is love. Solid lines indicate inheritance, dashed lines indicate the opposite.

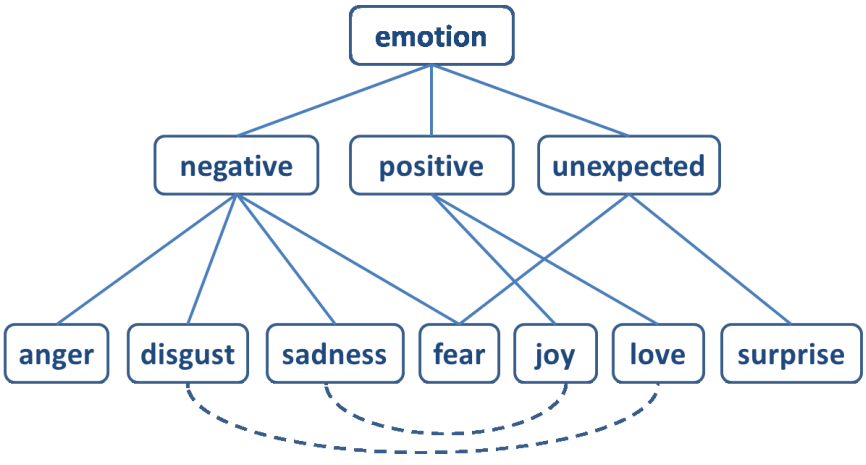


Figure 2.2: The emotion ontology for the six Ekman emotions (plus LOVE) by (Roberts, 2012)

2.1.2 Dimensional Model

Another model is the dimensional model. In emotional experience, this model indicates the significance of the fundamental dimensions of arousal and valence. The emotion theorists studied dimensional model and evidence was provided that the existence of more than two fundamental dimensions of emotional experience which were arousal and valence (Russell, 2003). The arousal and valence were general primitives in emotional experience. The arousal and valence named the feeling at any point on this two-dimensional space core affect. Dimensional model is better used as the method to visualize the emotions in a psychologically meaningful space compare to feature space (Calvo, 2013). Dimensional model would be difficult to interpret and it is less beneficial in designing graphical user interfaces. De Choudhury analysed the emotional states of users in Twitter using dimensional model after collected the emotion tweets via emotion hashtags in Twitter (De Choudhury, 2012). Figure 2.3 presents the Circumplex Model of Affect that is studied by (Russell, 2003).

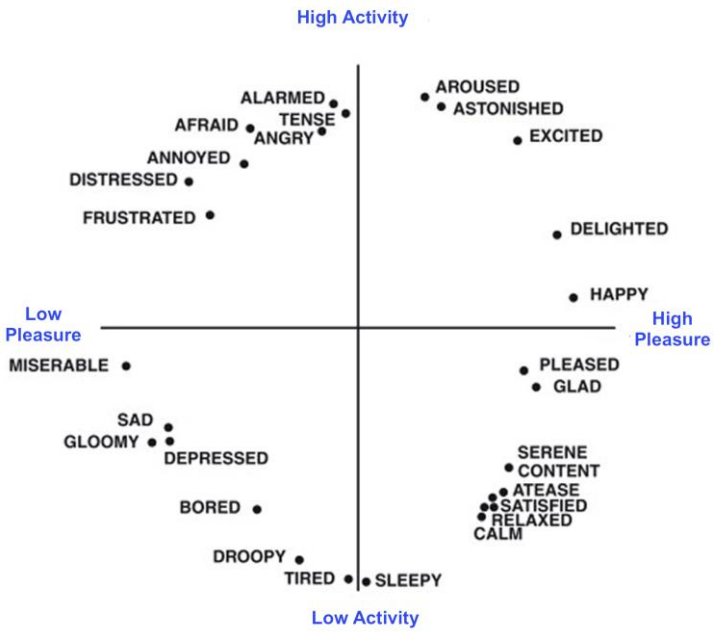


Figure 2.3: Circumplex Model of Affect including 28 affect words by J. A. Russell, 1980

2.2 Dimension Reduction

Dimension reduction is the process of reducing the number of random attributes in data. It makes machine learning algorithms analyse data much faster and easier. In feature space, dimension reduction technique is established to create a low-dimensional subset of useful features (Abualigah, 2017). It used a different strategy to choose the subset features in a dataset to produce different feature sublists (Bharti, 2015). Dimension reduction can be divided into feature selection and feature extraction.

2.2.1 Feature Selection

Feature selection is a method that selects the suitable subset of informative text features in order to improve the performance of clustering algorithm while preserving the element of the text (Abualigah, 2017). There are a lot of different type of feature selection methods such as document frequency (DF), term variance (TV), bi-normal separation (BNS) and others (Mahdieh Labani, 2018). In DF, a specific term will be evaluated when it is contained in a number of document (Liu L. e., 2005). TV will determine the higher variance values of features contain valued information. For BNS, a normal distribution will be modelled by the appearance of a term in each document (Forman, 2003). It is evaluated as an importance of a term when the corresponding area under the curve that over a threshold value.

2.2.2 Feature Extraction

Feature extraction is a method to transform a higher dimensional feature space to a lower dimension in order to reduce the size of feature vector (Mahdieh Labani, 2018). There are a few numbers of feature extraction methods to deduct the dimensionality of informative text. For instance, Singular Value Decomposition (SVD) is the method to learn and represent the relations among natural text documents and the huge amount of words (Li, 2007). While Independent Component Analysis (ICA) is a technique to discover the k component which effectively holds maximum variability of the informative data (Kolenda, 2000). Non-negative Matrix Factorization (NMF) is doing a great job in decomposing the term by sentence matrix which is created from the informative text (Batcha, 2013).

2.3 Lexicon-based Approach

Lexicon-based approach is the approach that use one or more lexical resources to extract the sentiment from text. In lexicon-based, there is a lot of lexical resources such as WordNet Affect ((Strapparava, 2008), (Wang W. e., 2012), (Balahur A. H., 2013)), WordNet synsets ((Calvo, 2013), (Roberts, 2012), (Desmet, 2013)), EmotiNet ((Balahur A. H., 2013)) and others. Lexicon-based approach is distinguished into three parts, which are keyword-based, ontology-based and statistical-based.

2.3.1 Keyword-based

Keyword-based approach is implemented where it predetermines a group of terms to annotate the text with emotion categories (Canales, 2014). This approach recognises the emotion by utilizing the use of words and their same state with words that have clearly expressed the affective meaning (Strapparava, 2008). Other than that, this approach is needed to identify the words by referring to emotional states and indirectly referring to context (Strapparava, 2008). WordNet Affect is implemented to be the affective labels for annotating the emotion to the text. Every tweet annotated the emotion automatically according to its emotion hashtag (i.e., #sadness) and the hashtag symbol is deleted from the tweet (Wang W. e., 2012).

2.3.2 Ontology-based

Ontology-based approach is also a technique where it used EmotiNet, which is the resources for emotion detection according to the affective consequence and interaction (Balahur A. H., 2011). This approach captured and facilitated the knowledge which was shared and reused by communities. EmotiNet ontology and knowledge base is used to model the actual circumstances and their relevant emotional effect as chains of actions. EmotiNet identified the emotion expressed in new examples pertaining to the categories in International Survey of Emotional Antecedents and Reaction, by calculating the similarity among the EmotiNet emotion chain and the emotion chain of the new situation (Balahur A. H., 2013). Figure 2.4 shows the concept of EmotiNet and the variables (Balahur A. H., 2013).

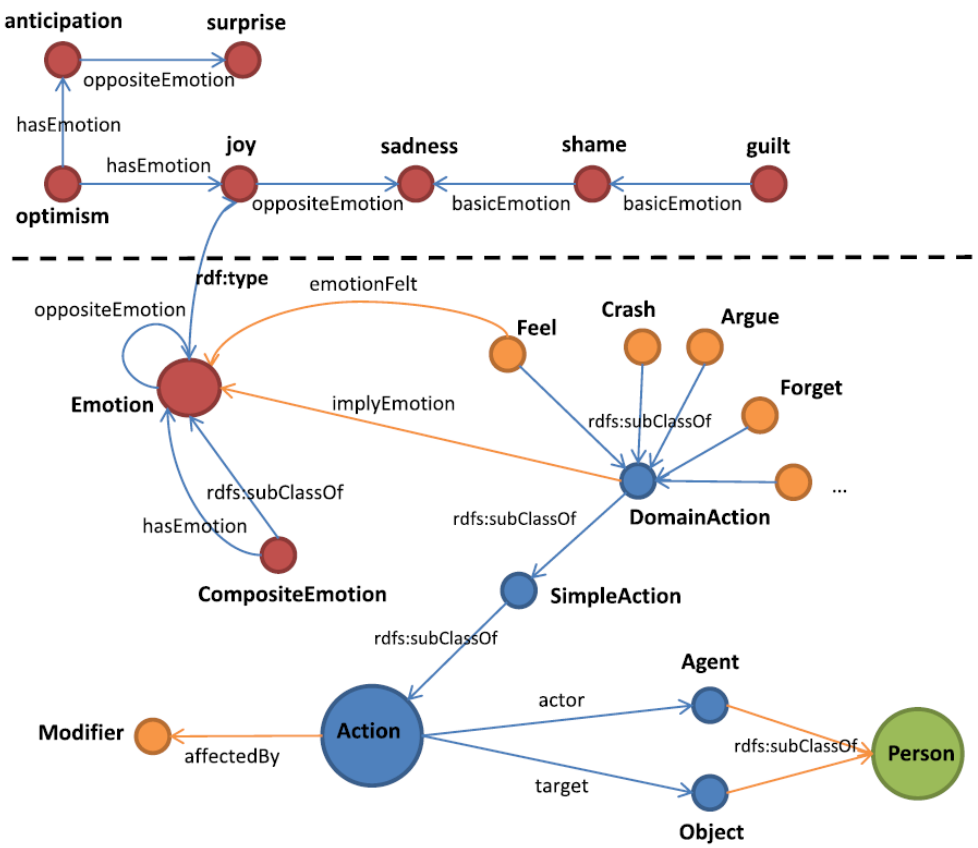


Figure 2.4: The concept of EmotiNet and the variables (Balahur A. H., 2013)

2.3.3 Statistical-based

In this approach, Latent Semantic Analysis (LSA) is implemented to inspect the relationships among a group of documents and the terms to produce a set of useful pattern (Deerwester, 1990). The latent semantic space is the semantic similarity that maps the documents into a vector space by implementing LSA. Latent semantic space is computed automatically between the texts and emotions keywords by using LSA and Hyperspace Analogue to Language (HAL) (Gill, 2008). An improved version of LSA algorithm is proposed recently for text emotion classification on ISEAR dataset (Wang X. &., 2013). The distance between the documents and the category of emotion is measured by using LSA, Probabilistic Latent Semantic Analysis (LSA) and Non-Negative Matrix Factorization (NMF) (Calvo, 2013). Latent Dirichlet allocation (LDA) is used to discover similarities between tweets even when tweets have no words in common (Roberts, 2012).

2.4 Machine Learning Approach

Machine learning is a method of data analysis which can learn from data by dealing with the model and the study of algorithms. The system not only learns the meaning of the keywords, but also take into the details of other arbitrary keywords, word co-occurrence frequency, and punctuation by training machine learning algorithm with a huge corpus of annotated text (Cambria, 2014). In machine learning, it can be categorised to supervised and unsupervised learning.

2.4.1 Supervised Learning Approach

Supervised learning approach relies on the supervised training data. The training data contained a set of training examples and it is used to teach the machine. The supervised algorithm was implemented to analyse the training data and inferred an information for mapping the new patterns (Mohri, 2012). A labelled dataset is a large and structured text which is already labelled with emotion in order to implement into the supervised learning. However, the annotation process is a time-consuming task especially in a huge amount of data.

Supervised learning approach is applied to classify automatically the short text messages such as tweets, corresponding to a fine-grained category of the emotions (Hasan, 2014). Rusell's Circumplex Model of Affect is implemented as a model of emotion. Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree and Naïve Bayes were used to compare the accuracy of classifying tweet messages (Hasan, 2014).

Supervised learning approach adopted as a baseline for automatically annotating emotions for tweets (Roberts, 2012). Roberts is using SVM classifiers to detect the emotions annotated in the corpus. A tweet is allowed to be annotated with more than one of the emotion by using the combination of the separate classifiers, which is also can be called single multi-label classifier. This approach is used to estimate the existence of emotions in hidden sentences (Desmet, 2013). Figure 2.4 shows the system used to identify emotion in the corpus (Roberts, 2012).

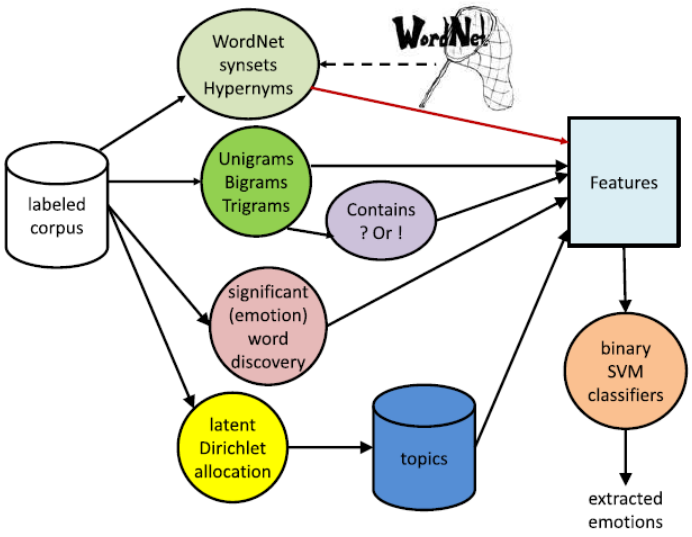


Figure 2.4: System used for identifying emotion automatically in the labelled corpus (Roberts, 2012)

2.4.2 Unsupervised Learning Approach

Unsupervised learning approach is a technique in which the algorithm uses only the predictor attribute values (Mac Kim, 2011). This approach needs to discover hidden patterns in unlabelled data to construct the models for classification of emotion (Mohri, 2012). Unsupervised learning approach was proposed by integrating Latent Semantic Analysis with WordNet Affect to work based on categorical emotion model (Strapparava, 2008). The reason behind this decision is to focus the study of lexical resources and prevent the bias of the categorization of participants.

Unsupervised learning approach is applied to the dimensional emotion model and categorical model into the experiments (Calvo, 2013). The emotional interpretations of a text should involve more than one annotators, which is preferred by unsupervised methods. The limitation of the knowledge base of lexical approach affected the accuracy when the vocabulary is not fit to the corpora.

Unsupervised learning approach is used to calculate the score of semantic relatedness among an emotion concept and the word (Agrawal, 2012). The purpose of this decision is because the number of emotion categories is not restricted. This approach also does not rely on annotated training data or affect dictionaries (Agrawal, 2012). It relied on general cues and can ignore the strong evidence potentially drawn from the topic-specific documents (Mac Kim, 2011).

2.5 Summary

A comparison is made to show the technique, model and resources that have been used by the researchers. The table for strength and weakness for each paper also made. Table 2.1 shows the strength and the weakness of each paper that have been studied. Table 2.2 presents the comparison of each paper in technique, emotion model and lexical resources.

Paper	Strength	Weakness
Emotions in text: Dimensional and Categorical Models (Calvo, 2013)	- No need to annotate datasets - Emotional interpretations of a text are more than one annotator needed, which is preferred by unsupervised methods.	- Limitation of the lexical approach background knowledge
EmpaTweet: Annotating and Detecting Emotions on Twitter (Roberts, 2012)	- Easy to be trained on the annotations in order to automatically extract emotions from tweets.	- Time-consuming for annotating the emotion
Learning to identify emotions in text (Strapparava, 2008)	- Avoid biasing participants toward simple "text categorization" approach	- There is no training will be provided. Study of emotion lexical semantics is focused.
Harnessing Twitter 'Big Data' for Automatic Emotion Identification (Wang W. e., 2012)	- Emotion is more reliable and natural as the emotion is provided by writers	- The emotion may not accurate as the writer maybe want to express in opposite way - Time-consuming to manually annotate the data
Emotion detection in suicide notes (Desmet, 2013)	- Estimate the existence of emotions in hidden sentences.	- The modification is not captured by bag-of-words, which is influenced the

		significant sequences.
Detecting implicit expressions of affect in text using EmotiNet (Balahur A. H., 2013)	- High degree of flexibility because relevant emotional effect be the chain of actions	- Low expressivity of the resource - The modifier is not included as attributes in the roles.
Unsupervised Emotion Detection from Text using Semantic and Syntactic Relations (Agrawal, 2012)	- No limitation to the number of emotion categories as no dictionaries or annotated training data has relied on	- The corpus determined the semantic relatedness scores
Using YouTube comments for text-based emotion recognition (Hajar, 2016)	- No labelling is required in the data corpus which is a time-consuming task	- Limited range of target emotions

Table 2.1: Strength and weakness of every paper that has been studied

Paper Name	Technique					Emotion Model		Resource / Dictionary			
	Lexicon- based			Machine Learning							
	Knowledge-based	Ontology-based	Statistical based	Supervised	Unsupervised	Dimension	Categorical	WordNet Synsets	WordNet Affect	EmotiNet	Others
Calvo, R. A., & Mac Kim, S. (2013). Emotions in text: dimensional and categorical models. *Computational Intelligence* , *29* (3), 527-543.			✓		✓	✓	✓	✓			
Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., & Harabagiu, S. M. (2012, May). EmpaTweet: Annotating and Detecting Emotions on Twitter. In *LREC* (Vol. 12, pp. 3806-3813).	✓			✓			✓	✓			
Strapparava, C., & Mihalcea, R. (2008, March). Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 1556-1560). ACM.	✓				✓		✓		✓		
Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012, September). Harnessing twitter" big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)* (pp. 587-592). IEEE.	✓			✓			✓		✓		MPQA Lexicon, Part-of-Speech (POS), LIWC Dictionary
Desmet, B., & Hoste, V. (2013). Emotion detection in suicide notes. *Expert Systems with Applications* , *40* (16), 6351-6358.			✓	✓			✓	✓			SentiWorkNet, Part-of-Speech (POS),Bags-of-words
Balahur, A., Hermida, J. M., Montoyo, A., & Muñoz, R. (2013). Detecting implicit expressions of affect in text using EmotiNet and its extensions. *Data & Knowledge Engineering* , *88* , 113-125.		✓		✓			✓		✓	✓	VerbOcean, ConceptNet, SentiWordNet, Core WordNet, Linguistic Inquiry and Word Count (LIWC)
Agrawal, A., & An, A. (2012, December). Unsupervised emotion detection from text using semantic and syntactic relations. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01* (pp. 346-353). IEEE Computer Society.			✓		✓		✓				
Hajar, M. (2016). Using YouTube Comments for Text-based Emotion Recognition. *Procedia Computer Science* , *83* , 292-299.			✓		✓		✓				

Table 2.2: The technique, emotion model and lexical resource that have been implemented by each paper

Chapter 3 : Methodology

3.1 Theoretical Framework

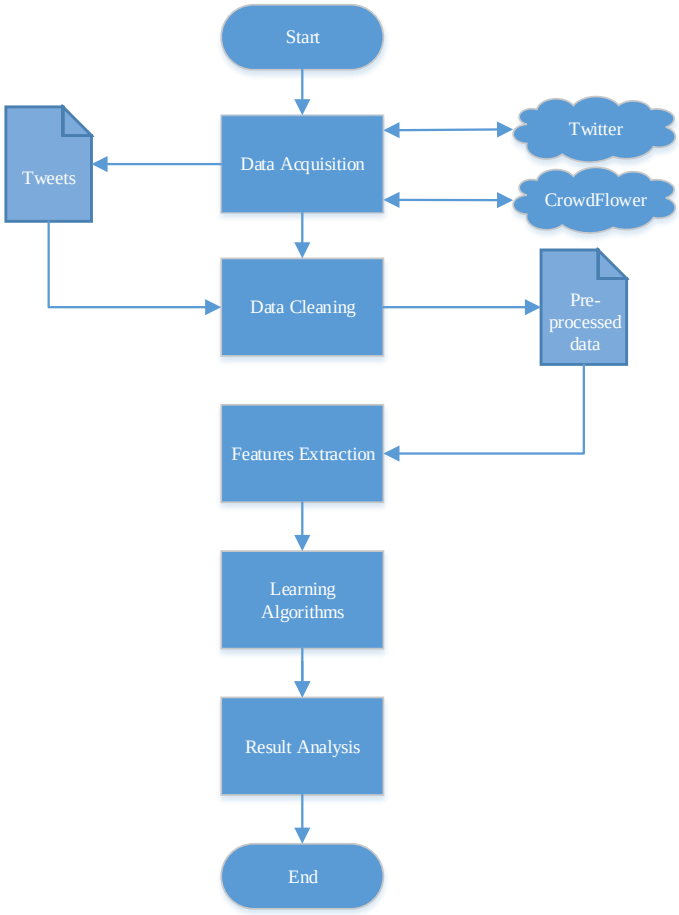


Figure 3.1: The steps involved in order to complete this project

In this project, the training dataset is obtained from CrowdFlower. This dataset is created from Twitter API. Four attributes which were tweet id, sentiment, author and content were included in this dataset. In this dataset, the tweets were annotated with 13 emotions which were enthusiasm, fun, happiness, love, anger, boredom, hate, relief, sadness, worry, empty, neutral and surprise. Table 3.1 shows the attribute of tweets dataset.

Number	Feature	Description	Type
1	tweet_id	ID of Twitter users	integer
2	sentiment	Annotated emotion	string
3	author	Username of Twitter	string
4	content	Tweets	string

Table 3.1: The attribute of tweets dataset

Section 3.2 describes the data cleaning and feature extraction in more details.

In data modelling, three experiments are carried out. The machine learning algorithms which are Support Vector Machine (SVM), Random Forest and Naïve Bayes were adopted in the experiments. Firstly, the pre-processed training dataset is split into 70 percent of train set and 30 percent of test set for evaluating the model fit on the train set.

After that, the pre-processed training dataset will be fully adopted as train set and pre-processed real data will be the test set to predict the outcomes by using the predictive model. Other than that, k-fold cross-validation is also implemented to estimate how well the model fit on the dataset. Area Under the Receiver Operating Characteristic Curve (ROC AUC) is used as the one of the data evaluation.

3.2 Data Pre-processing

In data pre-processing, Pandas are used to load the CSV files in data frame structure. A copy of dataset will be stored for model fitting. After that, regular expression operation (REGEX) is used to search the matching strings from the text and perform the operation on it.

Firstly, emoticons will be replaced with words by using REGEX and saved for the second dataset. For example, ":-)" will become happy, ":-(" will become sad. Next, Twitter username, hashtag symbol and &amp are removed from the dataset. While URL is changed to "website" word. The letters of the dataset are converted to lowercase. Then the dataset will be saved as the third dataset.

Other than that, punctuation is removed from the data by using REGEX. Then the dataset is stored as the fourth dataset. Furthermore, abbreviation will be reconstructed and saved as the fifth dataset. For example, "btw" become "by the way", "lol" become "laughing out loud".

After that, stop words package from Natural Language Toolkit (NLTK) is used to remove the stop words. For example, "I", "he", "she", "has", "have" and others. After removing stop words, the dataset will be saved as the sixth dataset.

The next data pre-processing step is word correction. Vocabulary resources will be adopted to correct the words. If the word is missing some letter, the word will be corrected. For instance, "speling" become spelling, "freakin" become freaking. After correction, the

dataset is stored as the seventh dataset. Lastly, texts are reduced to the root form by stemming package and it will save as the eighth dataset.

3.3 Feature Extraction

In feature extraction, Term Frequency-Inverse Document Frequency (TF-IDF) will be implemented to transform the eight types of the dataset into a matrix of TF-IDF feature which can be read by the classification algorithm. The maximum and the minimum of document frequency is set for an optimal matrix before fitting the model. The sublinear term frequency function and inverse-document-frequency function will be adjusted for a better matrix.

3.4 Data Modelling

In this stage, the dataset will be split into 70 percent of train set to fit the model and 30 percent of test set to test the predictive model. Other than that, k-fold cross validation will also be implemented for the data validation.

For classification algorithms, Support Vector Machine (SVM), Random Forest and Naïve Bayes are adopted for data modelling. In SVM, there are four types of Support Vector Classification (SVC), which are SVC with linear kernel, LinearSVC (linear kernel), SVC with RBF kernel and SVC with polynomial kernel. In this project, SVC with linear kernel will be adopted as the algorithm of SVM.

In Random Forest, the number of trees will be set 100 to be used in the forest. In Naïve Bayes, there are three types of event models, which are Gaussian Naïve Bayes, Multinomial Naïve Bayes and Bernoulli Naïve Bayes. While Bernoulli Naïve Bayes will be taken as the algorithm for data modelling.

3.5 Evaluation Matrix

The performance of emotion detection is evaluated by precision, recall and F-score. Firstly, precision and recall are explained in the figure below. Figure 3.2 shows the measures for precision and recall.

$$Precision = \frac{Categories\ found\ and\ correct}{Total\ categories\ found} = \frac{TP}{TP + FP}$$

$$Recall = \frac{Categories\ found\ and\ correct}{Total\ categories\ correct} = \frac{TP}{TP + FN}$$

Figure 3.2: The measures for precision and recall

True Positive (TP) is the number of data that are correctly categorised to the class, while False Positive (TP) is the number of incorrectly categorised instances. False Negative (FN) stands for the number of instances incorrectly classified as negative.

Other than that, F-Score/F-measure is used to be a measure of the effectiveness of classification. Figure 3.3 presents the definition of F-score.

$$F\ score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Figure 3.3: The definition of F-score

F-score is implemented to balance the weight of precision and recall. The values of F-score are in between 0 and 1. The bigger the values of F-score, the higher the quality of classification.

Other than that, Area Under the Receiver Operating Characteristic Curve (ROC AUC) is applied to compare the True Positive Rate (TPR) and False Positive Rate (FPR). TPR, also named sensitivity will describe the correct predicted of positive among all positive sampling during testing. While FPR, also named specificity defines the incorrect predicted of positive among all negative sampling during testing. A good classification is determined when ROC skewed to the upper left, which is the higher rate of True Positive.

$$TPR = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$FPR = \frac{False\ Positive}{False\ Positive + True\ Negative}$$

Figure 3.4: The formula of TPR and FPR

3.6 System Requirements

3.6.1 Hardware

This project is running under i7-4500U processors, which the processor base frequency is up to 2.40 GHz. The RAM which is 8GB is implemented to execute the emotion detection. 2GB of NVIDIA GeForce 820M is adopted for the GPU part.

3.6.2 Software

In this project, Jupyter Notebook is mainly used for the programming part. While Anaconda Navigator is used for launching the Jupyter Notebook, handling Conda packages and managing the environment. Other than that, Microsoft Word 2016 is used for the documentation while Microsoft Visio 2016 is implemented for flowchart diagram. Next, Microsoft Excel 2016 is used for constructing the table. Microsoft Project 2016 is used to create Gantt Chart.

3.4.3 Package and library included in Python

Firstly, Pandas are used for the data structures and file reading. Tabulate is very useful especially in pretty print a small table. Numpy package is used for array processing and the matrices. NLTK is a great library when processing raw text by using their packages such as stop words package, stem package and others. NLTK also provided packages for calculating the frequencies of words. Other than that, Pickle is an effective method to serialize and de-serialize Python object structure. Word Cloud package is a powerful visualization to highlight the important textual data points. While Matplotlib is a plotting library which provides a series of quality graphing. For machine learning part, Scikit-learn is the best library for data analysis and data mining.

3.7 Implementation Plan

Figure 3.5 shows the implementation plan for this project.

ID	Task Name	Duration	Start	Finish
1	**Final Year Project**	**155 days**	**Mon 10/07/**	**Fri 09/02/18**
2	**Phase One**	**58 days**	**Mon 10/07/**	**Wed 27/09/17**
3	Literature Revie	22 days	Mon 10/07/	Tue 08/08/17
4	Proposed Soluti	19 days	Wed 09/08/	Mon 04/09/17
5	Documentation	12 days	Tue 05/09/1	Wed 20/09/17
6	Presentation	1 day	Wed 27/09/	Wed 27/09/17
7	**Phase Two**	**60 days**	**Mon 20/11/**	**Fri 09/02/18**
8	Implementatio	21 days	Mon 20/11/	Mon 18/12/17
9	Result Analysis	16 days	Tue 19/12/1	Tue 09/01/18
10	Paper Writing	11 days	Wed 10/01/	Wed 24/01/18
11	Documentation	12 days	Thu 25/01/1	Fri 09/02/18

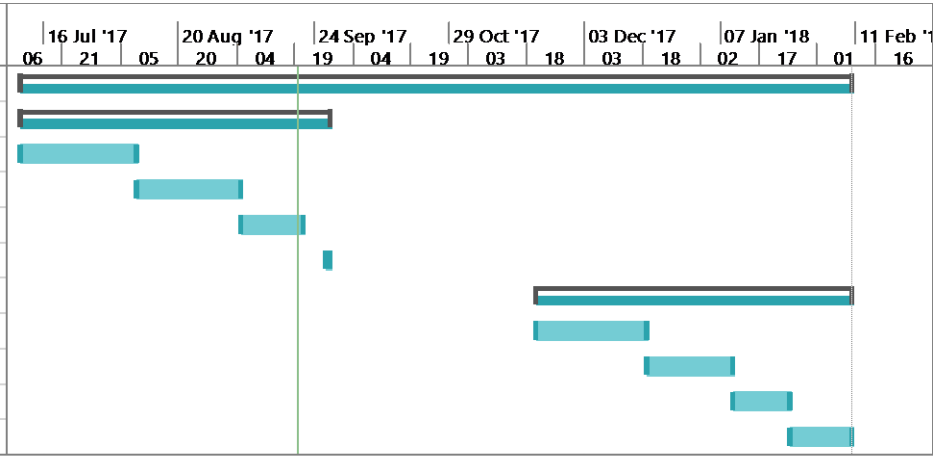


Figure 3.5: Gantt Chart

Chapter 4 : Result and Discussion

4.1 Introduction

In this chapter, a comprehensive analysis is carried out for the experimental dataset, data pre-processing, feature extraction and data modelling. A few of graphs will be demonstrated to describe the overview of the variables in order to understand the dataset. Other than that, comparison tables will be presented to show the difference of the result for a better understanding in this chapter.

4.2 Experimental Dataset

The dataset is implemented into data visualization to easily understand the significance of dataset. The size of the dataset is 40,000 of tweets. The dataset is annotated with 13 emotions which are enthusiasm, fun, happiness, love, anger, boredom, hate, relief, sadness, worry, empty, neutral and surprise. Figure 4.1 shows the number of tweets for every emotion in the dataset.

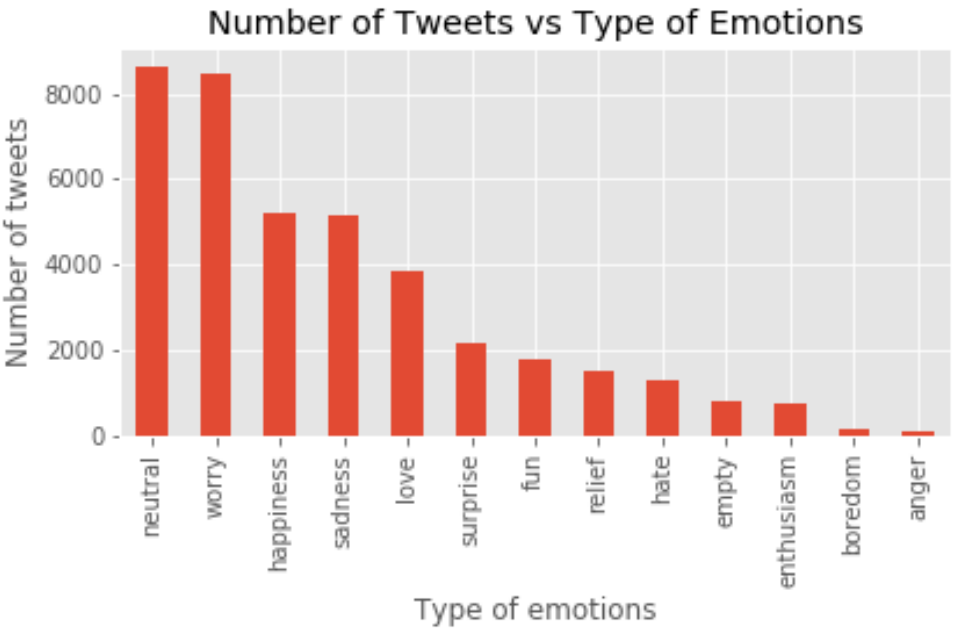


Figure 4.1: The number of tweets for each emotion

The emotion of dataset is categorized into three categories, which are positive, negative and neutral. Figure 4.2 presents the percentage of each sentiment.

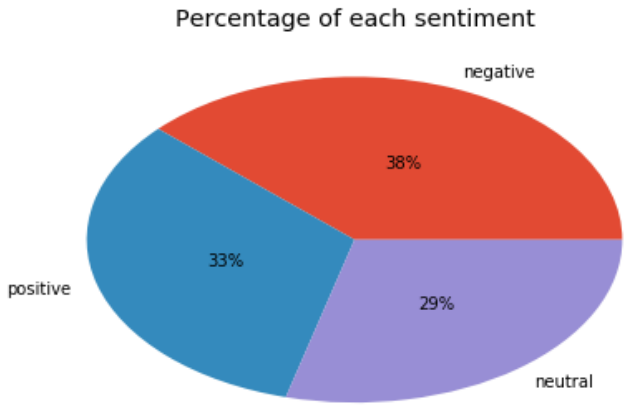


Figure 4.2: The percentage of each sentiment in the dataset

In our dataset, the neutral emotion has empty, neutral and surprise. Figure 4.3 shows the number of tweets for neutral emotion.

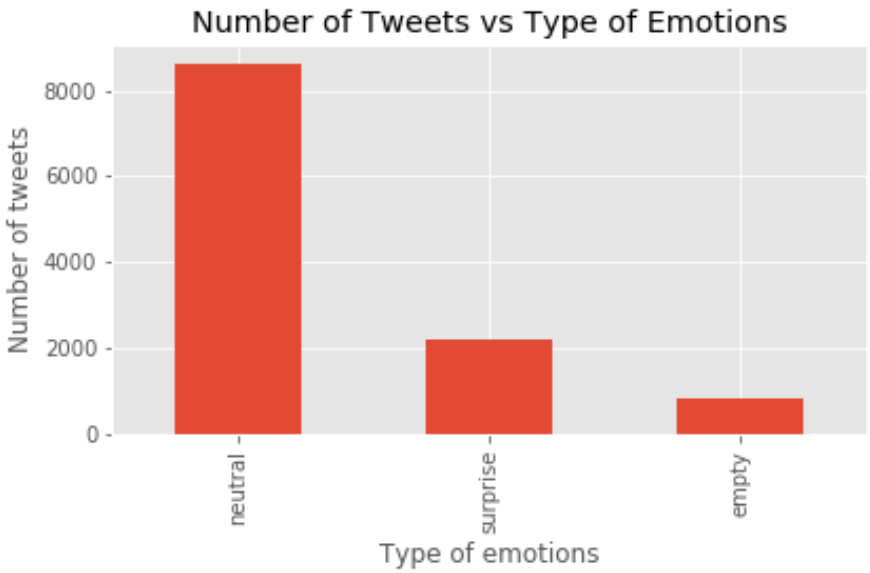


Figure 4.3: The number of tweets for neutral emotion

While positive emotion includes enthusiasm, fun, happiness, love and relief. Figure 4.4 presents the number of tweets for positive emotion.

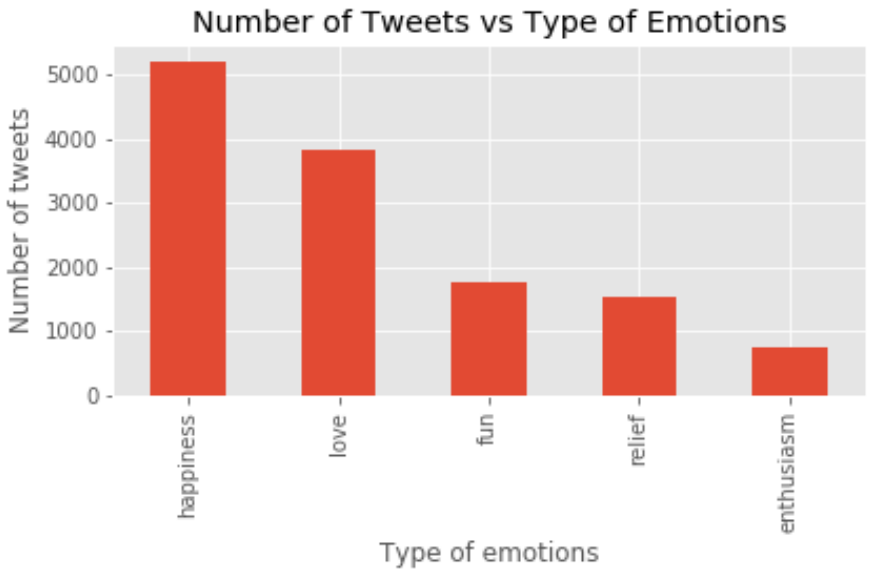


Figure 4.4: The number of tweets for positive emotion

Negative emotion has anger, boredom, hate, sadness and worry. Figure 4.5 presents the number of tweets for negative emotion.

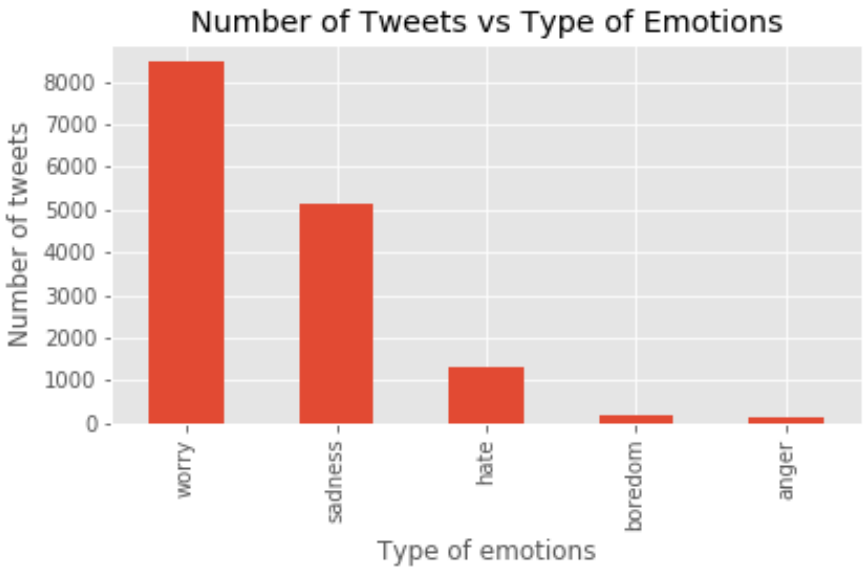


Figure 4.5: The number of tweets for negative emotion

4.3 Data Pre-processing Result

In this project, 40,000 of tweets are obtained from CrowdFlower. The dataset contained four attributes which are tweet id, sentiment, author and content. Only two attributes which are sentiment and content will be used in this project. After each data pre-processing step, the dataset will be stored for next data pre-processing steps and performance comparison of different types of data pre-processing step. All figure below will be shown 10 rows of the dataset only. Figure 4.6 shows the raw data before applying any data pre-processing.

```
    sentiment    content
--  -----------  ------------------------------------------------------------------------------------------------------------
----------------------
 0  empty        @tiffanylue i know  i was listenin to bad habit earlier and i started freakin at his part =[
 1  sadness      Layin n bed with a headache  ughhhh...waitin on your call...
 2  sadness      Funeral ceremony...gloomy friday...
 3  enthusiasm   wants to hang out with friends SOON!
 4  neutral      @dannycastillo We want to trade with someone who has Houston tickets, but no one will.
 5  worry        Re-pinging @ghostridah14: why didn't you go to prom? BC my bf didn't like my friends
 6  sadness      I should be sleep, but im not! thinking about an old friend who I want. but he's married now. damn, &amp; he w
ants me 2! scandalous!
 7  worry        Hmmm. http://www.djhero.com/ is down
 8  sadness      @charviray Charlene my love. I miss you
 9  sadness      @kelcouch I'm sorry  at least it's Friday?
```

Figure 4.6: The raw data before applying any data pre-processing

Firstly, emoticons are converted to informative text. Figure 4.7 shows the data after the emoticons converted to text.

	sentiment	content
0	empty	@tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part sad
1	sadness	Layin n bed with a headache ughhhh...waitin on your call...
2	sadness	Funeral ceremony...gloomy friday...
3	enthusiasm	wants to hang out with friends SOON!
4	neutral	@dannycastillo We want to trade with someone who has Houston tickets, but no one will.
5	worry	Re-pinging @ghostridah14: why didn't you go to prom? BC my bf didn't like my friends
6	sadness	I should be sleep, but im not! thinking about an old friend who I want. but he's married now. damn, & he wants me 2! scandalous!
7	worry	Hmmm. http://www.djhero.com/ is down
8	sadness	@charviray Charlene my love. I miss you
9	sadness	@kelcouch I'm sorry at least it's Friday?

Figure 4.7: The data after the hashtag symbol is removed

Twitter username mention, &amp and hashtag symbol are filtered from the content. While URL will be changed to “website” word. The letter will be converted to lowercase. Figure 4.8 presents the result after applying these five types of text-processing.

	sentiment	content
0	empty	i know i was listenin to bad habit earlier and i started freakin at his part sad
1	sadness	layin n bed with a headache ughhhh...waitin on your call...
2	sadness	funeral ceremony...gloomy friday...
3	enthusiasm	wants to hang out with friends soon!
4	neutral	we want to trade with someone who has houston tickets, but no one will.
5	worry	re-pinging why didn't you go to prom? bc my bf didn't like my friends
6	sadness	i should be sleep, but im not! thinking about an old friend who i want. but he's married now. damn, he wants me 2! scandalous!
7	worry	hmmm. website is down
8	sadness	charlene my love. i miss you
9	sadness	i'm sorry at least it's friday?

Figure 4.8: The result after applying these five types of text-processing

After that, punctuation will be fully removed from the data. Figure 4.9 shows the result after removing punctuation.

```
    sentiment    content
--  -----------  ------------------------------------------------------------------------------------------------------------------
---------
 0  empty        i know  i was listenin to bad habit earlier and i started freakin at his part sad
 1  sadness      layin n bed with a headache  ughhhhwaitin on your call
 2  sadness      funeral ceremonygloomy friday
 3  enthusiasm   wants to hang out with friends soon
 4  neutral      we want to trade with someone who has houston tickets but no one will
 5  worry        repinging  why didnt you go to prom bc my bf didnt like my friends
 6  sadness      i should be sleep but im not thinking about an old friend who i want but hes married now damn  he wants me 2 s
candalous
 7  worry        hmmm website is down
 8  sadness      charlene my love i miss you
 9  sadness      im sorry  at least its friday
```

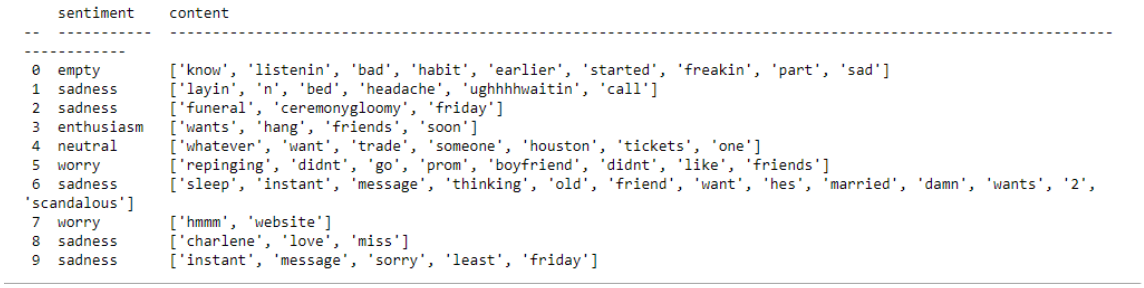
Figure 4.9: The result after removing punctuation

The short form of words is restructured into full sentences in tweets. Figure 4.10 presents the result after the short form of words is restructured into full sentences.

```
    sentiment    content
--  -----------  ------------------------------------------------------------------------------------------------------------------
----------------------
 0  empty        i know  i was listenin to bad habit earlier and i started freakin at his part sad
 1  sadness      layin n bed with a headache  ughhhhwaitin on your call
 2  sadness      funeral ceremonygloomy friday
 3  enthusiasm   wants to hang out with friends soon
 4  neutral      whatever want to trade with someone who has houston tickets but no one will
 5  worry        repinging  why didnt you go to prom because my boyfriend didnt like my friends
 6  sadness      i should be sleep but instant message not thinking about an old friend who i want but hes married now damn  he
wants me 2 scandalous
 7  worry        hmmm website is down
 8  sadness      charlene my love i miss you
 9  sadness      instant message sorry  at least its friday
```

Figure 4.10: The result after the short form of words is restructured into full sentences

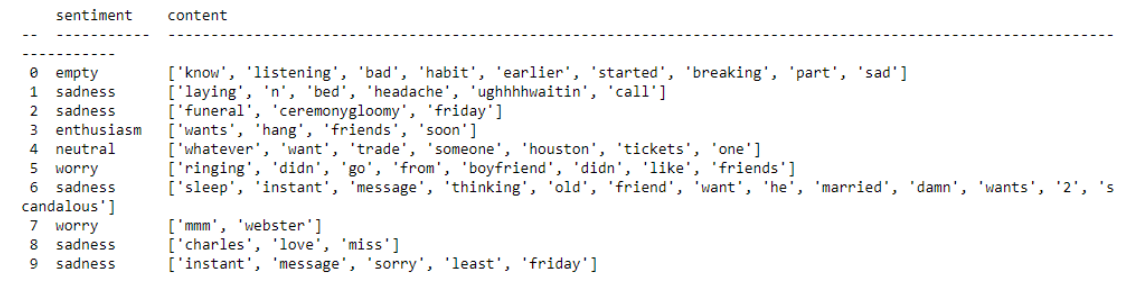
After that, stop words are removed from the data. Before that, the words will be tokenized and it will remain in tokenized states. Figure 4.11 shows the result after stop words are removed.



	sentiment	content
0	empty	['know', 'listenin', 'bad', 'habit', 'earlier', 'started', 'freakin', 'part', 'sad']
1	sadness	['layin', 'n', 'bed', 'headache', 'ughhhhwaitin', 'call']
2	sadness	['funeral', 'ceremonygloomy', 'friday']
3	enthusiasm	['wants', 'hang', 'friends', 'soon']
4	neutral	['whatever', 'want', 'trade', 'someone', 'houston', 'tickets', 'one']
5	worry	['repinging', 'didnt', 'go', 'prom', 'boyfriend', 'didnt', 'like', 'friends']
6	sadness	['sleep', 'instant', 'message', 'thinking', 'old', 'friend', 'want', 'hes', 'married', 'damn', 'wants', '2', 'scandalous']
7	worry	['hmmm', 'website']
8	sadness	['charlene', 'love', 'miss']
9	sadness	['instant', 'message', 'sorry', 'least', 'friday']

Figure 4.11: The result after stop words are removed

Next, the words will be corrected if the words are missing some letter. Figure 4.12 shows the data after the words which are missing letter are corrected.



	sentiment	content
0	empty	['know', 'listening', 'bad', 'habit', 'earlier', 'started', 'breaking', 'part', 'sad']
1	sadness	['laying', 'n', 'bed', 'headache', 'ughhhhwaitin', 'call']
2	sadness	['funeral', 'ceremonygloomy', 'friday']
3	enthusiasm	['wants', 'hang', 'friends', 'soon']
4	neutral	['whatever', 'want', 'trade', 'someone', 'houston', 'tickets', 'one']
5	worry	['ringing', 'didn', 'go', 'from', 'boyfriend', 'didn', 'like', 'friends']
6	sadness	['sleep', 'instant', 'message', 'thinking', 'old', 'friend', 'want', 'he', 'married', 'damn', 'wants', '2', 'scandalous']
7	worry	['mmm', 'webster']
8	sadness	['charles', 'love', 'miss']
9	sadness	['instant', 'message', 'sorry', 'least', 'friday']

Figure 4.12: The data after the words which are missing letters are corrected

Another data pre-processing is stemming. The dataset will be stemmed in root form. Figure 4.13 presents the result after the words stemmed in root form.

```
    sentiment    content
--  -----------  ------------------------------------------------------------------------------------------------------------
-
 0  empty        ['know', 'listen', 'bad', 'habit', 'earlier', 'start', 'break', 'part', 'sad']
 1  sadness      ['lay', 'n', 'bed', 'headach', 'ughhhhwaitin', 'call']
 2  sadness      ['funer', 'ceremonygloomi', 'friday']
 3  enthusiasm   ['want', 'hang', 'friend', 'soon']
 4  neutral      ['whatev', 'want', 'trade', 'someon', 'houston', 'ticket', 'one']
 5  worry        ['ring', 'didn', 'go', 'from', 'boyfriend', 'didn', 'like', 'friend']
 6  sadness      ['sleep', 'instant', 'messag', 'think', 'old', 'friend', 'want', 'he', 'marri', 'damn', 'want', '2', 'scanda
l']
 7  worry        ['mmm', 'webster']
 8  sadness      ['charl', 'love', 'miss']
 9  sadness      ['instant', 'messag', 'sorri', 'least', 'friday']
```

Figure 4.13: The result after the words stemmed in root form

4.4 Feature Extraction

In feature extraction, Term Frequency-Inverse Document Frequency (TF-IDF) is implemented to convert the dataset to a matrix of TF-IDF features. The maximum of document frequency is adjusted to five while the minimum of document frequency is adjusted to 0.8 for building the vocabulary in the matrix. It enabled the sublinear term frequency function for scaling the term frequency in logarithmic scale. Other than that, it is also enabled the inverse-document-frequency function for reweighting. Figure 4.14 shows partial matrix after dataset conversion.

```
(0, 2877)     0.19544676277
(0, 5612)     0.167579269882
(0, 3036)     0.407006276161
(0, 5248)     0.0988504847676
(0, 491)      0.225914032014
(0, 2287)     0.427004916501
(0, 1635)     0.329089194344
(0, 296)      0.123340490347
(0, 4851)     0.309787359519
(0, 2024)     0.349885085726
(0, 411)      0.165560146656
(0, 2438)     0.250507387978
(0, 3773)     0.304725132148
(1, 569)      0.300967084252
(1, 5739)     0.205181090258
(1, 2361)     0.370149177574
(1, 5437)     0.523925258854
(1, 5584)     0.494012379312
(1, 3668)     0.187453859326
(1, 5870)     0.238963533696
(1, 867)      0.345717649458
(2, 2071)     0.539752244312
(2, 948)      0.57347301423
(2, 2152)     0.504814119809
(2, 2037)     0.353495291555
```

Figure 4.14: Partial matrix after dataset conversion

4.5 Data Modelling

4.5.1 Introduction

The selected learning algorithms to run data modelling are Support Vector Machine (SVM), Random Forest and Naïve Bayes. Eight types of dataset are fitting into the training model to compare the performance of dataset and identify which dataset in which learning algorithm produced a better result. The dataset will be split into 70 percent of train set and 30 percent of test set. Other than that, the result will be shown after 10-fold cross validation is applied. For the ROC curve part, One-Vs-Rest Classifier is used for fitting the classifier with the class and it will against all the other classes. A curve graph will be presented based on the 13 classes. Table 4.1 shows the description of each class in ROC Curve.

Class	Description
Micro-average ROC curve	Assuming the element of label indicator matrix as a binary prediction
Macro-average ROC curve	Providing the actual values to the classification of each label
ROC curve of class 0	neutral emotion
ROC curve of class 1	surprise emotion
ROC curve of class 2	empty emotion
ROC curve of class 3	happiness emotion
ROC curve of class 4	love emotion
ROC curve of class 5	fun emotion
ROC curve of class 6	relief emotion
ROC curve of class 7	enthusiasm emotion
ROC curve of class 8	worry emotion
ROC curve of class 9	sadness emotion
ROC curve of class 10	hate emotion
ROC curve of class 11	boredom emotion
ROC curve of class 12	anger emotion

Table 4.1: The description of each class in ROC Curve

4.5.2 Dataset 1

For dataset 1, there are no data-processing steps applied. Table 4.1 describes the result of SVM, Random Forest and Naïve Bayes for dataset 1. While Table 4.2 shows the result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes. Figure 4.15, Figure 4.16 and Figure 4.17 show the ROC curve of different algorithms.

	SVM			Random Forest			Naïve Bayes		
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
anger	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
boredom	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
empty	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
enthusiasm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fun	0.24	0.02	0.04	0.12	0.01	0.02	0.12	0.03	0.05
happiness	0.34	0.40	0.37	0.34	0.31	0.32	0.34	0.30	0.32
hate	0.47	0.16	0.24	0.52	0.13	0.21	0.22	0.05	0.08
love	0.52	0.38	0.44	0.48	0.37	0.41	0.45	0.35	0.39
neutral	0.33	0.55	0.41	0.31	0.54	0.40	0.32	0.60	0.42
relief	0.24	0.02	0.04	0.28	0.03	0.06	0.13	0.03	0.05
sadness	0.33	0.21	0.26	0.34	0.17	0.22	0.29	0.24	0.26
surprise	0.26	0.04	0.06	0.21	0.02	0.04	0.13	0.04	0.06
worry	0.34	0.50	0.40	0.31	0.53	0.39	0.35	0.43	0.38

Table 4.2: The result of SVM, Random Forest and Naïve Bayes

Type of Learning Algorithm	Mean Score
Support Vector Machine	0.34714
Random Forest	0.32991
Naïve Bayes	0.32569

Table 4.3: The result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes

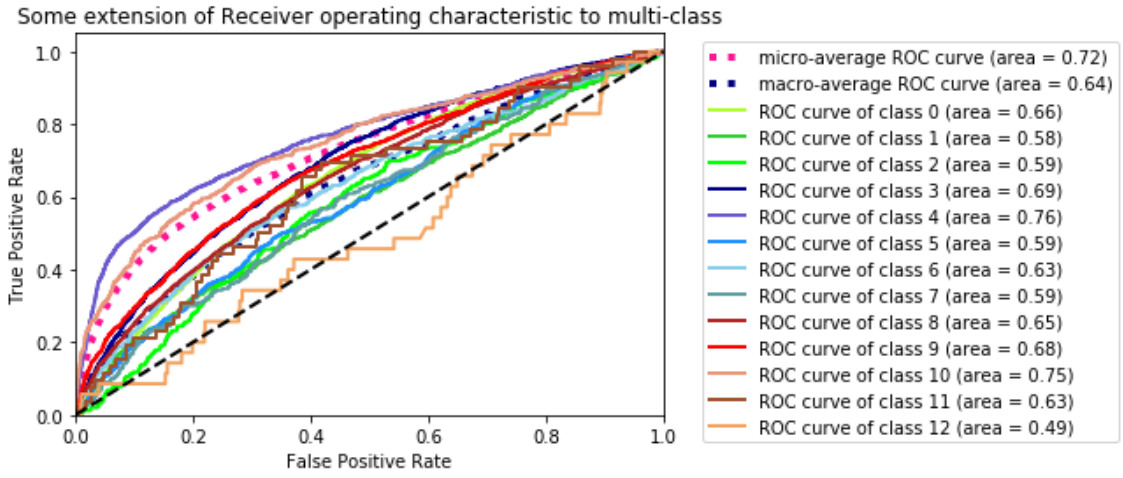


Figure 4.15: The ROC curve of Support Vector Machine

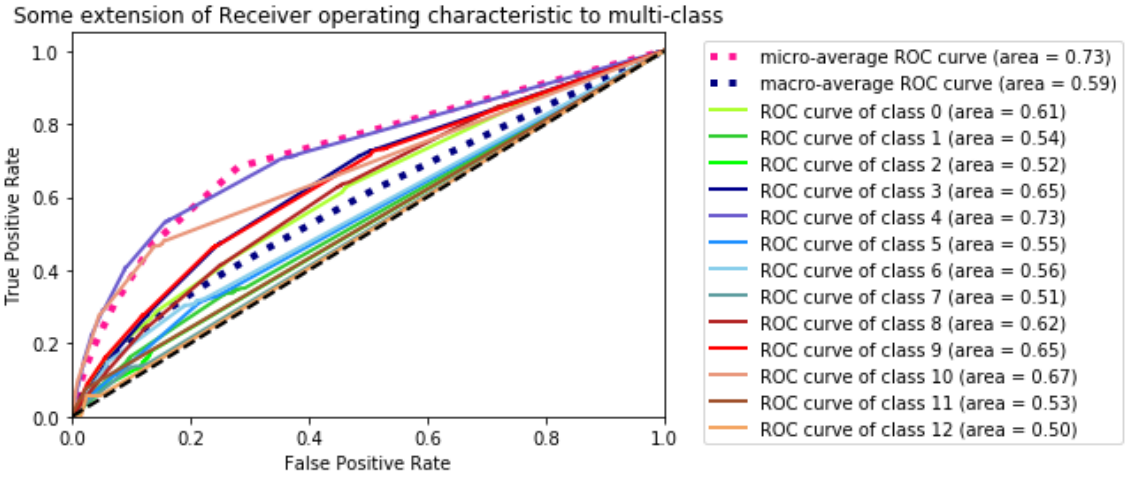


Figure 4.16: The ROC curve of Random Forest

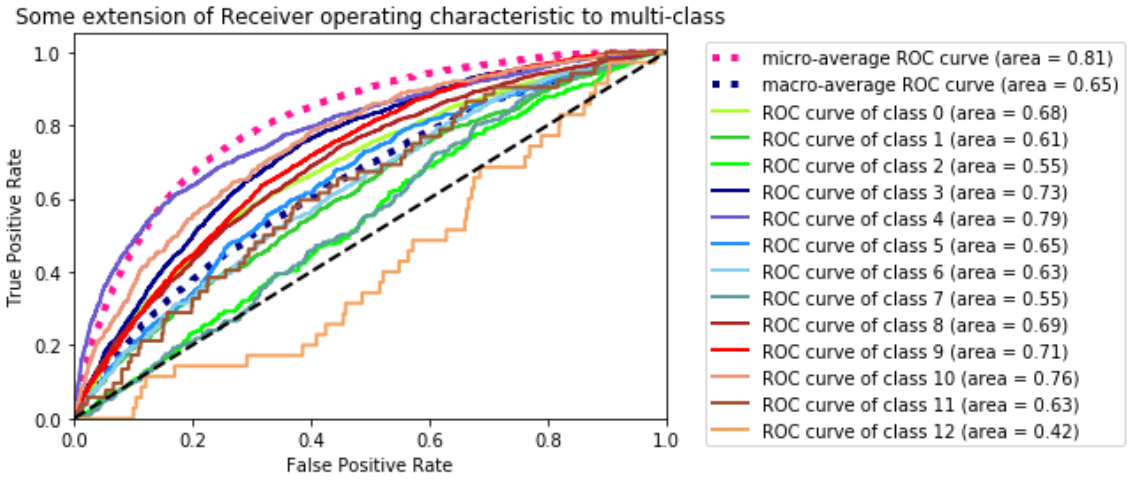


Figure 4.17: The ROC curve of Naïve Bayes

4.5.3 Dataset 2

For dataset 2, the emoticons are replaced with words. Table 4.3 describes the result of SVM, Random Forest and Naïve Bayes for dataset 2. While Table 4.4 shows the result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes. Figure 4.18, Figure 4.19 and Figure 4.20 shows the ROC curve of different algorithms.

	SVM			Random Forest			Naïve Bayes		
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
anger	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
boredom	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
empty	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
enthusiasm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fun	0.22	0.02	0.04	0.12	0.01	0.02	0.12	0.03	0.05
happiness	0.34	0.40	0.37	0.33	0.30	0.32	0.35	0.30	0.32
hate	0.46	0.16	0.24	0.52	0.11	0.19	0.23	0.05	0.08
love	0.52	0.38	0.44	0.48	0.38	0.42	0.45	0.35	0.39
neutral	0.33	0.55	0.41	0.31	0.54	0.40	0.32	0.60	0.42
relief	0.23	0.02	0.04	0.25	0.03	0.06	0.13	0.03	0.05
sadness	0.34	0.22	0.26	0.33	0.16	0.21	0.29	0.24	0.26
surprise	0.26	0.04	0.06	0.27	0.02	0.04	0.13	0.04	0.06
worry	0.34	0.50	0.41	0.31	0.52	0.39	0.35	0.43	0.39

Table 4.4: The result of SVM, Random Forest and Naïve Bayes

Type of Learning Algorithm	Mean Score
Support Vector Machine	0.347341
Random Forest	0.329536
Naïve Bayes	0.325562

Table 4.5: The result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes

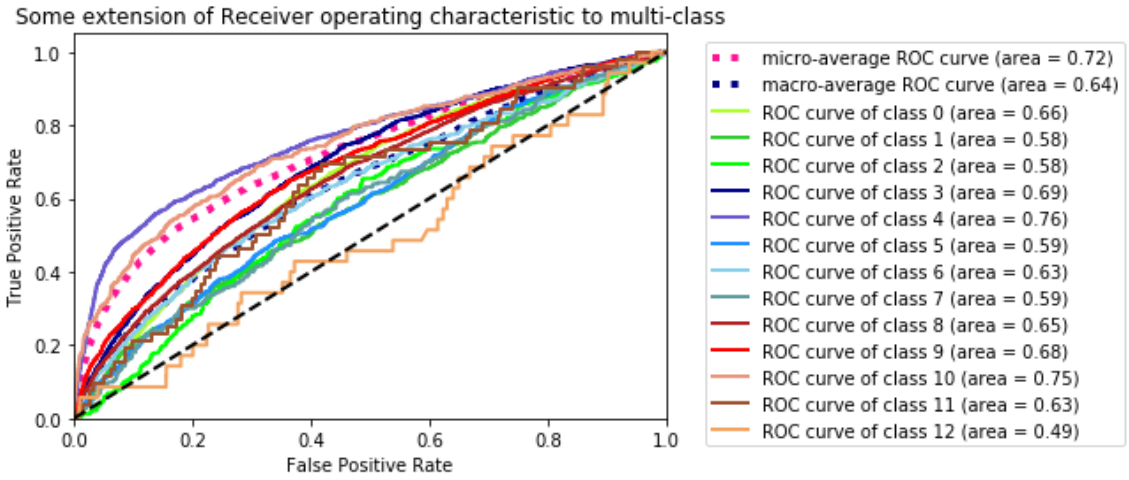


Figure 4.18: The ROC curve of Support Vector Machine

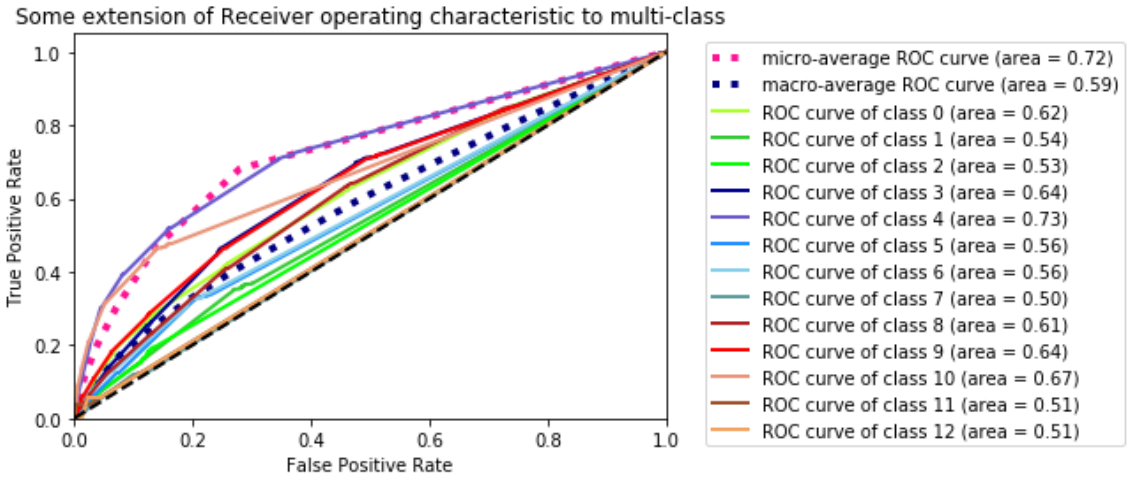


Figure 4.19: The ROC curve of Random Forest

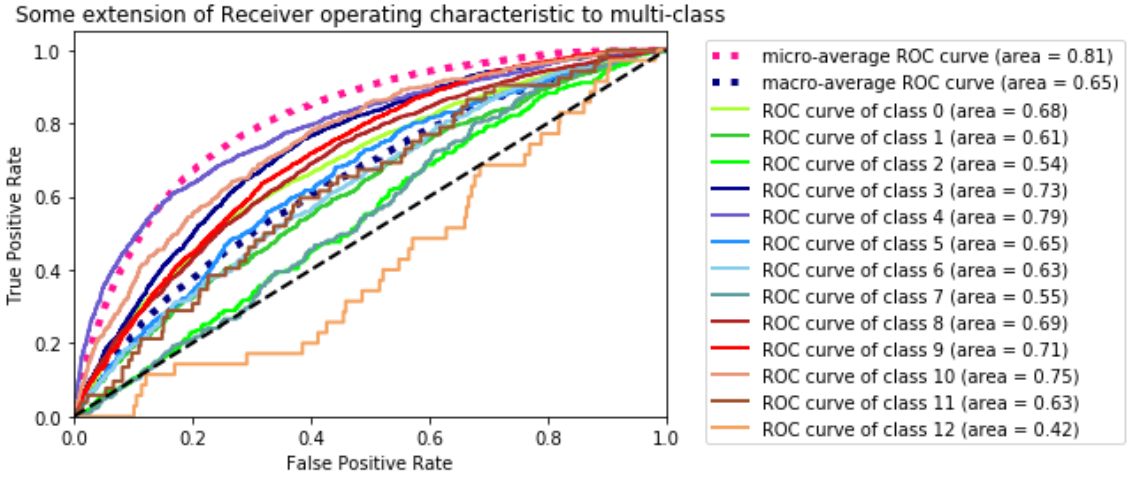


Figure 4.20: The ROC curve of Naïve Bayes

4.5.4 Dataset 3

For dataset 3, the emoticons are replaced with words and text processing carried out. Table 4.5 describes the result of SVM, Random Forest and Naïve Bayes for dataset 3. While Table 4.6 shows the result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes. Figure 4.21, Figure 4.22 and Figure 4.23 shows the ROC curve of different algorithms.

	SVM			Random Forest			Naïve Bayes		
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
anger	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
boredom	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
empty	0.00	0.00	0.00	0.10	0.01	0.02	0.07	0.00	0.01
enthusiasm	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.01
fun	0.15	0.02	0.03	0.16	0.01	0.02	0.11	0.04	0.05
happiness	0.35	0.40	0.37	0.35	0.32	0.33	0.35	0.29	0.32
hate	0.42	0.15	0.22	0.40	0.12	0.18	0.17	0.04	0.07
love	0.49	0.38	0.43	0.46	0.37	0.41	0.44	0.37	0.40
neutral	0.35	0.57	0.44	0.33	0.54	0.41	0.33	0.62	0.43
relief	0.39	0.03	0.06	0.25	0.03	0.05	0.20	0.04	0.07
sadness	0.36	0.21	0.27	0.36	0.16	0.22	0.31	0.23	0.26
surprise	0.27	0.02	0.03	0.30	0.02	0.04	0.18	0.05	0.08
worry	0.33	0.51	0.40	0.30	0.53	0.38	0.34	0.42	0.37

Table 4.6: The result of SVM, Random Forest and Naïve Bayes

Type of Learning Algorithm	Mean Score
Support Vector Machine	0.325562
Random Forest	0.331034
Naïve Bayes	0.326325

Table 4.7: The result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes

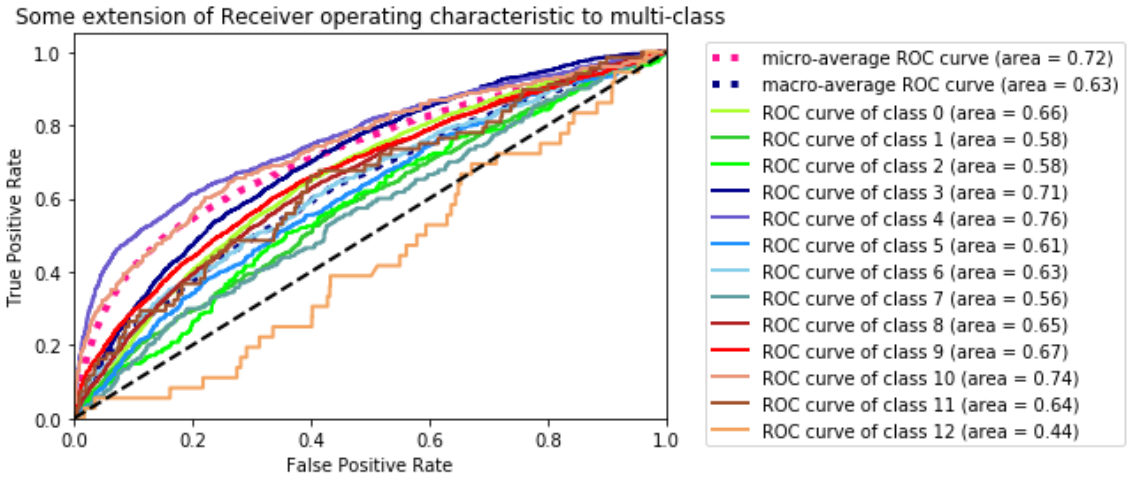


Figure 4.21: The ROC curve of Support Vector Machine

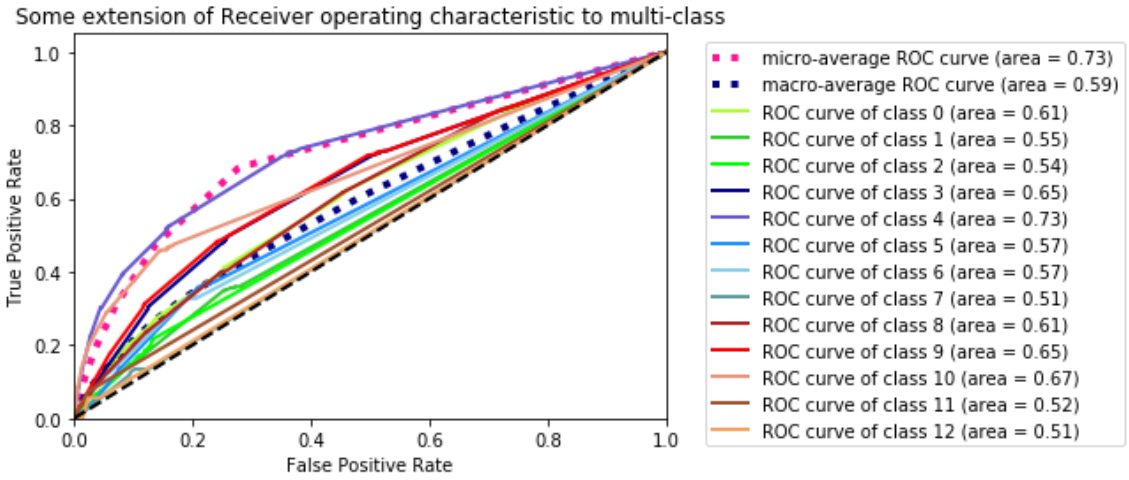


Figure 4.22: The ROC curve of Random Forest

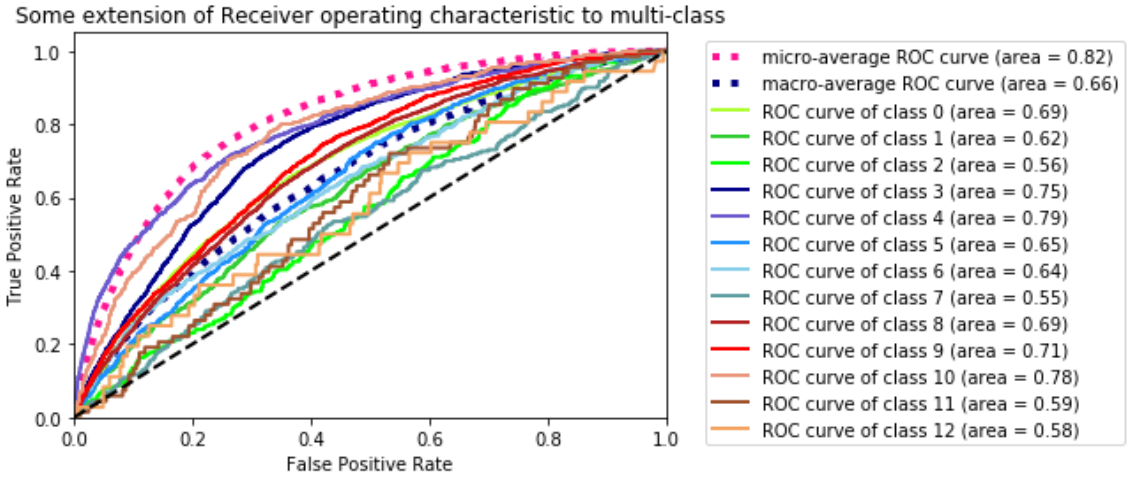


Figure 4.23: The ROC curve of Naïve Bayes

4.5.5 Dataset 4

For dataset 4, the emoticons are replaced with words and text processing carried out. Other than that, the punctuation is removed from the dataset. Table 4.7 describes the result of SVM, Random Forest and Naïve Bayes for dataset 4. While Table 4.8 shows the result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes. Figure 4.24, Figure 4.25 and Figure 4.26 shows the ROC curve of different algorithms.

	SVM			Random Forest			Naïve Bayes		
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
anger	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
boredom	0.00	0.00	0.00	1.00	0.01	0.03	0.00	0.00	0.00
empty	0.00	0.00	0.00	0.09	0.01	0.02	0.00	0.00	0.01
enthusiasm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
fun	0.16	0.01	0.03	0.25	0.02	0.04	0.12	0.04	0.06
happiness	0.34	0.41	0.37	0.36	0.33	0.34	0.35	0.29	0.32
hate	0.45	0.15	0.23	0.38	0.13	0.19	0.20	0.05	0.08
love	0.49	0.38	0.43	0.48	0.38	0.42	0.45	0.38	0.41
neutral	0.35	0.56	0.43	0.33	0.54	0.41	0.33	0.62	0.43
relief	0.34	0.02	0.04	0.25	0.03	0.05	0.17	0.03	0.05
sadness	0.36	0.21	0.26	0.36	0.16	0.22	0.30	0.22	0.26
surprise	0.31	0.02	0.04	0.30	0.02	0.04	0.19	0.05	0.08
worry	0.33	0.50	0.40	0.29	0.52	0.38	0.34	0.42	0.37

Table 4.8: The result of SVM, Random Forest and Naïve Bayes

Type of Learning Algorithm	Mean Score
Support Vector Machine	0.325562
Random Forest	0.331034
Naïve Bayes	0.326325

Table 4.9: The result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes

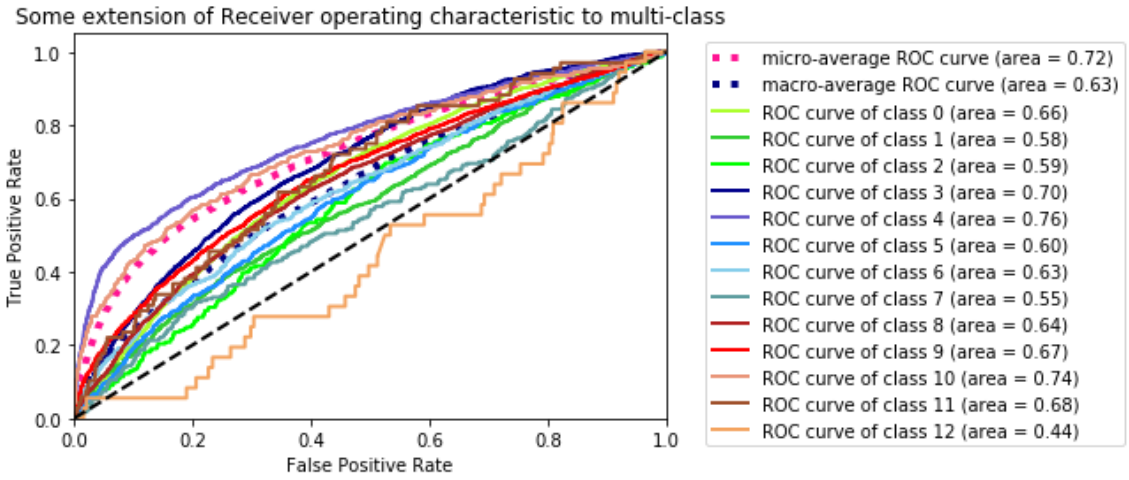


Figure 4.24: The ROC curve of Support Vector Machine

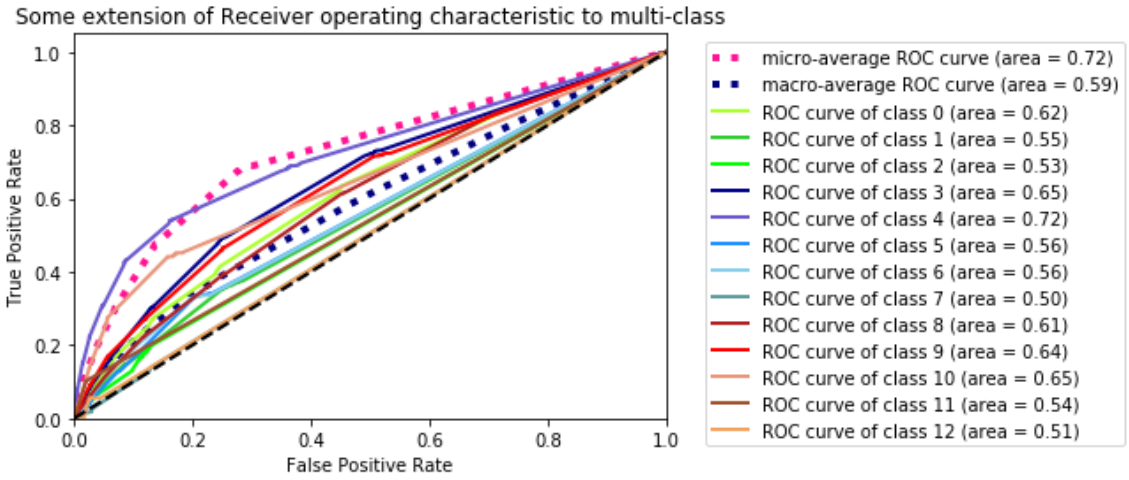


Figure 4.25: The ROC curve of Random Forest

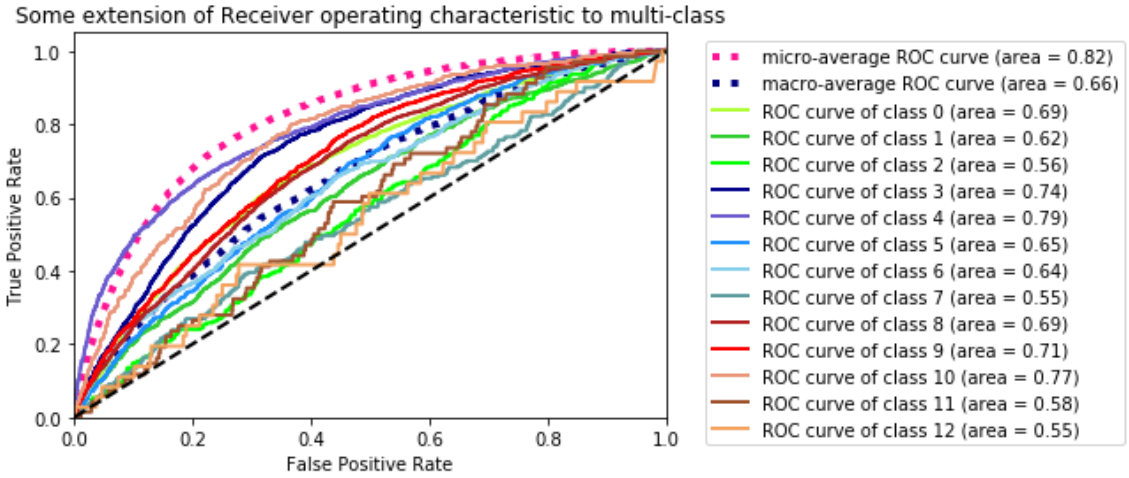


Figure 4.26: The ROC curve of Naïve Bayes

4.5.6 Dataset 5

For dataset 5, the emoticons are replaced with words and text processing carried out. Other than that, the punctuation is removed from the dataset and the abbreviation is reconstructed. Table 4.9 describes the result of SVM, Random Forest and Naïve Bayes for dataset 5. While Table 4.10 shows the result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes. Figure 4.27, Figure 4.28 and Figure 4.29 shows the ROC curve of different algorithms.

	SVM			Random Forest			Naïve Bayes		
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
anger	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
boredom	0.00	0.00	0.00	1.00	0.01	0.03	0.00	0.00	0.00
empty	0.00	0.00	0.00	0.07	0.00	0.01	0.00	0.00	0.00
enthusiasm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fun	0.15	0.01	0.02	0.22	0.02	0.04	0.13	0.05	0.07
happiness	0.34	0.41	0.38	0.34	0.32	0.33	0.35	0.28	0.31
hate	0.44	0.15	0.22	0.42	0.12	0.19	0.20	0.04	0.07
love	0.50	0.38	0.43	0.46	0.36	0.40	0.41	0.34	0.38
neutral	0.35	0.56	0.43	0.33	0.54	0.41	0.33	0.62	0.43
relief	0.38	0.03	0.05	0.20	0.02	0.04	0.14	0.03	0.04
sadness	0.36	0.21	0.26	0.35	0.16	0.22	0.31	0.23	0.27
surprise	0.26	0.02	0.03	0.31	0.03	0.06	0.19	0.06	0.09
worry	0.33	0.51	0.40	0.30	0.53	0.38	0.33	0.40	0.36

Table 4.10: The result of SVM, Random Forest and Naïve Bayes

Type of Learning Algorithm	Mean Score
Support Vector Machine	0.348145
Random Forest	0.325425
Naïve Bayes	0.320112

Table 4.11: The result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes

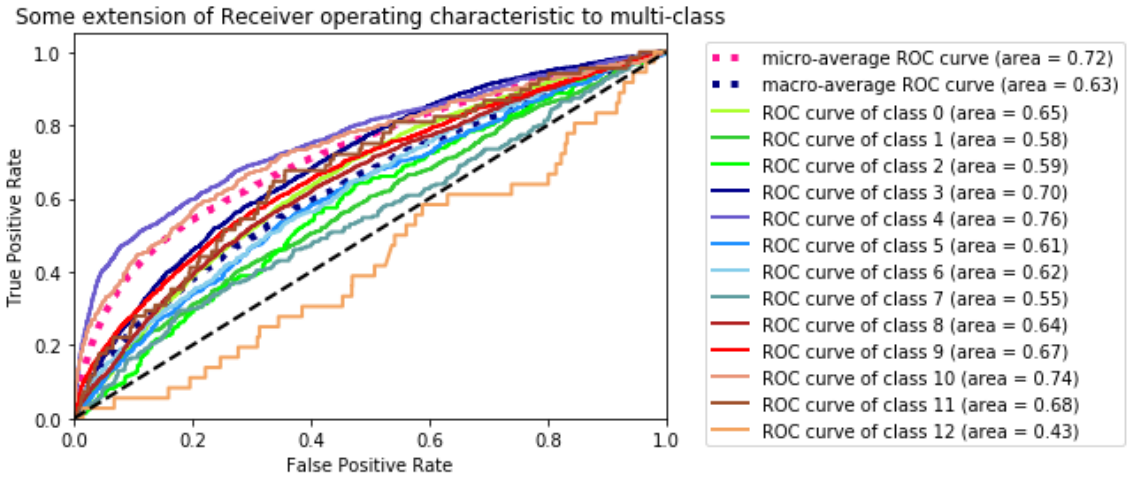


Figure 4.27: The ROC curve of Support Vector Machine

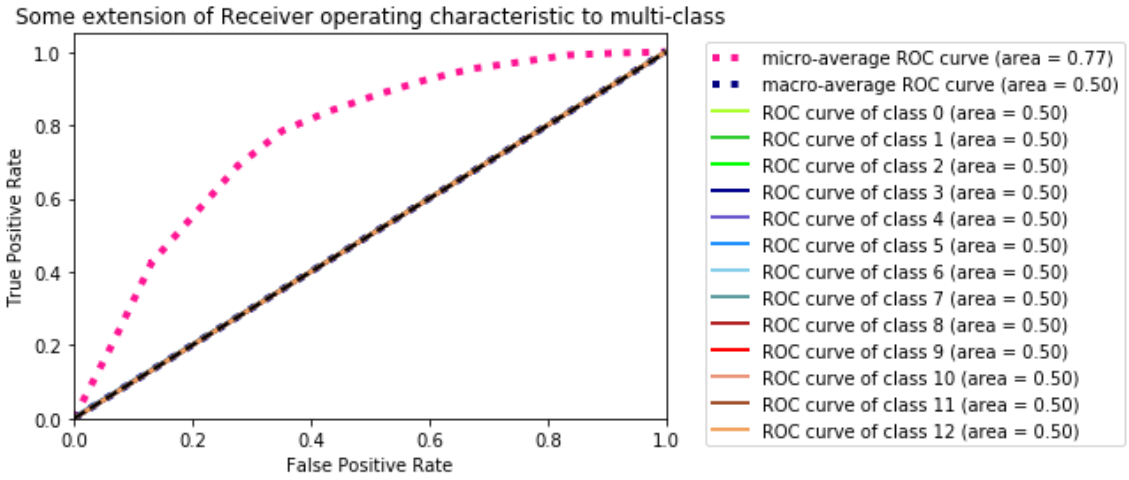


Figure 4.28: The ROC curve of Random Forest

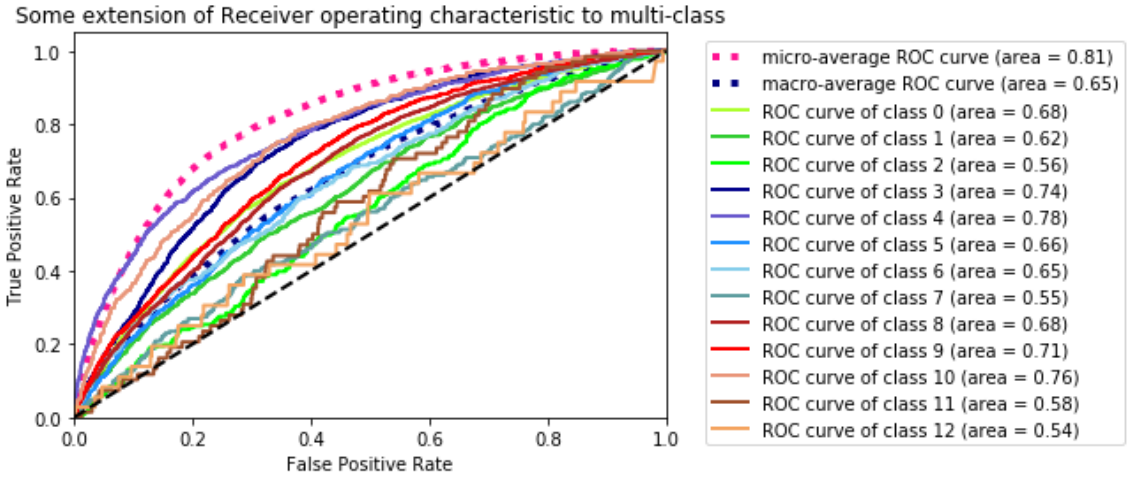


Figure 4.29: The ROC curve of Naïve Bayes

4.5.7 Dataset 6

For dataset 6, the emoticons are replaced with words and text processing carried out. Other than that, the punctuation and stop words are removed from the dataset and the abbreviation is reconstructed. Table 4.11 describes the result of SVM, Random Forest and Naïve Bayes for dataset 6. While Table 4.12 shows the result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes. Figure 4.30, Figure 4.31 and Figure 4.32 shows the ROC curve of different algorithms.

	SVM			Random Forest			Naïve Bayes		
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
anger	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
boredom	0.00	0.00	0.00	0.33	0.01	0.03	0.00	0.00	0.00
empty	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
enthusiasm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fun	0.14	0.02	0.03	0.13	0.02	0.04	0.09	0.03	0.04
happiness	0.33	0.41	0.37	0.31	0.36	0.33	0.33	0.29	0.31
hate	0.43	0.16	0.24	0.30	0.15	0.20	0.22	0.04	0.07
love	0.49	0.38	0.43	0.43	0.38	0.40	0.41	0.34	0.37
neutral	0.34	0.56	0.42	0.34	0.50	0.41	0.33	0.62	0.43
relief	0.33	0.03	0.05	0.20	0.04	0.06	0.15	0.02	0.04
sadness	0.34	0.20	0.25	0.33	0.20	0.25	0.30	0.22	0.25
surprise	0.34	0.02	0.05	0.19	0.04	0.06	0.18	0.05	0.08
worry	0.32	0.47	0.38	0.30	0.47	0.37	0.33	0.41	0.36

Table 4.12: The result of SVM, Random Forest and Naïve Bayes

Type of Learning Algorithm	Mean Score
Support Vector Machine	0.340554
Random Forest	0.324369
Naïve Bayes	0.320663

Table 4.13: The result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes

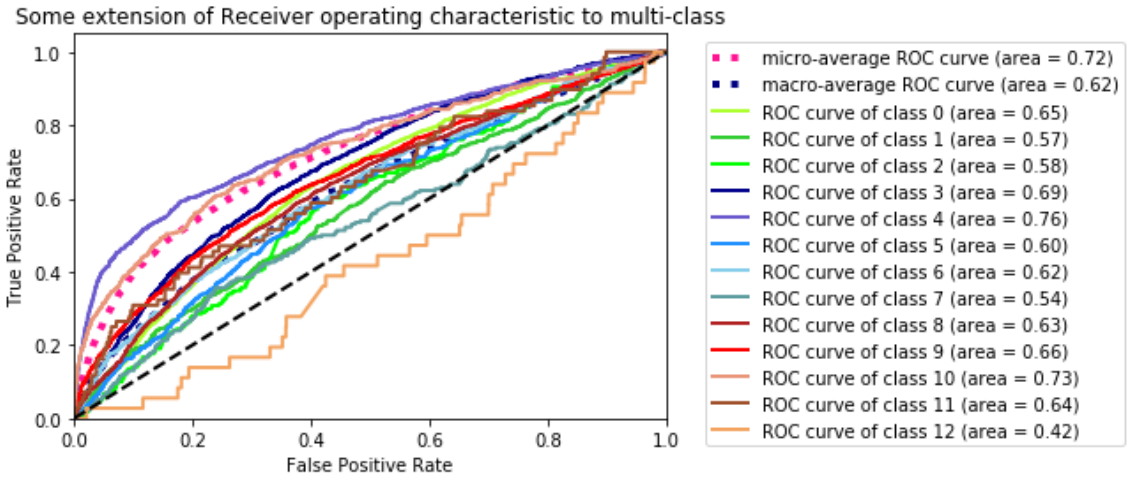


Figure 4.30: The ROC curve of Support Vector Machine

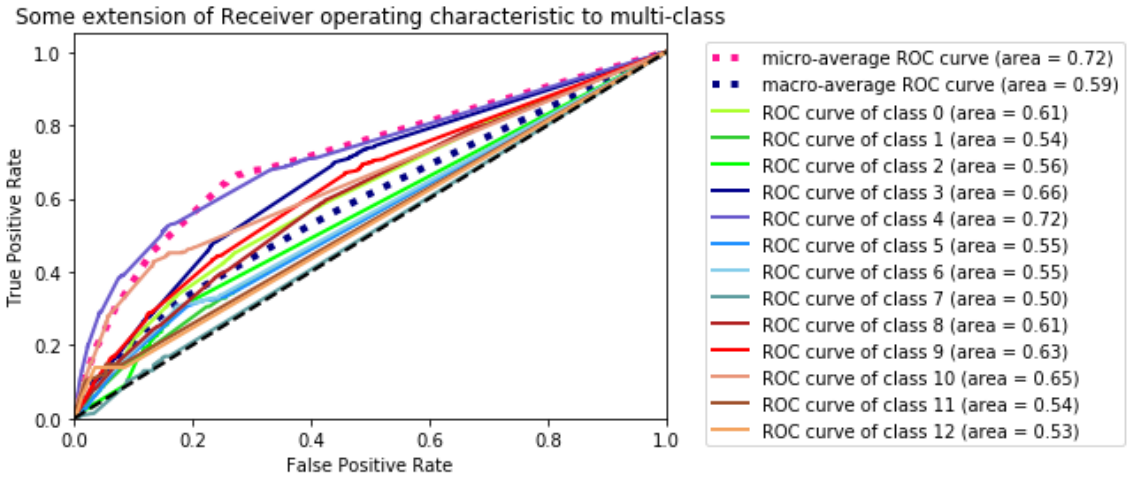


Figure 4.31: The ROC curve of Random Forest

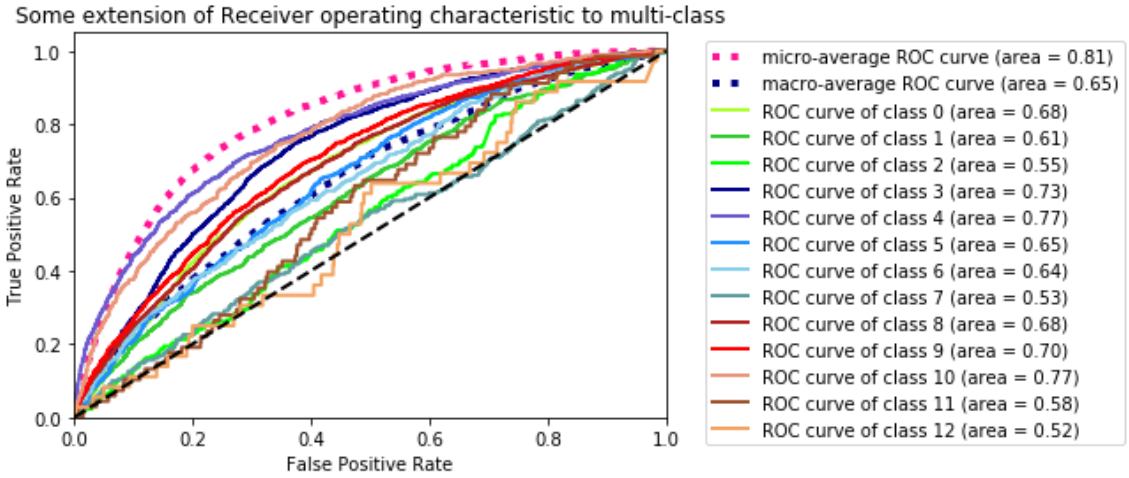


Figure 4.32: The ROC curve of Naïve Bayes

4.5.8 Dataset 7

For dataset 7, the emoticons are replaced with words and text processing carried out. Other than that, the punctuation and stop words are removed from the dataset and the abbreviation is reconstructed. Then the missing letters of the word are corrected. Table 4.13 describes the result of SVM, Random Forest and Naïve Bayes for dataset 7. While Table 4.14 shows the result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes. Figure 4.33, Figure 4.34 and Figure 4.35 shows the ROC curve of different algorithms.

	SVM			Random Forest			Naïve Bayes		
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
anger	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
boredom	0.00	0.00	0.00	0.33	0.01	0.03	0.00	0.00	0.00
empty	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01
enthusiasm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fun	0.17	0.02	0.04	0.11	0.02	0.03	0.10	0.03	0.05
happiness	0.33	0.39	0.36	0.30	0.34	0.32	0.33	0.28	0.30
hate	0.41	0.17	0.24	0.34	0.14	0.20	0.19	0.04	0.06
love	0.50	0.38	0.43	0.43	0.40	0.41	0.40	0.34	0.37
neutral	0.33	0.55	0.41	0.34	0.51	0.41	0.33	0.61	0.43
relief	0.31	0.03	0.05	0.22	0.04	0.07	0.11	0.02	0.04
sadness	0.33	0.20	0.25	0.34	0.18	0.23	0.30	0.21	0.25
surprise	0.29	0.02	0.04	0.13	0.02	0.04	0.17	0.05	0.08
worry	0.32	0.46	0.38	0.30	0.49	0.37	0.33	0.42	0.37

Table 4.14: The result of SVM, Random Forest and Naïve Bayes

Type of Learning Algorithm	Mean Score
Support Vector Machine	0.336493
Random Forest	0.32695
Naïve Bayes	0.316604

Table 4.15: The result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes

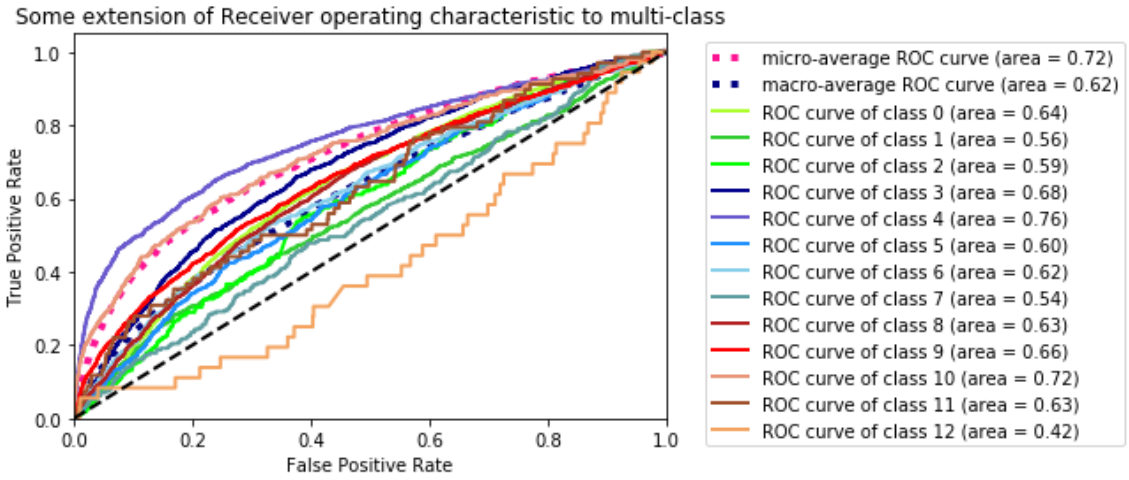


Figure 4.33: The ROC curve of Support Vector Machine

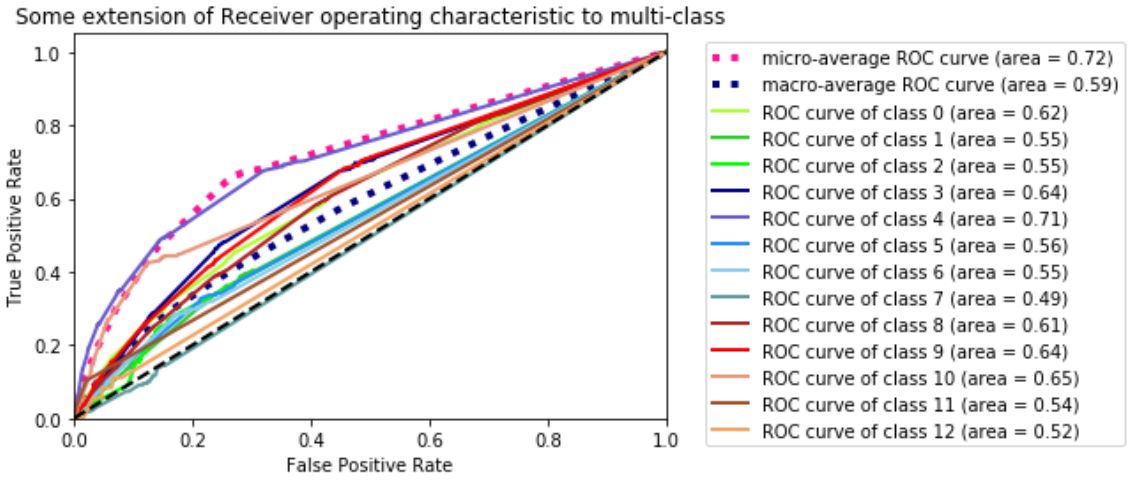


Figure 4.34: The ROC curve of Random Forest

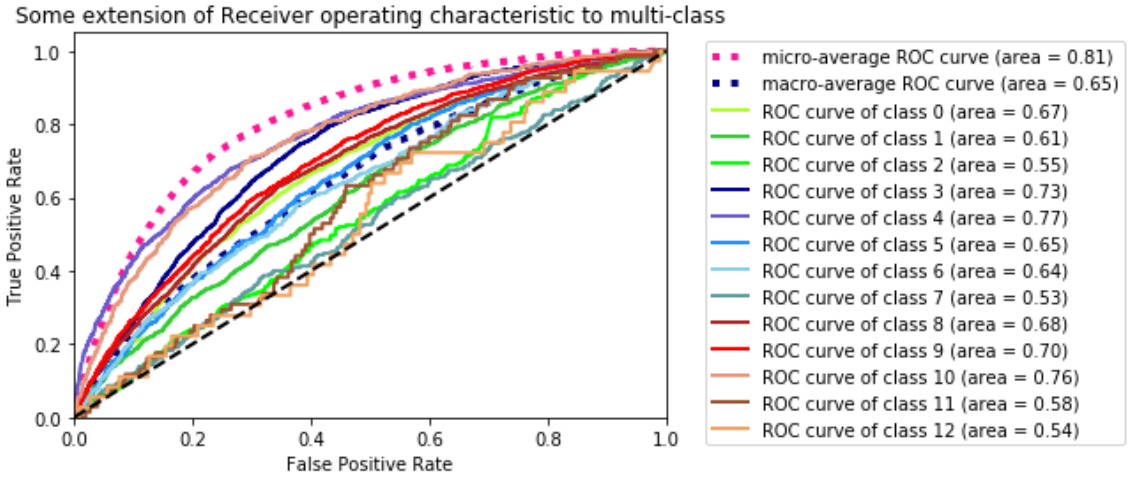


Figure 4.35: The ROC curve of Naïve Bayes

4.5.9 Dataset 8

For dataset 8, the emoticons are replaced with words and text processing carried out. Other than that, the punctuation and stop words are removed from the dataset and the abbreviation is reconstructed. Then the missing letter of the word are corrected and the words are stemmed. Table 4.15 describes the result of SVM, Random Forest and Naïve Bayes for dataset 8. While Table 4.16 shows the result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes. Figure 4.36, Figure 4.37 and Figure 4.38 shows the ROC curve of different algorithms.

	SVM			Random Forest			Naïve Bayes		
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
anger	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
boredom	0.00	0.00	0.00	0.33	0.01	0.03	0.00	0.00	0.00
empty	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.01	0.02
enthusiasm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fun	0.20	0.02	0.04	0.20	0.03	0.05	0.11	0.04	0.06
happiness	0.32	0.39	0.35	0.31	0.35	0.33	0.32	0.28	0.30
hate	0.44	0.17	0.25	0.38	0.16	0.23	0.20	0.06	0.09
love	0.49	0.39	0.43	0.44	0.39	0.41	0.39	0.35	0.37
neutral	0.34	0.56	0.42	0.35	0.51	0.42	0.33	0.59	0.43
relief	0.30	0.03	0.05	0.19	0.02	0.04	0.10	0.02	0.04
sadness	0.35	0.20	0.26	0.35	0.21	0.26	0.31	0.22	0.26
surprise	0.36	0.03	0.06	0.16	0.02	0.04	0.17	0.05	0.08
worry	0.33	0.47	0.39	0.31	0.48	0.38	0.33	0.41	0.36

Table 4.16: The result of SVM, Random Forest and Naïve Bayes

Type of Learning Algorithm	Mean Score
Support Vector Machine	0.339925
Random Forest	0.326826
Naïve Bayes	0.319234

Table 4.17: The result of 10-fold cross-validation on SVM, Random Forest and Naïve Bayes

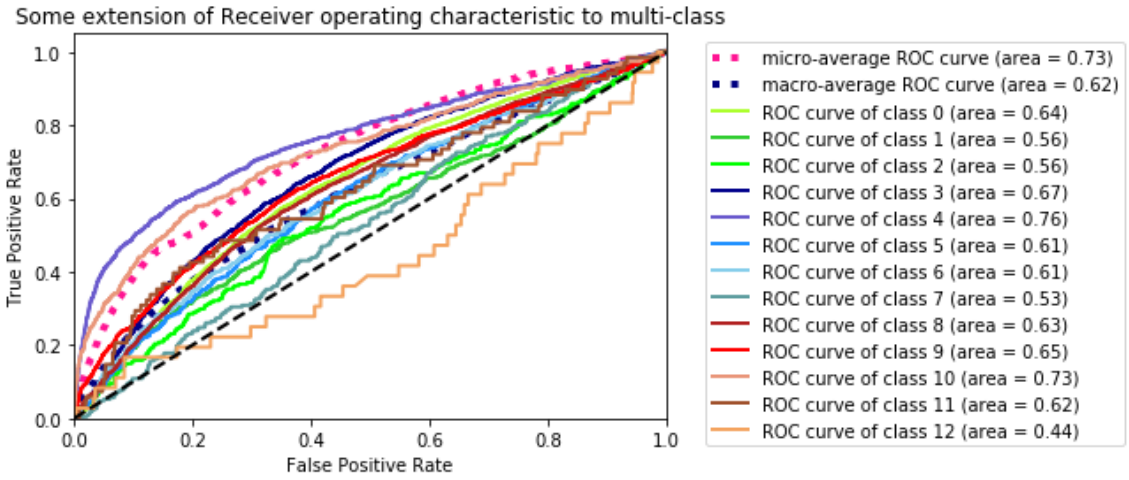


Figure 4.36: The ROC curve of Support Vector Machine

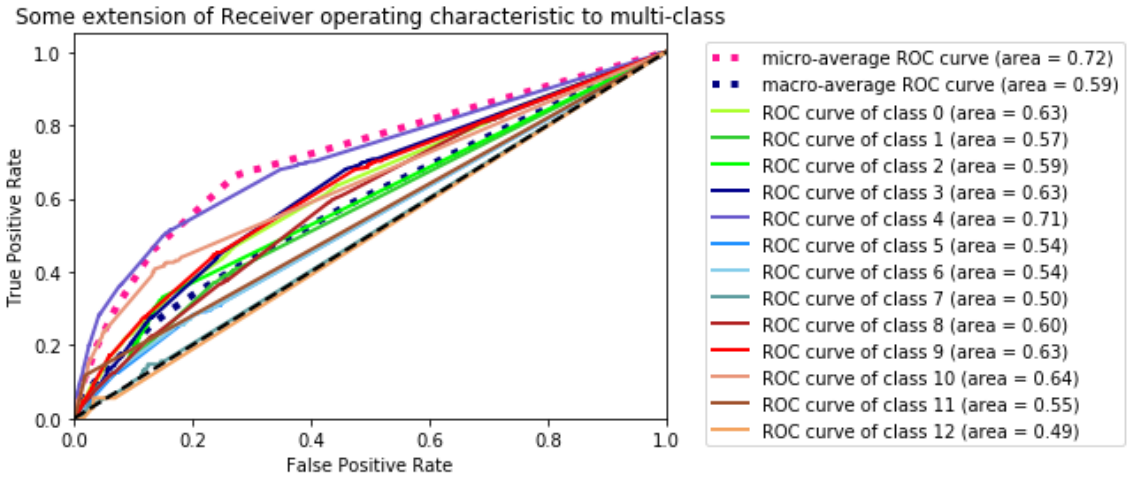


Figure 4.37: The ROC curve of Random Forest

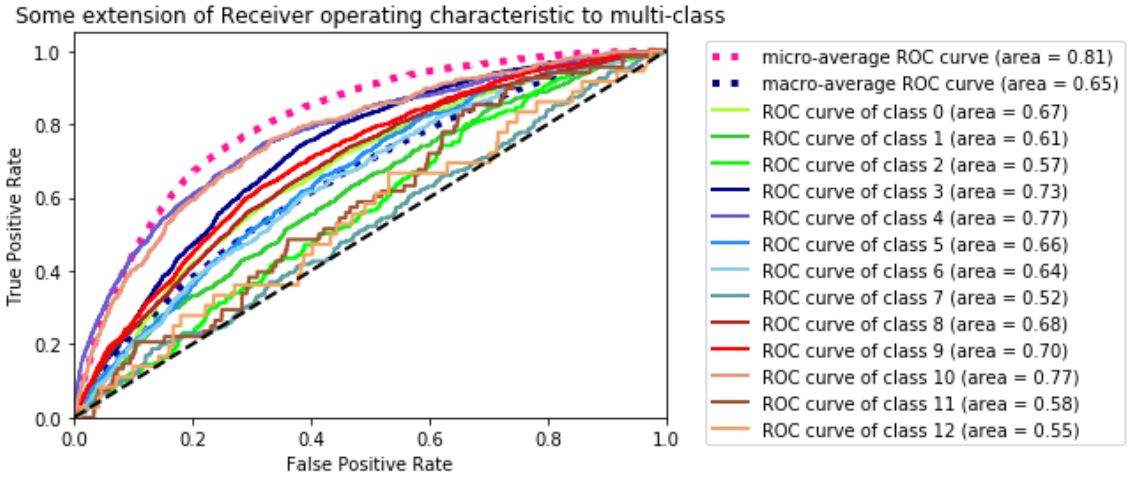


Figure 4.38: The ROC curve of Naïve Bayes

4.5.10 Summary

From the first table of each subsection, it is clearly shown that some of the emotions are zero. The reason behind this result is the size of the dataset of certain emotions are too small. For the 10-fold cross validation, the result of each learning algorithm is showing a low mean score. In order to solve the issue of imbalanced dataset, Area Under the Receiver Operating Characteristic Curve (ROC AUC) is implemented to identify the performance of the imbalanced dataset. From ROC curve, it is showing a higher score compared to previous two tables. From macro-average ROC curve, dataset 3 and 4 which are applied to Naïve Bayes Classifier are showing higher score compare with others. Hence, dataset 3 which applied into Naïve Bayes Classifier will be adopted as the predictive model to predict the real data.

4.6 Predictive Models On Real Data

4.6.1 Introduction

The real data is streaming from Twitter API. The size of the tweets is 100 and it's labelled with the emotions by my friends. The real data are streaming from Twitter API with a few of keywords, which are college, teenager and university. The real data will be tested directly on the predictive model without any data pre-processing steps.

4.6.2 Data Modelling

Figure 4.39 shows the ROC of real data which is predicted by Naïve Bayes.

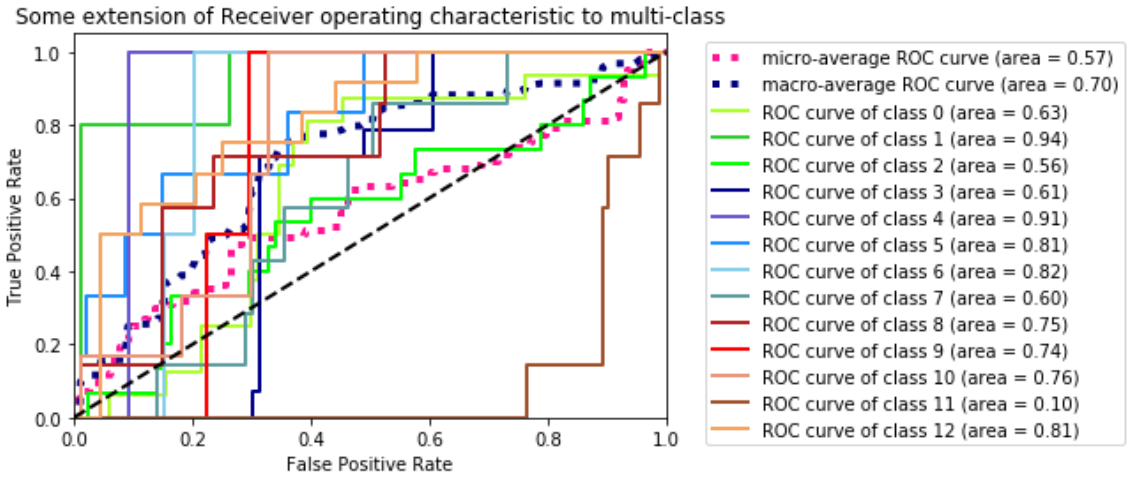


Figure 4.39: The ROC curve of real data

The test dataset is achieved 70% of average AUC by using dataset 3 which is fitted into Naïve Bayes as the predictive model. From Figure 4.39, the ROC curve of class 11 is clearly showing low true positive rate against false positive compare others. It is because the imbalanced of training dataset and domains of emotion words caused the predictive model predicted the wrong emotion.

4.6.3 Confusion Matrix

Figure 4.40 presents the normalized confusion matrix of real data.

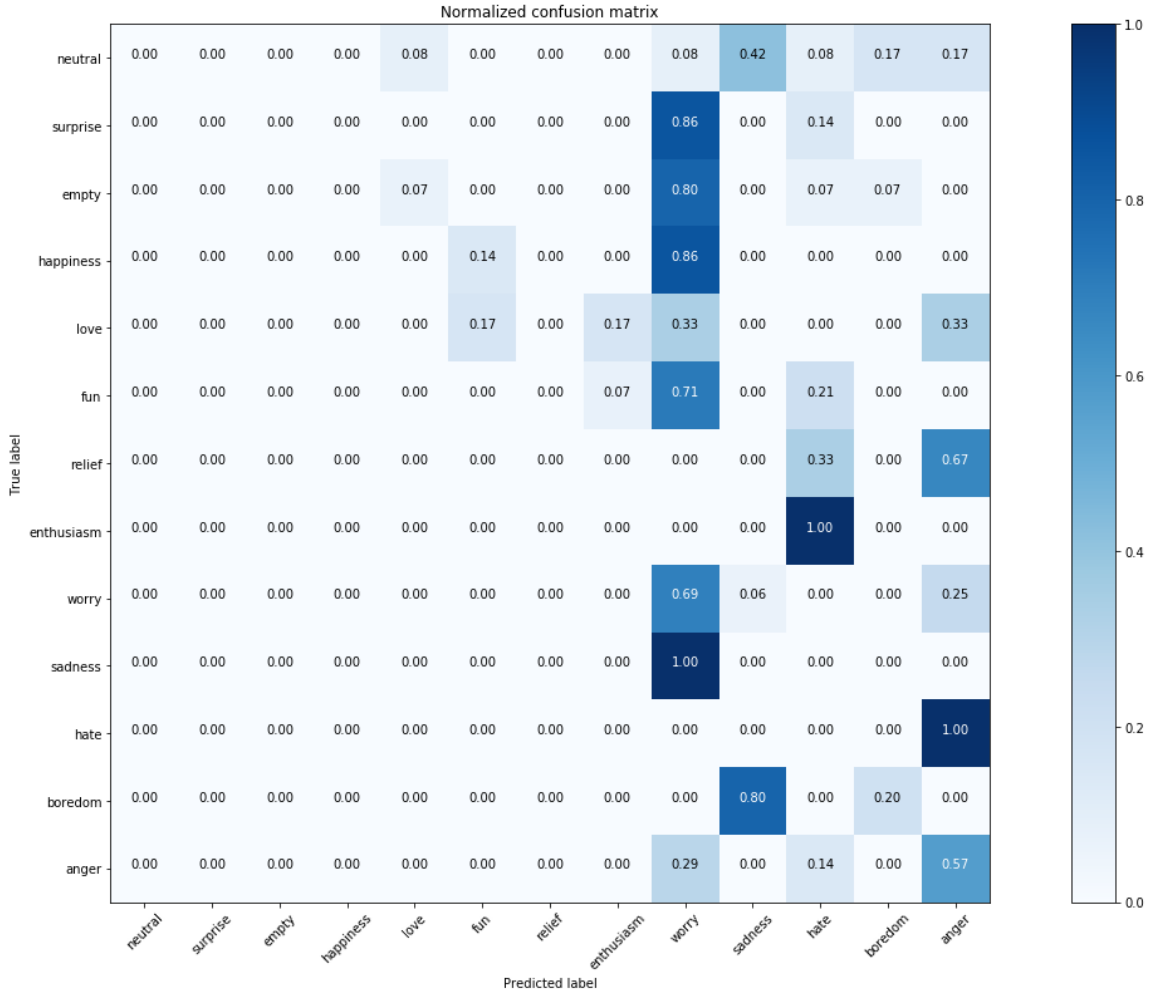


Figure 4.40: The normalized confusion matrix of real data

From Figure 4.40, it is showing that the percentage of boredom emotion which is correctly predicted only have 20 percent. It is because the size of data for the boredom emotion is too small and it caused the percentage of each emotion to be imbalanced, so the predictive

model wrongly predicted the emotion. The following subsection will be discussing the domain words of sadness emotion and boredom emotion in Word Cloud.

4.6.4 Word Cloud

Figure 4.41 shows Word Cloud of boredom emotion of real data. While Figure 4.42 presents Word Cloud of sadness emotion of training dataset.

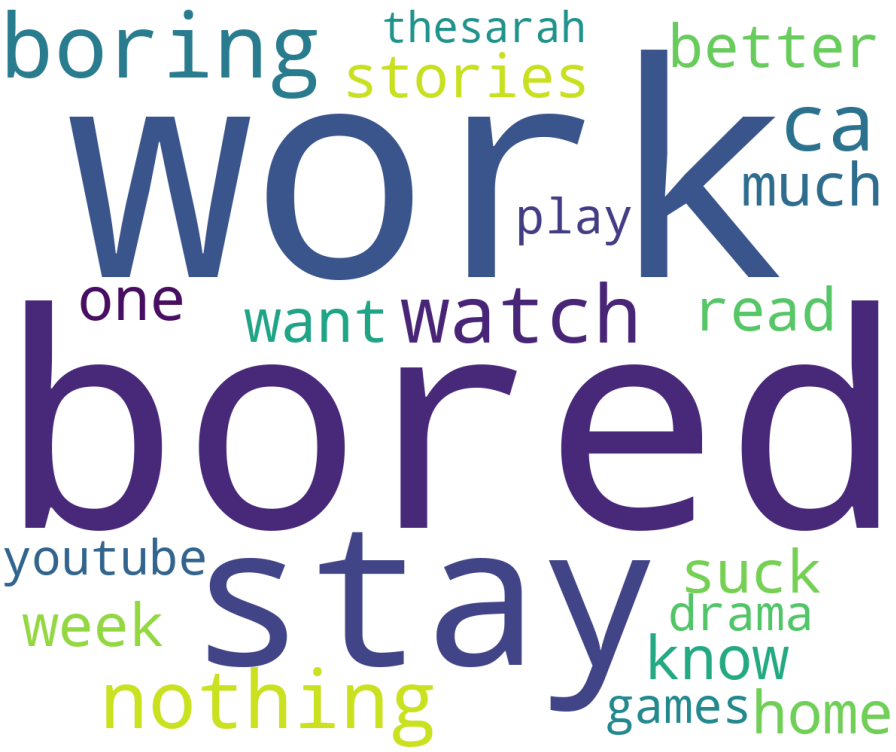


Figure 4.41: Word Cloud of boredom emotion of real data

Figure 4.42: Word Cloud of sadness emotion of training dataset

From Figure 4.41, “bored” and “work” words are the most frequently shown in boredom emotion of real data. In Figure 4.42, it is also showing that “work” and “bored” words are shown in sadness emotion of training dataset. The two most frequent words in boredom are shown in sadness emotion of training dataset, which will cause the prediction part become harder.

4.6.4 Summary

The predictive model has wrongly predicted the boredom emotion due to the smaller size of boredom emotion in training dataset and having a similar domain of emotion words between boredom emotion and sadness emotion. But however, the result of real data achieved 70 percent of average AUC by using training dataset as the predictive model.

Chapter 5 : Conclusion

5.1 Conclusion

In conclusion, this project has achieved the stated objectives as follow:

1. Data pre-processing steps is essential for data modelling. But the excess of data pre-processing steps will cause a worse result.
2. In feature extraction, Term Frequency-Inverse Document Frequency (TF-IDF) are doing a good job in transforming and extracting the data into a relevant matrix. It calculates the frequency of term and inverse document frequency before transforming the data.
3. Naïve Bayes Classifier has the best result among all the classification algorithm. It used the dataset which is going through the minimal of data-preprocessing steps.

5.2 Future Work

The result of the model can be improved if the dataset is implemented upsampling or downsampling or hybrid. Other than that, lexicon-based approaches can be considered for the comparison between machine learning approaches.

References

Abualigah, L. M. (2017). Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. *Expert Systems with Applications 84*, 24-36.

Agrawal, A. &. (2012). Unsupervised emotion detection from text using semantic and syntactic relations. *In Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01 (pp. 346-353)*.

Balahur, A. H. (2011). Detecting implicit expressions of sentiment in text based on commonsense knowledge. *In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (pp. 53-60).*, 60.

Balahur, A. H. (2013). Detecting implicit expressions of affect in text using EmotiNet and its extensions. *Data & Knowledge Engineering, 88*.

Batcha, N. K. (2013). CRF based feature extraction applied for supervised automatic text summarization. *Procedia Technology 11* , 426-436.

Bharti, K. K. (2015). Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Systems with Applications 42.6*, 3105-3114.

Brown, J. (2015, October 14). *Companies using sentiment-analysis software to understand employee concerns*. Retrieved from ciodive: https://www.ciodive.com/news/companies-using-sentiment-analysis-software-to-understand-employee-concerns/407357/

Calvo, R. A. (2013). Emotions in text: dimensional and categorical models. *Computational Intelligence, 29(3), 527-543.*, 527-543.

Cambria, E. &. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine, 9(2), 48-57.*, 48-57.

Canales, L. &.-B. (2014). Emotion Detection from text: A Survey. *Processing in the 5th Information Systems Research Working Days (JISIC 2014), 37.*, 43.

De Choudhury, M. C. (2012). Not all moods are created equal! exploring human emotional states in social media. *In Sixth international AAAI conference on weblogs and social media.*

Deerwester, S. D. (1990). Indexing by latent semantic analysis. 391.

Desmet, B. &. (2013). Emotion detection in suicide notes. *Expert Systems with Applications*.

Fernández, J. G.-B. (2014). A Supervised Approach for Sentiment Analysis using Skipgrams. *In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), number SemEval*.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research 3*, 1289-1305.

Giatsoglou, M. V. (2017). Sentiment analysis leveraging emotions and word embeddings.

Gill, A. J. (2008). Identifying emotional characteristics from short blog texts. *n Proc. 30th Ann. Conf. Cognitive Science Soc., BC Love, K. McRae, and VM Sloutsky, eds (pp. 2237-2242).*

Guru, D. S. (2018). An Alternative Framework for Univariate Filter based Feature Selection for Text Categorization. *Pattern Recognition Letters* .

Hajar, M. (2016). Using YouTube Comments for Text-based Emotion Recognition. *Procedia Computer Science*.

Hasan, M. R. (2014). Emotex: Detecting emotions in twitter messages.

Kolenda, T. L. (2000). Independent components in text. Advances in independent component analysis.

Li, C. H. (2007). Neural network for text classification based on singular value decomposition. *Computer and Information Technology, 2007. CIT 2007. 7th IEEE International Conference on. IEEE*, 47-52.

Lin, K.-C. e. (2016). Feature selection based on an improved cat swarm optimization algorithm for big data classification. *The Journal of Supercomputing 72.8*, 3210-3221.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*.

Liu, L. e. (2005). A comparative study on unsupervised feature selection methods for text clustering. *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*, 597-601.

Mac Kim, S. (2011). Recognising Emotions and Sentiments in Text.

Mahdieh Labani, P. M. (2018). A novel multivariate filter method for feature selection in text classification problems. *Engineering Applications of Artificial Intelligence*, 25-37.

Mohri, M. R. (2012). Foundations of machine learning. *MIT press.*

Newton, Casey. (2017, September 26). *Twitter just doubled the character limit for tweets to 280*. Retrieved from theverge: https://www.theverge.com/2017/9/26/16363912/twitter-character-limit-increase-280-test

Pak, A. &. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *In LREc*.

Pak, A. a. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *LREc. Vol. 10*.

Parth Vora, M. K. (2017). Classification of Tweets based on Emotions using Word Embedding and Random Forest Classifiers. *International Journal of Computer Applications (0975 – 8887)*.

Roberts, K. R. (2012). EmpaTweet: Annotating and Detecting Emotions on Twitter. *In LREC*.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*.

Strapparava, C. &. (2008). Learning to identify emotions in text. *In Proceedings of the 2008 ACM symposium on Applied computing (pp. 1556-1560). ACM.*, 5.

Wang, W. e. (2012). Harnessing twitter" big data" for automatic emotion identification. *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). IEEE*.

Wang, X. &. (2013). Text emotion classification research based on improved latent semantic analysis algorithm. *In Proceedings of the 2nd International Conference on Computer Science nd Electronics Engineering (ICCSEE 2013), number Iccsee (pp. 210-213).*

Weiyuan Li, H. X. (2013). Text-based emotion classification using emotion cause extraction. 8.

Appendix A: Meeting Logs

Appendix B: Supporting Documents