



Intelligent Twitter Agent

A dissertation submitted for the Degree of Master of
Computer Science

U.C.B.I Padmasiri

University of Colombo School of Computing
2018



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: U.C.B.I Padmasiri

Registration Number: 15440519

Index Number: 2015-MCS-051

.....
Signature

.....
Date

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

Mr. U.C.B.I Padmasiri

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

Certified by:

Supervisor Name: Dr. D.A.S.Atukorale

.....
Signature

.....
Date

Abstract

The rapid development of information technology has impacted the human behaviour in different ways. The invention of social media services has connected people from all around the world irrespective of the geographical locations. These platforms provide different ways of connecting with people via different contents. Unique features, popularity, UI/UX, business usage etc. of each social media services compel the users to use more than one social media service. Twitter is one of popular social media service where users connect with messages called tweets which limited to 140 characters. Twitter user's twitter-feed is constructed from the tweets from followers of a specific user. All tweets of followers are shown to the user irrespective of preferences since Twitter does not provide a feature to provide feedback on the preference of tweets.

The Thesis describes a computerized system that gathers Twitter user preference for tweets, analyze the preferences of tweets using Natural language processing and machine learning techniques and generate Twitter content based on the user preference.

When considering the high-level architecture of the system, It includes several modules such as API gateway module that integrate the Twitter API with the system, Core application module developed using python for the analysis purposes and the mobile application that used to gather and display user preferred tweets. The system implementation was carried out by using multiple languages, Java and Python. Python language was very effective when processing text and it includes various libraries and platforms in machine learning and classifications.

The evaluation was carried out for each data pre-processing techniques. Thus the implementation could use the best and efficient data pre-processing steps without losing data. Some of the data pre-processing steps such as removing stop words were not performed because it reduces the accuracy of the classification. The basic requirement was to select the best classification algorithm to perform in a small amount of data. Thus The suitable classifier was also selected after evaluating multiple classifiers. After performing the evaluation, logistic regression classifier was selected for classification. It could gain around 70% of accuracy in the classification.

Acknowledgements

It is my greatest pleasure to remember all those who extended their assistance and support to accomplish this project.

First and foremost, I owe my deepest gratitude to my supervisor, Dr Ajantha Atukorale, for his valuable guidance and advice which continually and convincingly conveyed a spirit of adventure in regards to the research. This project would not have been possible without him because he has always shown the way of research and helped me to overcome most of the obstacles that I had encountered.

I would also thank my parents who encouraged me to complete this research. Finally, I thank all of my colleagues, who supported me in many ways during the completion of the research.

Contents

Declaration	i
Abstract	iii
Acknowledgements	v
List of Figures	xi
List of Tables	xiii
List of Abbreviations and Acronyms	xv
1 Introduction	1
1.1 Introduction	1
1.1.1 General Tweets	2
1.1.2 Mentions	2
1.1.3 Replies	2
1.1.4 Non Following Tweets	3
1.2 Problem Definition	3
1.3 Motivation	8
1.4 Aim and Objective	8
1.5 Scope	9
1.6 Thesis Outline	10
2 Background and Related Work	11
2.1 Background	11
2.2 Natural Language Processing	11
2.2.1 Development of NLP	11
2.2.2 NLP Process	12

	Data Pre-processing	12
	Remove Stop Words	12
	N-Gram	14
	Stemming or Lemmatization	14
2.3	Machine Learning	16
2.3.1	Bag of Words	16
2.3.2	Count Vectorizer	16
2.3.3	TFIDF	16
2.4	Tweet Content Analysis	17
2.4.1	Hashtags	18
2.4.2	Emoticons	19
2.4.3	Images	20
2.5	Spam and Irony Detection	20
2.5.1	Spam Detection	20
2.5.2	Irony Detection	21
2.6	Summary	22
3	Analysis and Design	23
3.1	Introduction	23
3.2	System Analysis and Feasibility	23
3.2.1	Integrating with Twitter API	23
3.2.2	Feasibility	24
3.2.3	Twitter Data	25
3.3	System Design	25
4	Implementation	29
4.1	Introduction	29
4.2	Technologies, Libraries and Frameworks	29
4.2.1	Programming Languages	29
4.2.2	Libries and Frameworks	29
4.3	System Implementation	30

4.3.1	Mobile Application Implementation	30
4.3.2	Application Core module implementation	31
	Data Pre-processing	31
5	Evaluation	35
5.1	Introduction	35
5.2	Data and Resources	35
5.3	Data Pre-processing and Evaluation	36
	Define Baseline	36
	Removing Stop Words	37
	Expanding Contractions	38
	Converting Emoj in to Text	40
	Converting Emoticons into Text	41
	lemmatize	42
	Remove Numeric Characters	43
5.4	Data Pre Processing and Evaluation- Small Dataset	44
	Define Baseline	44
	Removing Stop Words	45
	Expanding Contractions	46
	Converting Emoji in to Text	46
	Converting Emoticons into Text	47
	lemmatize	48
	Remove Numeric Characters	49
5.5	Comparison of Classifiers	50
5.6	Improving Classifier Accuracy using TFIDF vectorization	52
5.7	Selecting N-Gram Value	53
5.8	Parameter Tuning in Classifier	54
6	Conclusion and Future Work	57
6.1	Conclusion	57
6.2	Future Work	58

Bibliography	59
A Code Samples	65
A.1 Twitter API Timeline JSON Response	65
B Data Cleanning and pre-processing Evaluation	69
B.1 Accuracy on control datasets	69
B.2 Stopwords removed datasets accuracy vs Number of features	69
B.3 Contractions expanded datasets accuracy vs Number of features	69
B.4 Emoji converted datasets accuracy vs Number of features	69
B.5 Emoticons converted datasets accuracy vs Number of features	69
B.6 Lemmatized dataset accuracy vs Number of features	69
B.7 Numeric characters removed dataset accuracy vs Number of features	69
B.8 TFIDF vectorizer classification accuracy	70
B.9 Count vectorizer classification accuracy	70
B.10 N-Gram accuracy calculation	70
B.11 Parameter tunning	70

List of Figures

1.1	General tweet from Oracle Twitter account	3
1.2	Mention type of tweet posted by Oracle mentioning two another Twitter accounts	4
1.3	Reply type of tweet from a sender to another Twitter user's tweet	5
1.4	Forbes Tech news tweets	6
1.5	National Geographic tweets	7
2.1	NLP Process	13
2.2	POS Trigger output - Sentence one	15
2.3	POS Trigger output - Sentence two	15
2.4	Tweet Content	18
3.1	System high level architecture	26
3.2	System high level architecture decomposed	27
4.1	Authorization Request	30
5.1	Control dataset accuracy vs Stop words removed dataset accuracy	38
5.2	Baseline dataset accuracy vs Contractions expanded dataset accuracy	39
5.3	Baseline dataset accuracy vs Emoji converted dataset accuracy	41
5.4	Baseline dataset accuracy vs Emoticons converted dataset accuracy	42
5.5	Baseline dataset accuracy vs Lemmatized dataset accuracy	43
5.6	Baseline dataset accuracy vs Numeric characters removed dataset accuracy . .	44
5.7	Baseline datasets average accuracy vs Stop words removed datasets average accuracy	45
5.8	Baseline datasets average accuracy vs Contractions expanded datasets average accuracy	46
5.9	Baseline datasets average accuracy vs Emoji converted dataset average accuracy	47

5.10	Baseline datasets average accuracy vs Emoticons converted dataset average accuracy	48
5.11	Baseline dataset average accuracy vs Lemmatized dataset average accuracy . .	49
5.12	Baseline dataset average accuracy vs Numeric characters removed dataset average accuracy	50
5.13	Classification accuracy summary	52
5.14	TFIDF vectorization average accuracy vs Count Vectorizer average accuracy . .	53
5.15	N-Gram average accuracy	54
5.16	Classifier accuracy vs <i>Cvalue</i>	55

List of Tables

2.1	Stemming or Lemmatization Output	14
2.2	Count Vectorizer in a corpus of three sentences	16
4.1	Cleaned Twitter data	33
5.1	Labeled Twitter data	35
5.2	Baseline Dataset Accuracy vs Number of features	37
5.3	Stopwords removed dataset accuracy vs Number of features	38
5.4	Contractions expanded dataset accuracy vs Number of features	39
5.5	Emoj converted dataset mean accuracy vs Number of features	40
5.6	Emoticons converted dataset accuracy vs Number of features	42
5.7	Lemmatized dataset accuracy vs Number of features	43
5.8	Numeric characters removed dataset accuracy vs Number of features	44
5.9	Comparison of classifiers accuracy	51
B.1	Accuracy on control datasets	71
B.2	Stopwords removed datasets accuracy vs Number of features	72
B.3	Contractions expaned datasets accuracy vs Number of features	73
B.4	Emoj converted datasets accuracy vs Number of features	74
B.5	Emoticons converted datasets accuracy vs Number of features	75
B.6	Lemmatized datasets accuracy vs Number of features	76
B.7	Numeric characters removed dataset accuracy vs Number of features	77
B.8	TFIDF vectorizer classification accuracy	78
B.9	Count vectorizer classification accuracy	79
B.10	n-gram (1,1)	80
B.11	n-gram (1,2)	81
B.12	n-gram (1,3)	82
B.13	c value vs n-gram	83

List of Abbreviations and Acronyms

API	Application Programming Interface
HTML	Hypertext Markup Language
IOS	iPhone Operating System
ITA	Intelligent Twitter Agent
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
POS	Part-Of-Speech
TFIDF	Term Frequency-Inverse Document Frequency
URL	Uniform Resource Locator

Chapter 1

Introduction

1.1 Introduction

Social media services play a vital role in the current technological world. There are numerous social media platforms developed for various kind of content sharing amongst people who interact with each other. Twitter is one of online social media service. Twitter is a service for people to communicate and stay connected through the exchange of quick, frequent messages called "tweets". Twitter is not only a social media service but also a microblogging service which let users create and read tweets. These tweets are restricted to 140 characters [1]. In order to create tweets, users have to register for the Twitter service, but unregistered users can read tweets posted by registered users.

Twitter was created in the year 2006 by four friends named Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams [2]. Twitter started as a platform for sending short updates to a network of friends through text messages, standard text message size was 160. Twitter capped each tweet at 140 characters, leaving other space for to usernames etc. Gradually Twitter service was updated to add links which include content previews, pictures, and videos. Therefore, users can see a preview from an external site before clicking on a tweet.

In the year 2012 Twitter reached the milestone of 100 million monthly active users [3]. Twitter released that it has 320 million active users excluding accounts that have not been active for long time periods. According to Alhabash and Ma [4], Twitter service is visited by around 1 billion unique visits per month in the year 2016.

When a Twitter user opens the Twitter application in mobile or web browser, the Twitter feed is populated with tweets from accounts that the user is following. There are several types of tweets from following accounts that are available in users Twitter feed. A Twitter user can be followed by other Twitter users. In this scenario, the user who is been followed by others

become the sender. Sender posts tweets. Users who follow a certain sender becomes recipients and in their feeds, they will see the tweets posted by senders.

Tweets that are populated in users Twitter feed are categorized into different categories as following [5].

1.1.1 General Tweets

A general tweet is the basic type of tweet, it's a message posted to Twitter containing text, photos, a GIF, and/or video. It first appears in the sender's profile which is the creator's feed. Then it will appear in the recipient's feeds who follows the sender. If you are interested to know about trends of the technology, you can follow a user who tweets regarding technology. And you will see all his tweets. Assume if certain Twitter users are interested in a particular subject such as 'Java technologies', they can follow Twitter users who generally post 'Java' related tweets. Then the users feed will be populated with general tweets which are related to 'Java Technology'. Figure 1.1 is an example of a general tweet.

1.1.2 Mentions

Mentions are tweets that contain another user's username. Mentions are done by adding '@' symbol preceding to the username of the twitter user. For example: "Hello @Oracle!" Mention type of tweets appears on the Senders public tweet feed. When a Twitter user (Sender) mentions another Twitter user, the mentioned user will receive a notification regarding the mentioning. Also, this tweet will appear in other Twitter users who follow sender account. Figure 1.2 is an example of mentions type tweet.

1.1.3 Replies

When a Twitter user replies to another user's tweet using reply option, these tweets are identified as replies. Reply tweets also appear in senders home and will be available in followers' feeds who follow the sender. Also, a notification will be sent to the Twitter user who's the tweet has been replied by a sender. Figure 1.3 is an example of reply type tweet.



FIGURE 1.1: General tweet from Oracle Twitter account

1.1.4 Non Following Tweets

There are random tweets feeds from Twitter. You will get random Tweets regarding almost everything happening in the world. Twitter identify a tweet, an account to follow, or other content that's popular or relevant, they may add it to your timeline. This means you will sometimes see tweets from accounts you don't follow. Twitter select each tweet using a variety of signals, including how popular it is and how people in your network are interacting with it [6].

1.2 Problem Definition

Twitter users follow other Twitter accounts, personalities based on the different areas that the user is interested in. Once the user follows another Twitter account, the user's feed is populated with followed Twitter accounts tweets. And this followed Twitter account will post many tweets from different broader subjects. Twitter does not provide an option to filter and read preferred

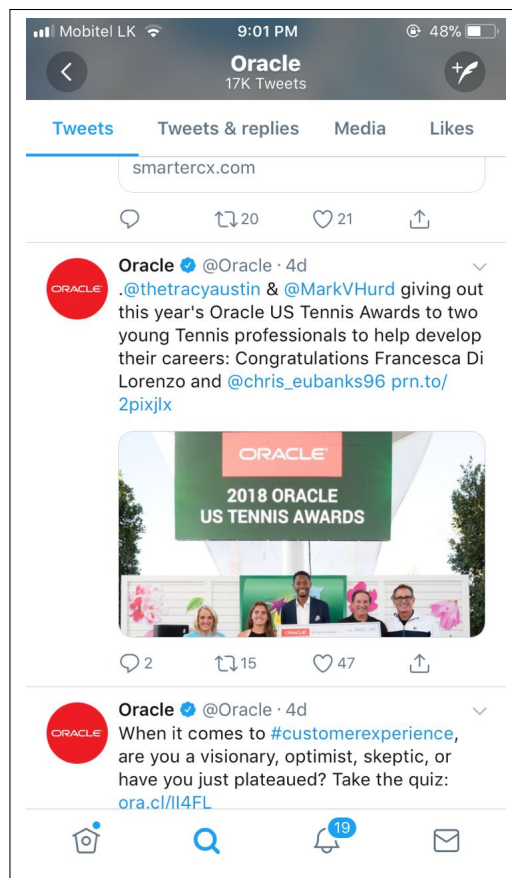


FIGURE 1.2: Mention type of tweet posted by Oracle mentioning two another Twitter accounts

types of Tweets from followed users. Hence the users have to go through all the unimportant tweets that are populated in their feed and it is a waste of time.

If you are interested to get updates about latest technologies you can follow a user who frequently tweets about latest technology achievements. That user may tweet about mechanical engineering achievements, Spacecraft, Satellites and Space stations. If your only interest in latest achievements on Satellites technology it will be a waste of time to go through all unnecessary tweets. When considering a large number of tweets appear from the user from different subjects, it is hard for you to manage time for reading all of them and filter the ones with Space technologies.

Following are some of the practical scenarios where this problem is presented.

Forbes Tech news is a famous Twitter who tweets about latest technology trends and technology updates. Figure 1.4 in page number 6 contains four images of their recent tweets.



FIGURE 1.3: Reply type of tweet from a sender to another Twitter user's tweet

When considering those four tweets I am very much interest in the tweet about the Cryptocurrency, but I have no option to provide my preference on the Cryptocurrency related tweets and see more tweets related Cryptocurrency on my Twitter feed from Forbs Tech News.

National Geographic is followed by 22.3 million other Twitter users. They post tweets related to Environment, Wildlife, Space, Human Nature, Bio-Technology, Climate and Natural Disasters etc. Figure 1.5 in page number 7 contains four images of their recent tweets. A user is very much interested in Climate Change related tweets, but the user has to go through the all tweets in own Twitter feed or National Geographic Twitter home feed in order to search for tweets regarding Climate Change. If there was an option to provide the preference of the user to Twitter, Twitter can filter the tweets with Climate Change subject and populate in specific Twitter user feed.

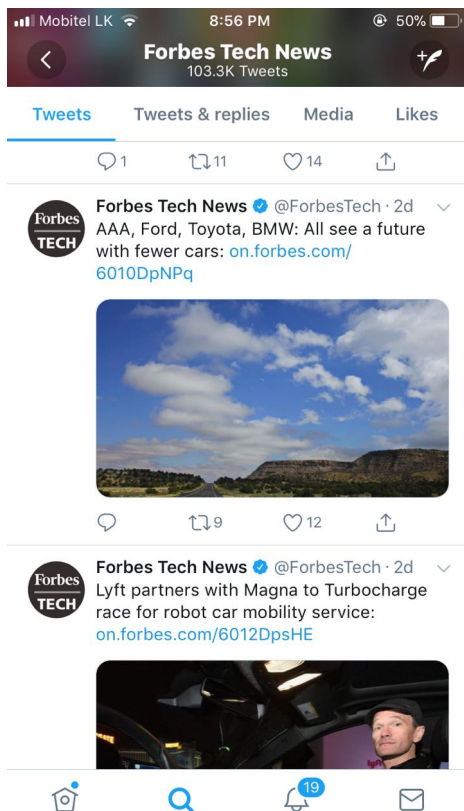
When the users Twitter feed is filled with unwanted and non-favourite tweets from followers, the Twitter user may consider unfollowing certain users. According to Statistics [7] National



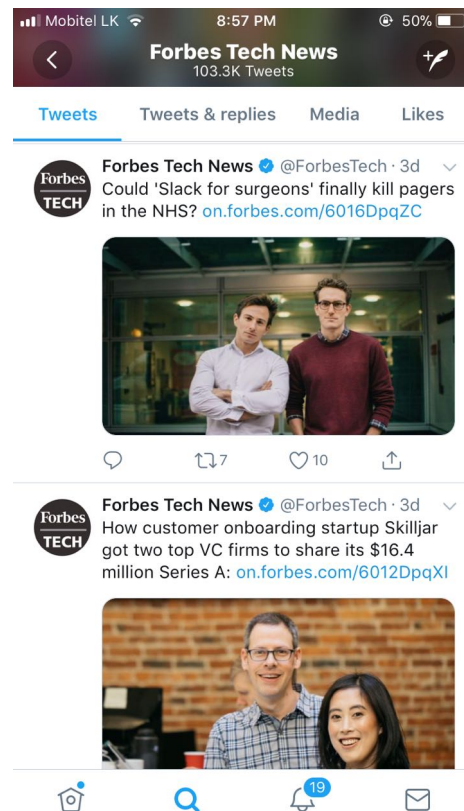
(A) fig 1



(B) fig 2

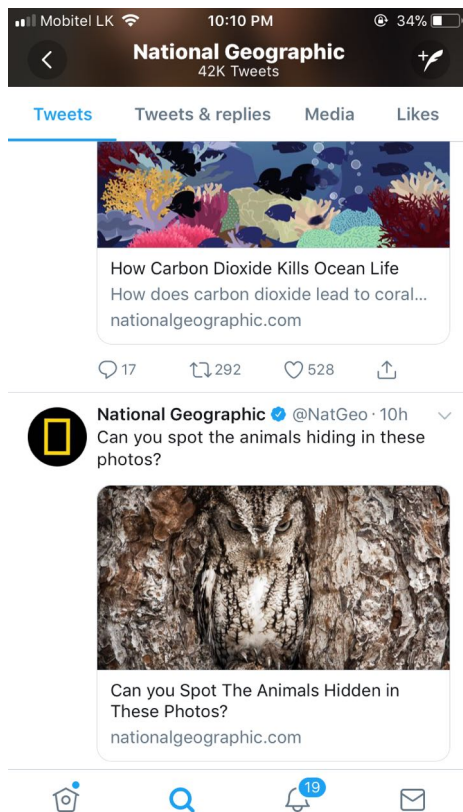


(c) fig 3

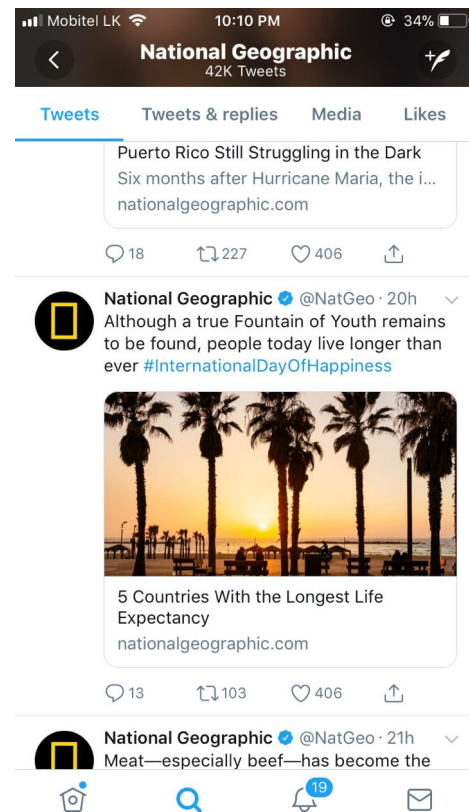


(D) fig 4

FIGURE 1.4: Forbes Tech news tweets



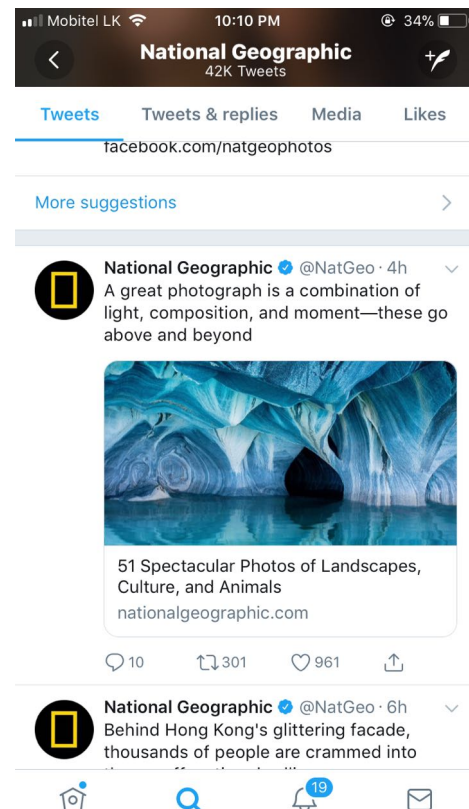
(A) fig 1



(B) fig 2



(c) fig 3



(d) fig 4

FIGURE 1.5: National Geographic tweets

Geographic posts around 20+ tweets per day and Forbs Tech News posts around 22+ tweets per day [8]. Among these, there can be one or two preferred subject's related tweets. But the users Twitter feed is filled with all tweets posted by both accounts. In this scenario, a Twitter user may contemplate unfollowing both users since the user mostly sees unwanted tweets from these tweeters.

If the user has a way of providing Positive or a negative feedback for the tweets in their Twitter feed and gets more preferred types based on the preferred subjects, they can manage time used in Twitter social media and have a better Twitter experience.

1.3 Motivation

The current world is moving fast and the time is the most precious thing that humans have. With the short time, they have users intend to get more and more relevant news and information. Acting as a one of the major news and information provider, Twitter should only provide user preferred types of tweets.

The research is focused on developing an "Intelligent Twitter Agent" that can identify the user preference and provide preferred tweets for the user. "Intelligent Twitter Agent" is a system that capable of gathering user feedback on tweets, analyzing the provided feedback using machine learning technologies, identifying user preference for tweets and populate the preferred tweets in the users Twitter feed.

1.4 Aim and Objective

The objective of the research is to analyze user's feedback for the tweets using efficient natural language processing algorithms and identify the user preference. By identifying user preferred types of tweets, it will be able to provide more and more preferred type of tweets to the user.

When you logged in to Twitter you may see both important and not important tweets on your Twitter feed. There can be spams, tweets that for advertisements, promotions or maybe tweets containing totally uninterested facts. There is no way that you can give a feedback for a particular tweet regarding your preference. There is no way to tell the application these are the type of tweets that I am interested in.

If we can get positive or a negative feedback from the user, we can analyze the feedback that with the content of the tweet. This analysis is done using effective algorithms by feeding the feedback. From the analysis, we would be able to identify what are the type of tweets the user interest on. By analyzing the tweet content and the user feedback, we will be able to filter only the necessary, preferred and important tweets to the user.

Users will be able to follow more and more Twitter users but will only receive his or her preferred types of tweets. For an example, the user can follow several tweeters who tweet about programming languages. If he is interested only about JAVA programming language, he will receive java related tweets only. All the other irrelevant tweets will be filtered out.

1.5 Scope

The scope of this project is to create an application to identify user preferred types of tweets. The application will gather feedback from the users about the tweets and analyze those feedbacks. From the analysis, it can identify the preference of the user. Afterwards, the application will display tweets according to the user's preference.

Users can give positive or negative feedback for tweets, for an example thumbs-up or a thumbs-down. According to what I have mentioned in the project problem section user can mark thumbs-down when he or she receives tweets regarding "Mechanical Engineering Achievements". After gathering that feedback from the user, the application will analyze and determine that user does not prefer tweets regarding "Mechanical Engineering achievements". Next time user will receive less amount of tweets regarding "Mechanical Engineering achievements". The more feedback application gets will increase the accuracy of the analysis.

There can be several approaches to determine the importance of the tweet. The knowledge that gathered with above analysis process will be used to provide only the Important, necessary tweets to the particular users. This will lead the user to have a better twitter experience and effective time management with the Twitter.

1.6 Thesis Outline

Chapter 2 Contains the literature review of this project. It will describe the related research efforts which have taken place earlier.

Chapter 3 Contains the high-level system architecture, the feasibility of research development, constraints of the research, steps taken to address those constraints of the project in order to complete the research. Also, the system components are described in detail in this section.

Chapter 4 Describes the research implementation activities. Used programming languages, methods of data gathering, data cleansing approaches, classifications used to determine the preference using machine learning and mobile application development to populate preferred tweets based on the preferences are included in this chapter.

Chapter 5 Contains the evaluation of the research implementation. Results of data preprocessing are explained with each of used approaches, results of analysis explained with the classifiers used and evaluating of the efficient approach to achieve maximum success are described in the chapter.

Chapter 6 Contains conclusion of the research project and possible future enhancements.

Chapter 2

Background and Related Work

2.1 Background

In this chapter, I will discuss the background and previous work related to this project. The literature on natural language processing and its applications, machine learning, Twitter content analysis focusing on social media related sentiment analysis, visual sentiment analysis, contextual analysis domains, spam and irony detection is included.

2.2 Natural Language Processing

Natural Language Processing(NLP) is an area of computer science that used to explore how the computers understand and manipulate the human language. NLP researchers research on how the humans process data and information and try to program the computers to process the data and information as humans. By doing so, computer systems will be able to perform, function human-like behaviour in language processing. The information or data might be in forms of texts, keyboard inputs or spoken language. NLP will translate the inputs into another language, analyze the content of the input and present results of the analysis by building different datasets or summaries. It is difficult to say the extent of how much the developed computer system ‘Understand’ the language compared to humans. What we can understand is how much the system appears to understand the language to perform a given activity successfully. The output of the NLP is depended on the desired task that requires to be performed by the processing.

2.2.1 Development of NLP

The first use of computers to manipulate natural languages was in the 1950s with attempts to automate translation between Russian and English in the World War II [9]. These systems were

dramatically unsuccessful requiring human Russian-English translators to pre-edit the Russian and post-edit the English. Based on World War II code-breaking techniques, they took individual words in isolation and checked their definition in a dictionary. But this technique was unsuccessful at that time as the translation was quite incorrect. One of the famous tales of this translations is mistranslating the phrase "hydraulic ram" translated as "water goat".

By the 1960s NLP systems improved to examine sentence structures compared to the translations techniques used in 1950s. The systems were based on pattern matching in the provided inputs and few systems derived representations of the meaning. In the mid-1970s as systems started to use more general approaches and attempt to formally describe the rules of the language they worked with and it dramatically improved the applying NLP in systems [9]. In the 1990s, NLP has started to focus on specific, limited domains due to difficulties in developing a universal system to understand general language.

2.2.2 NLP Process

This section, I will discuss natural language processing related literature. The simplified view of NLP emphasizes four logical stages of the process which occur as separated but sequential activities. The first activity of the NLP is feeding the input stream to the system and system will perform morphological processing of the inputs, then the syntax and semantic analysis where words and grammar will be understood by provided rules. Lastly, pragmatic analysis interprets the results of semantic analysis from the perspective of a specific context in which the inputs are processed for. The following Figure 2.1 displays the four logical stages NLP process.

Data Pre-processing

The input information is required to be preprocessed before feeding it to the system. For the preprocessing, different processing techniques can be used and the used technique have a significant impact on the accuracy of the results of NLP process.

Remove Stop Words

Removing stop words is a common data cleaning approach taken by most of the NLP applications. Stop words refers to the most common words in a language. This approach removes the

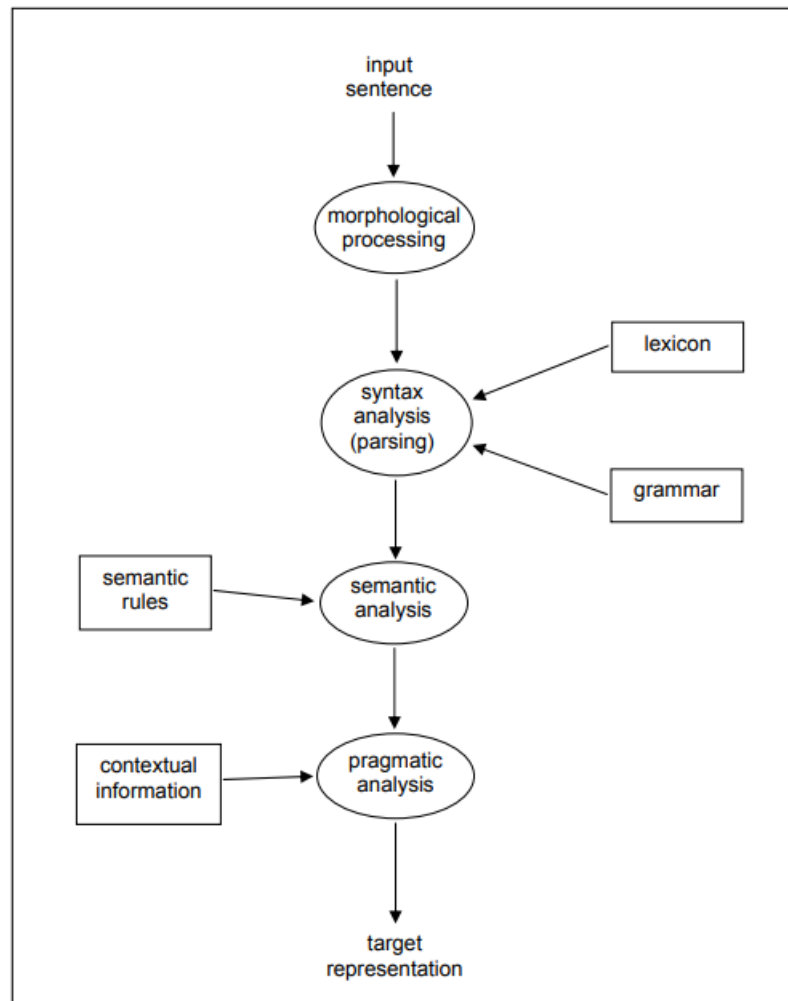


FIGURE 2.1: NLP Process

words of a data set which is identified as stop words.

For an example in English language the most common words that are used are "The" and "a". Zipf's law is an empirical law formulated using mathematical statistics [10]. It states that, if t_1 is the most common term in the collection, t_2 is the next most common, and so on, then the collection frequency cf_i of the i th most common term is proportional to $1/i$.

$$cf_i \propto \frac{1}{i}$$

According to Zipf's law, when applied to a language it states that top 20% of the most frequently used words in a corpus large enough will make up 80% of it.

N-Gram

n-grams are the all combinations of adjacent words in length on n, for an example, if we consider the word "I am a boy" the combination of words for the n=1 will be, "I", "am", "a" and "boy". It is the word distribution. When considering the n value of 2 the possible combination will be 'i am", "am a" "a boy". one of the purposes of using n-gram concept is to statistically determine the next word. For an example, in a text message, the next word suggestion is given by n-gram frequency

Stemming or Lemmatization

This approach reduces multiple forms of words and derives related forms of a word to a common base [11]. Stemming is the process that removes the ends of words in the hope of reducing inflectional forms and Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. According to the online resource, *Stemming and lemmatization* [11] Table 2.1 displays results when Stemming or Lemmatization apply to some words in the English language,

Words	After applying Stemming or Lemmatization
Am, are , is	be
Car, Cars, Car's, Cars'	car

TABLE 2.1: Stemming or Lemmatization Output

When differentiating Stemming vs Lemmatization, stemming usually does the derivation properly with the use of a vocabulary and morphological analysis of words. Lemmatization use Part-Of-Speech(POS) tagging to retrieve vocabulary and morphological analysis. POS Tagger is a piece of software that reads the text in some language and assigns parts of speech to each word (and other tokens), such as noun, verb, adjective, etc [12]. Following is how the Stanford CoreNLP POS [13] trigger works when a sample paragraph is provided.

Sample paragraph: Jane is a girl. She doesn't like cats and dogs.

First, POS Trigger breaks the paragraph into sentences and display the results. As per the sample paragraph, it contains two sentences. Figure 2.2 express the tagging of the first sentence.

Sentence #1									
Tokens									
Id	Word	Lemma	Char begin	Char end	POS	NER	Normalized NER	Speaker	Sentiment
1	Jane	Jane	0	4	NNP	PERSON		PERO	
2	is	be	5	7	VBZ	O		PERO	
3	a	a	8	9	DT	O		PERO	
4	girl	girl	10	14	NN	O		PERO	
5	.	.	14	15	.	O		PERO	

FIGURE 2.2: POS Trigger output - Sentence one

POS trigger tokenizes the first sentence into five distinct tokens including the period sign. In Lemma column, we can see the output result of each word when lemmatization applied. The Lemma of token #2 'is' is 'be'. In POS column, token #1 tagged as a proper noun, token #2 as a verb, token #3 a detriment, token #4 as a noun and token #5 as sentence final punctuation [14].

Figure 2.3 illustrate the POS tagging of the second sentence of the sample paragraph.

Sentence #2									
Tokens									
Id	Word	Lemma	Char begin	Char end	POS	NER	Normalized NER	Speaker	Sentiment
1	She	she	16	19	PRP	O		PERO	
2	does	do	20	24	VBZ	O		PERO	
3	n't	not	24	27	RB	O		PERO	
4	like	like	28	32	VB	O		PERO	
5	cats	cat	33	37	NNS	O		PERO	
6	and	and	38	41	CC	O		PERO	
7	dogs	dog	42	46	NNS	O		PERO	

FIGURE 2.3: POS Trigger output - Sentence two

The second sentence contains six words. But when Lemmatization is used, it breaks the sentence into seven tokens by breaking 'doesn't' contraction into two words as we can see in the token #2 and #3. Also, lemma of token #5 and #7 is the singular nouns of the plural words in the sentence. As per the POS tagging, token #1 tagged as a personal pronoun, #2 as the verb and token #3 as an adverb, token #4 as a verb base form, token #6 as a conjunction and token #5 and #7 as plural nouns [14].

Rather Stemming, Lemmatizing does a full morphological analysis to accurately identify the lemma for each word. But the efficient method of data preprocessing depends on the task on hand which the NLP is used.

	I	Love	Cats	Hate	and	Driving	is	my	Job	Passion
Sent 1	1	1	1							
Sent 2	1		1	1	1	1				
Sent 3					1	1	1	2	1	1

TABLE 2.2: Count Vectorizer in a corpus of three sentences

2.3 Machine Learning

Machine learning is the science of getting computers to act without being explicitly programmed where computers learn and act like humans do, and improve their learning over time in independently, by feeding data and information in the form of observations and real-world interactions. To use texts in machine learning algorithms, texts are required to convert into a numerical representation. Here are some of the textual data converting methods.

2.3.1 Bag of Words

This model ignores grammar and order of words in sentences. Once the corpus (text data) is received, a list of vocabulary is created based on the entire corpus. Then each document or data entry is represented as numerical vectors based on the vocabulary built from the corpora.

2.3.2 Count Vectorizer

The count vectorizer counts the occurrence of a word in a corpus. Assume we have three sentences in the corpus. ‘I Love Cats’, ‘I Hate Cats and Driving’ and ‘Driving is my Job and my Passion’ When a vocabulary is built from above three sentences, it will be as following Table 2.2.

2.3.3 TFIDF

TFIDF is another way to convert text data to a numeric form and is short for Term Frequency-Inverse Document Frequency. The vector value it produces is the product of TF and IDF. Following is a scenario how the Term Frequency-Inverse Document Frequency (TFIDF) works. Let’s say we have two documents,

1. I Love Cats

2. I Hate Cats and Cooking

This is the equation to calculate the Relative Term Frequency for each word in the document,

$$TF(t, d) = \frac{\text{Number of times term (t) appears in document (d)}}{\text{Total number of terms in document}}$$

TF for the word 'I' in both documents as follows

$$1. TF(I, d1) = 1/3 = 0.3333333$$

$$2. TF(I, d2) = 1/5 = 0.2$$

Inverse Document Frequency (IDF) of 'I' is calculated as follows,

$$IDF(I, D) = \log(2/2) = 0$$

Once TF and IDF values are calculated, TFIDF values can be calculated by multiplying TF and IDF.

$$TFIDF(t, d, D) = TF(t, d) * IDF(t, D)$$

For the 'I' value, TFDIF is as follows.

$$TFIDF('I', d1, D) = TF('I', d1) * IDF('I', D) = 0.33 * 0 = 0$$

$$TFIDF('I', d2, D) = TF('I', d2) * IDF('I', D) = 0.2 * 0 = 0$$

2.4 Tweet Content Analysis

In this section, I will discuss regarding some characteristics in tweets and what kind of analysis was performed in literature, such as sentiment analysis and contextual analysis. Figure 2.4 displays different contents available in a tweet.



FIGURE 2.4: Tweet Content

2.4.1 Hashtags

The hashtag is a word or a phrase starting with the symbol "#" and it should be without shape or punctuations [15]. According to Ma *et al.* [16] usage of hashtags are increased significantly in past few years, for an example, one of eight tweets contains one hashtag. Twitter users prefer to use hashtags to categorize their tweets and help tweets to show more easily in Twitter search.

Understanding the hashtags and their relationship is quite a challenging work. Hashtags are not easy to make a scene of, for an example #GoT is not about the user got something, it is about Game of Thrones TV series [17]. Hashtag such as #September11 is also challenging to understand the relativity is not clear. It might be about the terrorist attack on America or it might be the user's birthday [18]. A user can create a hashtag saying #MJ for the famous signer Michel Jackson and another user can create a hashtag saying #MichelJackson and both of them

are referring to the same singer.

Latent Dirichlet Allocation is a probabilistic generative model it can be used to extract information, "latent topics", in a collection of documents. Ma *et al.* [16] have proposed an extension for Latent Dirichlet Allocation, which accounts with both hashtag and the content of the tweet. In the model Tag-Latent Dirichlet Allocation (TLDA), Ma *et al.* [16] have proposed to solve some questions that we might face in the research, such as

- Understand, interpret hashtags and the context in which they are used
- How can we discover the relationships and correlations between the hashtags

2.4.2 Emoticons

The first emoticon was used on September 19, 1982, by professor Scott Fahlman [19]. In his message, Fahlman proposed to use ":-)" and ":-(" to distinguish jokes from more serious matters. Currently, emoticon plays a big role in Twitter due to the character limits.

According to Hogenboom *et al.* [20], emoticon can be used basically in three ways. First, emoticons can be used to express sentiment, when the sentiment is not clear in positive or negative words. Second, emoticons can stress sentiment by intensifying. Third, emoticons can be used to disambiguate sentiment. However, today's lexicon-based approaches typically do not consider emoticons.

If we read a tweet without considering the emoticon, the meaning of that tweet might be misleading. For an example Let's take two sentences with emoticons.

1. They have increased my salary :-D
2. I love my work -_-

When considering the first sentence we can say the emoticon is intensifying the sentiment, the person is actually happy. But when we read the second sentence we can identify that the emoticon is disambiguating the sentiment, the person does not actually love the work he does.

Hogenboom *et al.* [20] discuss exploiting emoticons in sentiment analysis and proposed framework for automated sentiment analysis, by using both sentiment lexicon and an emoticon lexicon. The combined analysis of both sentiment lexicon and an emoticon lexicon provides a better result [20].

2.4.3 Images

There are millions of images sharing in Twitter. Due to character limitations, using images in the tweets has become popular. When analyzing the tweet content, the content of the image plays a big role. We cannot simply ignore the image content and identify the sentiment. The most re-tweeted tweet in twitter contains an image [21]. The tweeter was Ellen DeGeneres [22]. There was no much text on that tweet.

In recent years NLP and image processing has become important application domains in machine learning. Recently deep learning has added more advantage in above two contexts.

Twitter Application Programming Interface(API) gives us the URL(s) for the images. From extracting the images and analyzing for duplicates, Hare *et al.* [23] have developed systems to visualize the trending images in Twitter. Using these kinds of systems, we can identify and visualize trending images and provide Twitter users a better Twitter experience.

When considering literature in visual sentiment analysis, Siersdorfer *et al.* [24] has proposed a machine learning algorithm to predict the sentiment of images using pixel-level features. Convolutional Neural Networks (CNNs) become the common approaches for extracting visual features [25]. Dumoulin *et al.* [26] describe a hierarchical approach and the use of a deep CNNs model to analyze a movie database.

Images also can stress sentiment by intensifying or to disambiguate sentiment that comes from the text. You [25] have proposed a progressive CNN for visual sentiment analysis and joint visual-textual sentiment analysis method to analyze both visual and text data. According to You [25], when analyzing we have to consider the text content and the image together.

2.5 Spam and Irony Detection

2.5.1 Spam Detection

Spam refers to "Irrelevant or unsolicited messages sent over the Internet, typically to a large number of users, for the purposes of advertising, phishing, spreading malware, etc" [27]. When using Twitter spam can be reported by clicking on the "report as spam" link in the home page. Twitter developers have also implemented blacklist filtering in their detection system called BotMaker [28].

We can categorize spam detection into two. Based on Syntax Analysis and based on Feature Analysis.

Detection based on Syntax Analysis, in literature Thomas *et al.* [29], describe a machine learning based methods that related to spam detection as a binary classification problem. According to conventional supervised detection paradigm, a set of labelled not-spam and spam tweets are used to train a classification model. Afterwards, the classification model is used to detect spam in the next tweets.

There are many account-related features such as Twitter account age, number of tweets, hashtags, tweets containing URLs, number of followers etc. When considering Detection based on Feature Analysis, Anita, Gupta, and Kumar [30] propose a process to identify the spammers by studying and comparing the temporal behaviour of the particular user. Wu *et al.* [31] discuss the usage of Supporting Vector Machine method to identify both spam and spammers. Lee, Caverlee, and Webb [32] describe the process of using honeypots to gather spammer profile features, they trained machine learning algorithms such as Decorate and LogitBoost to detect spammers. In all above literature, some of the features related to the Twitter profile were used to identify the spammers and spams.

2.5.2 Irony Detection

The irony is known as a way of communicating the opposite of the literal meaning. In the context of Sentiment Analysis, the irony direction will be a challenging task. The common way of analysis is to clarify the text in "negative", "positive" or "neutral" but when considering the ironic statement, the result, in general, is a misinterpretation [33].

When considering literature in ironic detection Freitas *et al.* [33] has proposed thirteen patterns to detect ironic statements. The research was conducted using a collection of tweets under the domain "Fim do mundo"("End of the world"), composed of 2,779 tweets (55,663 words).

The phrase "End of the world" was chosen due to the fact it was largely used by the users in that time period. Researchers observed patterns related to ironic statements and these patterns were partially extracted from and some of them created with the help of specialist native speaker. In this study, thirteen patterns were implemented, and they were classified into six categories [33].

2.6 Summary

There are numbers of methods to classify documents as well as tweets, such as analysing tweet text content, tweet image content tweet emoticons and Emoji content. But there is not combined analysis performed for tweets. When continuing the research there is a possibility that we can combine those individual analysis methods together and also combine irony detecting spam detection features as well.

Chapter 3

Analysis and Design

3.1 Introduction

This chapter includes the high-level architecture of the system, modules and their usage. The module naming and boundaries can be changed alongside with the development. The chapter also discusses the feasibility of the application development, development constraints and the solutions taken to avoid them.

3.2 System Analysis and Feasibility

3.2.1 Integrating with Twitter API

Twitter provides two types of APIs to the outside world to interact with Twitter data.

- Twitter Rest API
- Twitter Streaming API

Twitter Rest API provides smart endpoints to extract twitter data. I have listed down several endpoints that might be helpful for gathering the research related data [34].

- GET statuses/home_timeline

Returns a collection of the most recent tweets posted by the authenticating user and the users they follow.

- GET statuses/user_timeline

Returns a collection of the most recent tweets posted by a user indicated from the screen_name or user_id parameters

- GET statuses/mentions_timeline

Returns the 20 most recent mentions (tweets containing a users's @handle) for the authenticating user.

Twitter Stream API and the REST API both provide data in well-formatted JSON format. Sample Twitter response is listed under Appendix A.1. The user authentication of Twitter is based on OAuth1 process. Any modern programming language can be used to process the data in JSON format. Because of that, the integration with the APIs and accessing the Twitter data becomes simple.

There are third-party libraries that also can be used to access Twitter data such as,

- Spring Social Twitter
- Twitter4j

Above libraries provides authentication and data processing functionalities with twitter [35], [36].

3.2.2 Feasibility

From using Twitter stream API we can gather random tweets for the analysis and from the REST API, we can gather specific users tweets from home timeline. Using above two APIs we can gather required data from the Twitter.

According to the *Overview — Twitter Developers* [34], above three endpoints returns the timeline of tweets of the respective categories, but there are limitations.

The most important restriction that we have to manage is the Rate Limiting of the Twitter API [37]. Rate limiting of the Twitter API is based on a per-user access token. For an example for a single user the GET statuses/home_timeline request is only allowed to invoke 15 times in a 15 minutes window [Requests / 15-min window (per user token) =15]. The rate limit depends on the API endpoint.

Thus we have to carefully manage the number of requests we make to the Twitter API. There were several recommendations given to avoid being rate limited. Such as, caching data and prioritizing user activities.

3.2.3 Twitter Data

Sample Twitter response is listed under Appendix section : A.1 , Twitter APIs provide data in JSON format. Data is well formatted and categorized. I have listed down some of the data that can be extracted from the JSON response.

1. Tweet Text
2. URLs
3. Image URLs
4. Hash-Tags
5. Created date and time
6. Details regarding tweeter
 - (a) Name
 - (b) Location
 - (c) Profile created date
 - (d) Profile image URL

The implemented system can process tweets from English Language only. Characters from any other language are considered as unwanted textual information.

3.3 System Design

According to the facts that have mentioned in the Introduction section, 82% of Twitter users access their twitter account using mobile. Because of that, providing a mobile application for the users becomes a must.

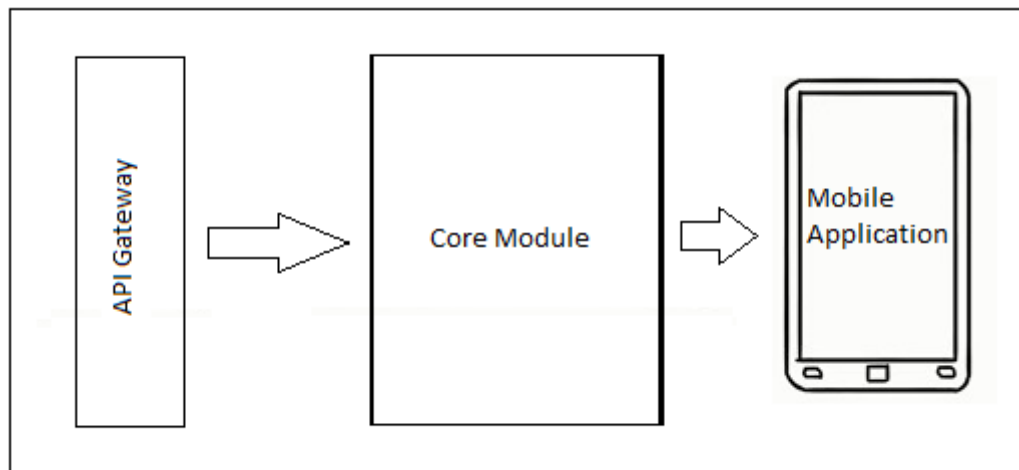


FIGURE 3.1: System high level architecture

Figure 3.1 contains a high level architecture of the application. There are 3 main components of the system.

1. Twitter API Gateway module
2. Core application module
3. Mobile application

The Twitter API Gateway module will basically interact with the external Twitter APIs and extract user timeline. The users will be authenticated from the Twitter APIs and an access token will be saved on behalf of the user. API Gateway module can use to both stream API and the REST API to extract data. Stream API can be used to gather random tweets from Twitter and the REST API can be used to extract specific user timeline.

Core application module is responsible for analyzing the tweets and learn about each user preference. The users give feedback for the tweets. That feedback is used to learn about each user. There are multiple sub-components in this core module to separately analyze the tweet content and the feedback.

The Mobile application and related module are used to gather user feedback. Initially, the users can view their normal Twitter home timeline. When the user starts giving feedbacks for the tweets, the timeline will be filtered accordingly. Figure 3.2 decompose the core application module, there will be several sub-components to perform specific analytical tasks.

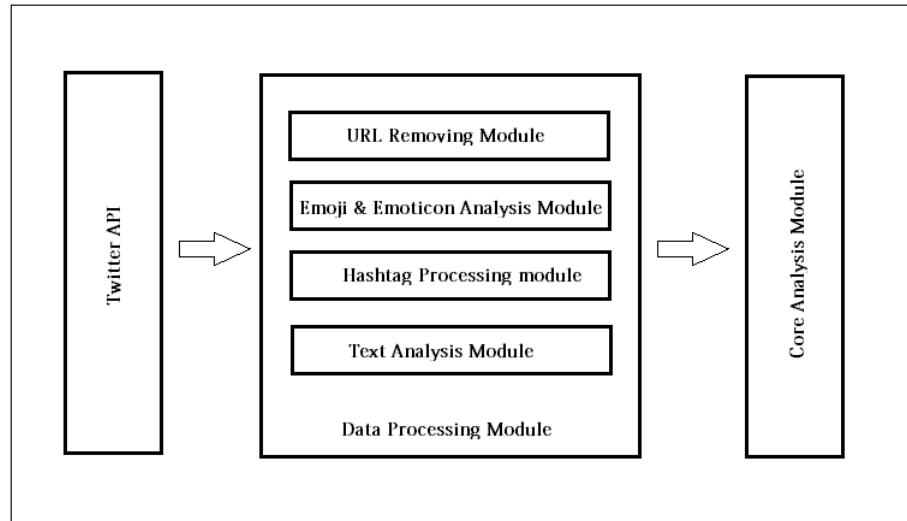


FIGURE 3.2: System high level architecture decomposed

Spam Detection Module This module will be used to filter spam tweets. Spams will be detected and removed from further processing. We can mainly use two approaches to identify spam tweets. The spam detection can be considered as a classification problem, a learning classifier system can be used to detect spams and filter next coming tweets accordingly. There are some profile related characteristics that also can be used to filter the spam tweets. Such as the number of tweets and age of the account number of frequent tweets etc. In this module, we can use one or both above-mentioned analysis to detect tweets.

Image Processing module This module will be used to extract images from the Uniform Resource Locator(URL) and predict the sentiment of images. The image data will be very useful when identifying the user preference. Images can stress sentiment by intensifying or to disambiguate sentiment that comes from the text. There are machine learning algorithms to predict the sentiment of images. By using those kinds of machine learning algorithms we can extract the sentiment of the image. These extracted data will be sent to core module to perform analysis alongside with the textual analysis.

URL Processing Module Due to the text restrictions in twitter, some tweeters use URLs and redirect users to other websites. The Twitter user is required to open the URL from via another third-party application. Most of the tweets contain short URLs. URL itself does not contain any

valuable information to process. The URL detection module detects the URLs, images from the tweet and removes them from the tweet.

Hashtags Processing module Hashtags are one of the important characteristics of tweets. Processing and identifying the Hashtags is important in identifying the user preference. This module will extract the Hashtags from the tweets and clean the data for processing. For an example, #MJ and #KingOfPop refers to the same individual called Michael Jackson. This module processes those kinds of hashtags into one unit and does the processing. The output of this module is sent to the core module to analyse with other module information.

Text Processing Module This is the most important analysis that we have to perform to identify the user preference. tweets have character limits to 140 characters, Due to these restrictions users intend to use Emoticons and Short words. The extracted text data should be cleaned (identify the short words) and processed.

Emoticon Processing Module The emoticon processing module extracts the emoji, emoticons from the tweet and converts them into its text representation.

Core Analysis Module The combined output of above-mentioned modules will be processed from the core module to give the final output. For an example, if a tweet contains an Image, URL, Text and Some Hashtags. Each of these modules extract the corresponding data and do necessary cleaning and pre-processing that to be done. The final output will be analysed from the Core module.

Chapter 4

Implementation

4.1 Introduction

This chapter describes the implementation process of the system, where it has described implementation tools and techniques in a number of steps. The implementation of each of modules was described in the chapter and the selection criteria of each of libraries and programming languages were also described.

4.2 Technologies, Libraries and Frameworks

The nature of the project is about natural language processing, therefore, programming languages and tools were selected which supports NLP.

4.2.1 Programming Languages

The main programming languages that were used was Python and Java. Python programming language heavily supports natural language processing. The JAVA API was used to integrate with twitter API and gather the relevant Twitter data.

4.2.2 Libraries and Frameworks

Natural Language Toolkit(NLTK) is a leading platform for building Python programs to work with human language data [38]. NLTK can provide word tokenization, stemming, tagging, and parsing functionalities that are advantageous with the language processing.

Scikit-learn is a simple and efficient tool for data mining and data analysis [39]. And it provides a rich toolset for natural language processing including vectorizers, classifiers etc. Scikit-learn was selected mainly because of NLTK and Scikit-learn both build on top of the

common Python language. Scikit-learn also includes examples and tools for classifications, regressions and also clustering.

4.3 System Implementation

4.3.1 Mobile Application Implementation

A vast majority of Twitter uses their mobile devices to access Twitter. Because of that Implementation of a mobile application was a definite need. A simple Twitter-like mobile application was developed to access tweets from the user and get user feedback for the tweets.

The user interface for the mobile application is same as the Twitter. The only difference is that it is given a thumbs up and thumbs down button to capture user feedback. The mobile application was developed as a hybrid mobile application using AngularJS and Apache Cordova. The advantage of using hybrid application was that we can easily create either Android or iPhone Operating System(IOS) application without having an in-depth knowledge of either platform.

The Mobile application interacts with the core application module using REST web services. Thus the communication between core application and mobile application was easily managed. After creating the AngularJS application, it was converted to Android mobile application using Apache Cordova.

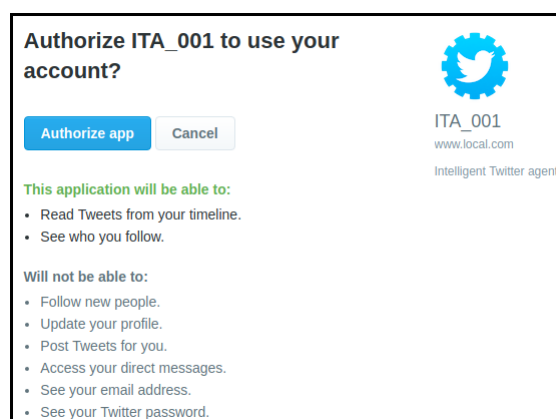


FIGURE 4.1: Authorization Request

According to the Figure 4.1, user will be asked to authorize and log in using Twitter login credentials. Twitter login was provided using Oauth1 authorization framework.

4.3.2 Application Core module implementation

The core application module was implemented using python, The application module contains submodules to data pre-processing and classification. Python has a vast range of libraries and frameworks that support NLP and also machine learning applications.

The core application module communicates with an API layer developed using JAVA language. JAVA gives good support in web service implementation and third-party API integrations. The communication between the Twitter API and Core module is being handled by the API gateway.

When a user initially logging to the application, the user will be redirected to the Twitter for login.in return Twitter provides an access token on behalf of the user. Intelligent Twitter Agent(ITA) API layer saves the access token and uses the access token for the further communication with Twitter.

For each user there will be access tokens, These access tokens will be used to access user home timeline. These user home access data will be stored in the system for classifier training purposes. After training the classifier the trained classifier model will be saved against the Twitter user [40].

After the training of the classifier no further data will be gathered from the user. The Twitter home feed will be directly filtered by using the user's trained classifier model.

Data Pre-processing

The tweets that were gathered required to be cleaned and processed before proceeding to the analysis. There are series of cleaning and preprocessing activities performed on top of each tweet.

As a data cleansing approach, the accented characters were removed from the tweets by converting the tweets into Unicode. This conversion makes all the tweets into a single form of Unicode and it will be easy to proceed with further data cleaning process.

Tweets can contain Emoji symbols, in the data cleaning process, the emoji symbols were converted to the text format. Emoji package included in Python was used to decode the emoji into text [41]. Python Emoji package contains the entire set of Emoji codes as defined by the Unicode consortium [42] and also it supports additional branch of aliases [43].

Let's take an example of a tweet contain following emoji " 🐵💋🍤👊 ". The tweet will be converted in to "speak no evil monkey mouth fried shrimp raised fist"

The extracted Tweets can contains Hypertext Markup Language(HTML) entities. Let's take an example "Whinging. My client& boss doesn't understand English well. Rewrote some text unreadable. It's written by v. good writer& reviewed correctly."

Above tweet is in the dataset that and it contains HTML entities & , The data cleaning process clean these HTML entities to "Whinging. My client & boss doesn't understand English well. Rewrote some text unreadable. It's written by v. good writer&reviewed correctly." Python in build HTML Unescape function was used to unescape HTML entities from the tweets [44].

Tweets also can contain HTML tags. Beautiful Soup [45], Python package was used to remove the HTML tags from the tweets. HTML URL tags also do not add important value to the tweets. Thus the URLs are also removed.

Due to character restrictions, users may prefer to use contractions in the tweets. Contractions, which are sometimes called "short forms", commonly combine a pronoun or noun and a verb, or a verb and not [46]. To extract the meaning of the tweets in the data cleaning process, I expanded the contractions. A list of contractions was mapped with the expanded word is used to perform this action.

Lemmatization is performed in the tweets in data cleaning process, the goal of lemmatization is to reduce inflectional forms of tweets and derive related forms of a word to a common base form [11]. For an example car, cars, car's, cars' can be derived as "car". NLTK WordNetLemmatizer [47] is used to lemmatize the tweets. Without a POS tagger [48], WordNetLemmatizer assumes everything that feed in as a noun. I used NLTK POS tagger to tag the tweets. The tagged tweets sent to Lemmatization function afterwards.

Tweets also contain emoticons, such as :) smile face, :(sad face. These kinds of emoticons were extracted from the tweets and converted into words. O'Connor [49] provided a rule-based implementation to identify the emotions, I have add extensions to the code and develop an improved version to extract emotions from the tweet.

If we post a tweet " I am a :) person, I have all I need", the extractor will extract the emoticons into words and tweet output will be "I am a happy person, I have all I need". After extracting emotions, emoji and expanding contractions we will be able to remove the special characters

from tweets. This step is performed at last because all emotions, emoji and expanding contractions rely on special characters. Numbers in the tweets also removed because the numbers also do not add value to the dataset.

Removing stop words, According to Zipf's Law, we only use a small number of words at all the time. Rest of the vast majority of words used rarely. The word "The" is the most used word in the English language. Stop words usually have little lexical content. Presence of a stop word in a text fails to distinguish it from other texts [50]. In natural language processing, as a data cleaning process the stop words AKA most commonly used words will be removed.

The Table 4.1 contains a sample of cleaned data. Data were cleaned by all above data cleaning procedures.

	text	target
0	sick damn snow	0
1	iampritty try late	0
2	lose game	0
3	disapointed cannot sign idol sweepstakes canad...	0
4	video feed go asot	0

TABLE 4.1: Cleaned Twitter data

Chapter 5

Evaluation

5.1 Introduction

In this chapter, I will discuss the research evaluation. The System evaluation is performed by using real user data. The most accurate classifier was selected according to the evaluation and included in the system to identify the user preferences and filter the tweets. The evaluation selects the best classifier and data preprocessing approach which performed on top of a small dataset. The selected classifier should be able to perform well with a small amount of training data. The evaluation was carried out in two different data sets, One on top of a large dataset (tweets contains 1,500,000) and a small dataset (tweets of 230).

5.2 Data and Resources

‘Sentiment140’ is a project, oriented from Stanford University [51]. The data contains 1.6 million classified tweets. The Table 5.1 contains a sample from that classified data, the information of each field as follows, the sentiment is the polarity of the tweet. 0 being negative and 4 being positive. id field is the ID of the tweet. The date is the date and time which the tweet was posted. The query string is the query used to search the tweet, user fields include the username of the Tweeter and lastly text field contains the content of the tweet. This dataset is used to evaluate most accurate data pre-processing approach and classification algorithm.

	sentiment	id	date	query_string	user	text
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....

TABLE 5.1: Labeled Twitter data

Next set of data will be gathered from Twitter via application users. According to Figure 4.1, the users will be notified that the application requires read-only access to the tweets. If the user accepts the request to access their data, the application will be able to gather the home timeline tweets of the user. After gathering the data, those will be stored in and filtered preference according to the results that previously observed best data pre-processing and classification algorithm.

5.3 Data Pre-processing and Evaluation

Data cleaning is an essential part of natural language processing and machine learning approaches. The data can contain outliers and noise. In this section, I have described the steps taken to clean and remove noise in the data. There were several approaches taken to clean the data. extract Emoji, extract Emoticons, remove stop words, data lemmatizing and expand constraints are the data cleaning approaches. The implementation of those data cleaning steps was described in detail in Chapter 4.

Each data clean approach was evaluated to study the effect of each approach. For an example, I have removed Stopwords for data cleaning purposes and evaluate how much of accuracy that we can gain from stop words removed data set. If a data cleaning approach decreases the accuracy of the classification we can identify and eliminate such approach.

Define Baseline

When comparing multiple data cleaning approaches, we have to provide a baseline as a point of reference to compare with. This baseline dataset was created by performing only the most necessary data cleaning activities. There are basic data cleaning activities that should apply on top of the data. Special characters, URL s, HTML tags and unnecessary spaces do not add a value to the data and they were removed with a basic data cleaning option such as escape URL, unescape HTML characters, escape HTML, lowercasing of text and removing unwanted spaces.

By removing such unnecessary data as a whole, some of the tweets data becomes completely unusable. We can identify some of the tweets originally in the dataset becomes null after data cleaning process. We can identify those original tweets only contains URL information or

only special characters. These NULL entries were removed from the cleaned dataset to reduce classification errors.

After performing basic data cleaning operations I have prepared set of tweets containing nearly 1,500,000 tweets. This Dataset will be taken as a point of reference to compare with other data cleaning functions (baseline dataset).

According to Scikit learn [52], with a small amount of data, Naïve Bayes classifier should perform well in text data classification. Thus the Naïve Bayes classifier was selected as a point of reference to compare with other classifiers. For the feature representation, I have used Count Vectorizer.

All the classification accuracy scores were calculated using k-fold cross-validation. Table 5.2 contains the accuracy values calculated against the number of featured used for the baseline dataset.

	No of Features	Mean Score	Standard Deviation
0	4000	0.77091	+/- 0.00052
1	14000	0.77894	+/- 0.00044
2	24000	0.78072	+/- 0.00068
3	34000	0.78155	+/- 0.00031
4	44000	0.78203	+/- 0.00036
5	54000	0.78227	+/- 0.00034
6	64000	0.78251	+/- 0.00032
7	74000	0.78266	+/- 0.00037
8	84000	0.78280	+/- 0.00026
9	94000	0.78278	+/- 0.00049

TABLE 5.2: Baseline Dataset Accuracy vs Number of features

According to the Table 5.2 we can identify that the accuracy increases with the number of features used.

Removing Stop Words

First data cleaning approach used is removing stop words in the baseline dataset. Afterword the dataset was classified by using the Naïve Bayes classifier. The Table 5.3 contains the cross-validated accuracy score of the stop word removed dataset against the number of features used.

When comparing against the baseline dataset we can clearly identify that the removing stop words have drastically decreased the performance of the classifier. Some of the important

features of the tweets may have been removed alongside with the stop words. The Figure 5.1 visualize the drastic decrease in the accuracy of classification.

	No of Features	Mean Score	Standard Deviation
0	4000	0.75915	+/- 0.00068
1	14000	0.76915	+/- 0.00075
2	24000	0.77128	+/- 0.00034
3	34000	0.77241	+/- 0.00048
4	44000	0.77293	+/- 0.00027
5	54000	0.77322	+/- 0.00026
6	64000	0.77352	+/- 0.00022
7	74000	0.77359	+/- 0.00021
8	84000	0.77371	+/- 0.00025
9	94000	0.77376	+/- 0.00042

TABLE 5.3: Stopwords removed dataset accuracy vs Number of features

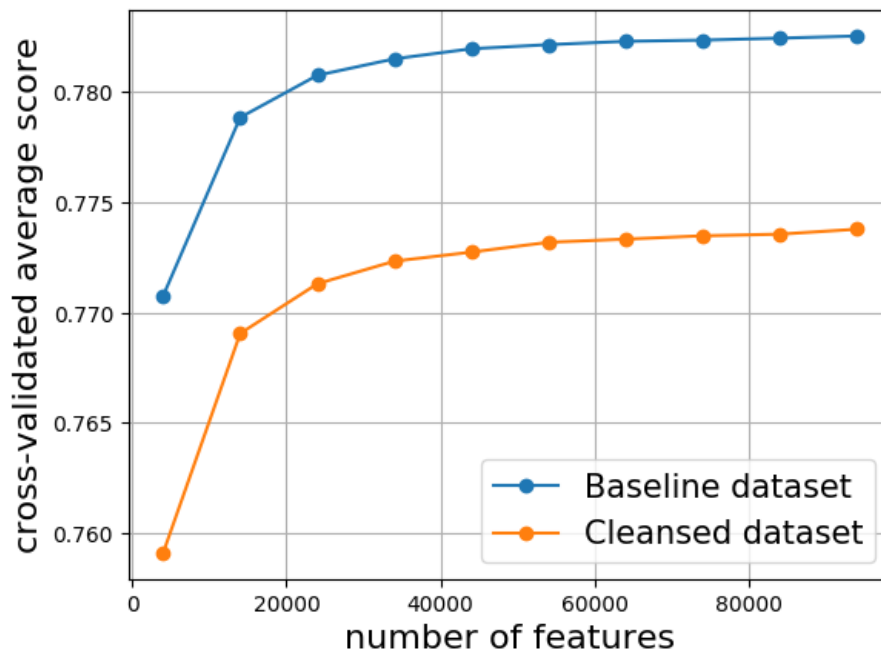


FIGURE 5.1: Control dataset accuracy vs Stop words removed dataset accuracy

Expanding Contractions

Expanding contractions is another data preprocessing approach. The baseline dataset was processed from the expanding contractions approach and then the dataset was classified using the

Naïve Bayes classifier. The Table 5.4 contains the cross-validated accuracy score of the expanding contractions used dataset against the number of features used.

	No of Features	Mean Score	Standard Deviation
0	4000	0.77074	+/- 0.00054
1	14000	0.77862	+/- 0.00113
2	24000	0.78050	+/- 0.00101
3	34000	0.78145	+/- 0.00073
4	44000	0.78189	+/- 0.00062
5	54000	0.78211	+/- 0.00055
6	64000	0.78234	+/- 0.00068
7	74000	0.78253	+/- 0.00064
8	84000	0.78260	+/- 0.00084
9	94000	0.78265	+/- 0.00076

TABLE 5.4: Contractions expanded dataset accuracy vs Number of features

When analyzing the results we can identify that the accuracy of the classification drops slightly when using expanding contractions as the data cleaning approach. The Figure 5.2 contains a summary of the accuracy.

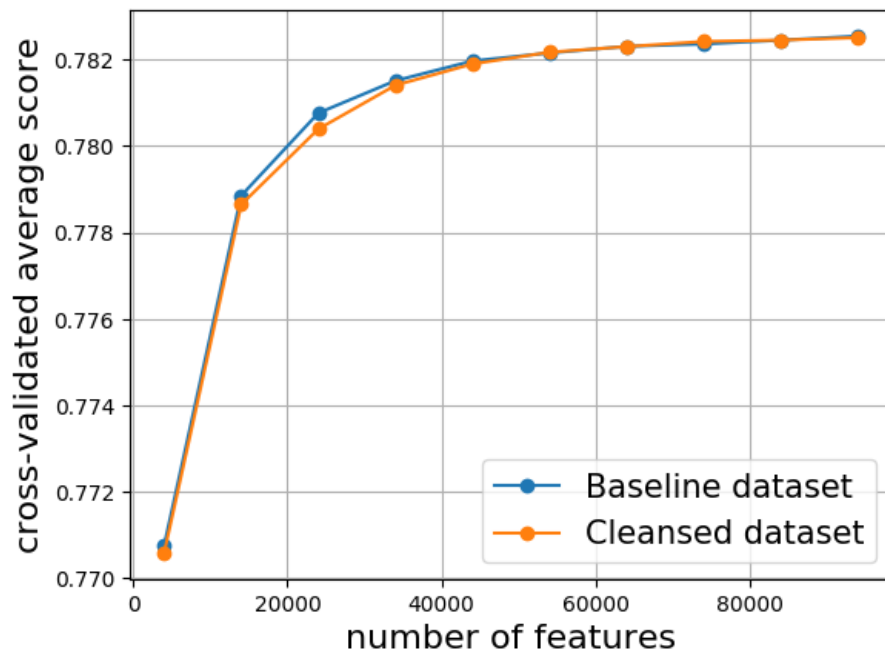


FIGURE 5.2: Baseline dataset accuracy vs Contractions expanded dataset accuracy

Converting Emoj in to Text

When considering tweets there are Emoji and also Emoticons. In this preprocessing function, I have considered about extracting Emoji into text form. The baseline dataset was processed and the emoji were converted into text. The processed data set is sent to the Naïve Bayes classifier and the results were compared with the baseline. Table 5.5 contains the Emoji converted dataset cross-validated accuracy score against the number of features used.

	No of Features	Mean Score	Standard Deviation
0	4000	0.77107	+/- 0.00023
1	14000	0.77905	+/- 0.00080
2	24000	0.78078	+/- 0.00064
3	34000	0.78169	+/- 0.00064
4	44000	0.78217	+/- 0.00035
5	54000	0.78242	+/- 0.00048
6	64000	0.78262	+/- 0.00054
7	74000	0.78281	+/- 0.00068
8	84000	0.78288	+/- 0.00077
9	94000	0.78296	+/- 0.00075

TABLE 5.5: Emoj converted dataset mean accuracy vs Number of features

The Figure 5.3 brings a summary of the comparison. The Emoji converted dataset has improved accuracy over the baseline dataset. We can conclude that extracting Emoji into the text format has increased the accuracy of classification.

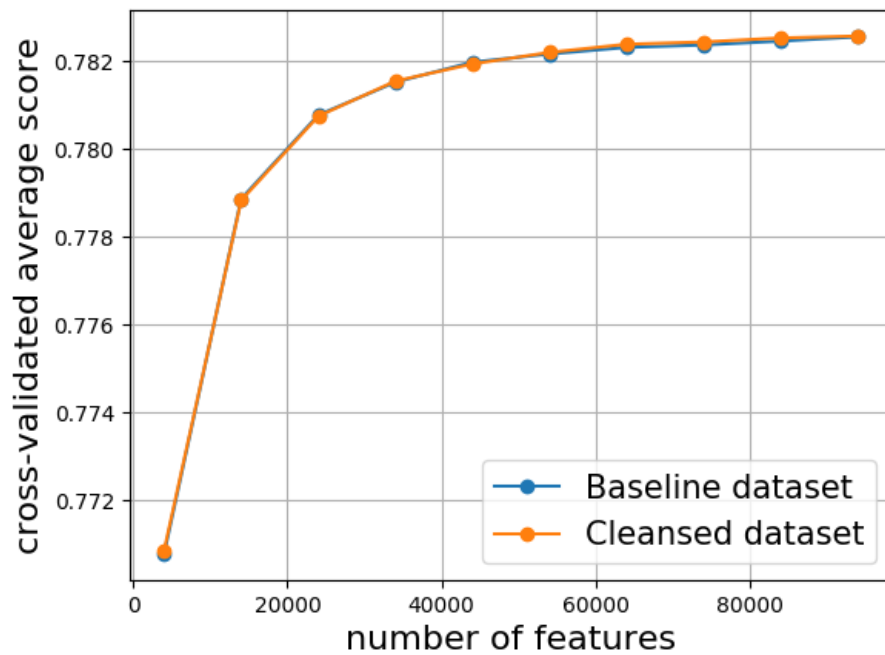


FIGURE 5.3: Baseline dataset accuracy vs Emoji converted dataset accuracy

Converting Emoticons into Text

In this preprocessing approach, the Emoticons are converted into text format. The baseline dataset was processed from this approach and all the Emoticons of the baseline dataset were converted into text format. The cleansed data is sent to the classifier.

The Table 5.6 contains the classifier cross-validated accuracy score against the number of features used. The Figure 5.4 brings a summary of the comparison. When comparing the results, we can identify by extracting the Emoticons has only a small effect on the accuracy of the classifier. Most of the time the accuracy of the classifier is being decreased by the Emoticons to text processing function.

	No of Features	Mean Score	Standard Deviation
0	4000	0.77064	+/- 0.00045
1	14000	0.77872	+/- 0.00067
2	24000	0.78054	+/- 0.00079
3	34000	0.78150	+/- 0.00102
4	44000	0.78207	+/- 0.00105
5	54000	0.78230	+/- 0.00100
6	64000	0.78238	+/- 0.00100
7	74000	0.78255	+/- 0.00091
8	84000	0.78263	+/- 0.00106
9	94000	0.78274	+/- 0.00112

TABLE 5.6: Emoticons converted dataset accuracy vs Number of features

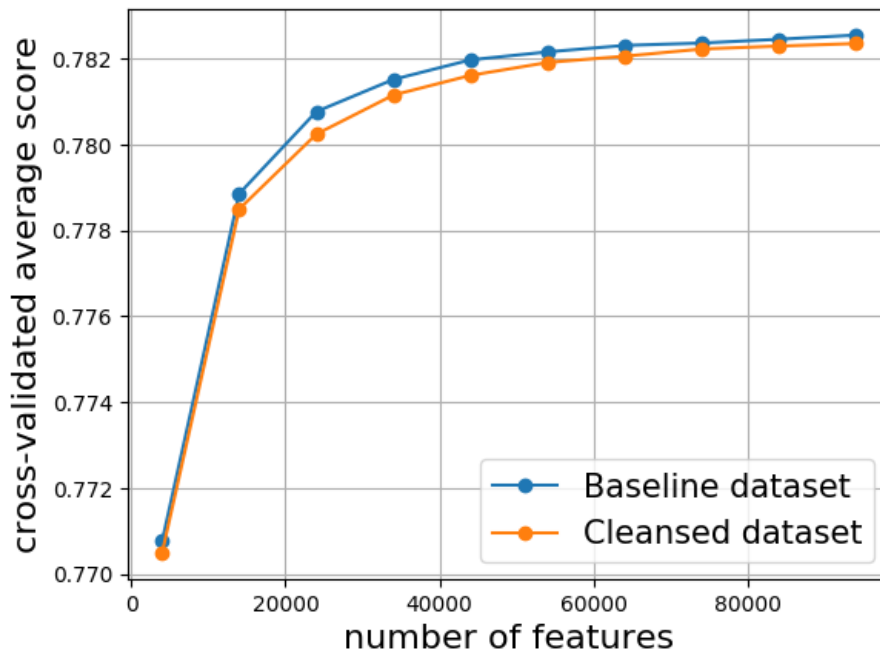


FIGURE 5.4: Baseline dataset accuracy vs Emoticons converted dataset accuracy

lemmatize

Lemmatizing data is another data preprocessing approach. The baseline dataset is lemmatized and sent to the classification. The cross-validated accuracy scores are listed in the Table 5.7. As per the result with the baseline dataset accuracy, we can clearly identify the accuracy of the classifier has been decreased.

The figure 5.5 contains a summary of the comparison. The figure clearly shows the decrease of the accuracy with the accuracy of the baseline dataset accuracy.

	No of Features	Mean Score	Standard Deviation
0	4000	0.76878	+/- 0.00090
1	14000	0.77559	+/- 0.00091
2	24000	0.77703	+/- 0.00080
3	34000	0.77771	+/- 0.00061
4	44000	0.77821	+/- 0.00072
5	54000	0.77840	+/- 0.00062
6	64000	0.77856	+/- 0.00061
7	74000	0.77861	+/- 0.00072
8	84000	0.77877	+/- 0.00087
9	94000	0.77892	+/- 0.00076

TABLE 5.7: Lemmatized dataset accuracy vs Number of features

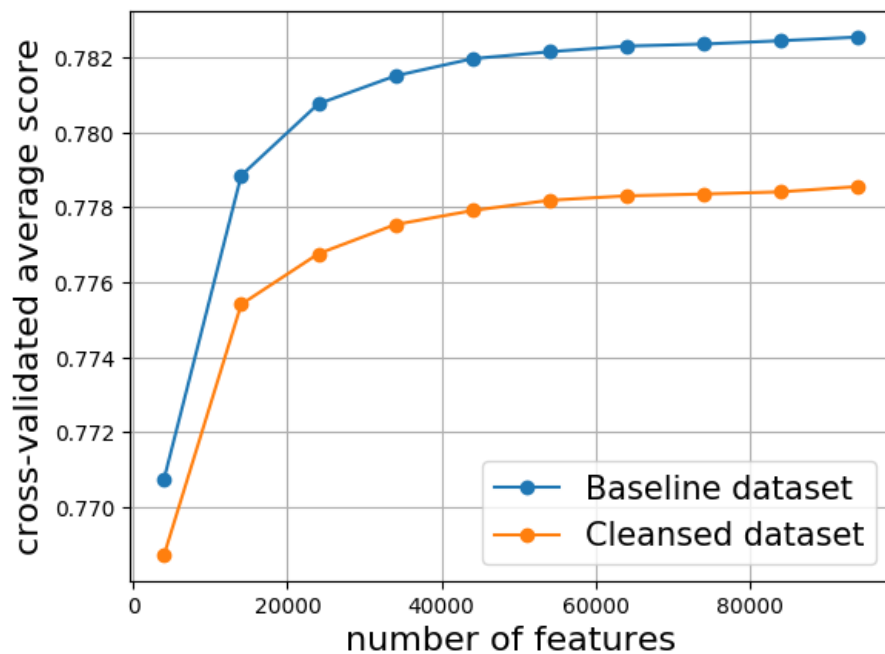


FIGURE 5.5: Baseline dataset accuracy vs Lemmatized dataset accuracy

Remove Numeric Characters

Numeric characters in a dataset usually do not add value to the dataset. To analyze this concept, the baseline dataset is processed with a function to remove numeric characters. The Table 5.8 contains the cross-validated accuracy scores.

The accuracy of the classification has increased from removing the numeric characters. The Figure 5.6 clearly shows the comparison with the base data set and when the number of features increased the processed dataset accuracy also increased.

	No of Features	Mean Score	Standard Deviation
0	4000	0.77095	+/- 0.00018
1	14000	0.77905	+/- 0.00072
2	24000	0.78076	+/- 0.00069
3	34000	0.78177	+/- 0.00056
4	44000	0.78227	+/- 0.00071
5	54000	0.78249	+/- 0.00071
6	64000	0.78266	+/- 0.00062
7	74000	0.78284	+/- 0.00079
8	84000	0.78304	+/- 0.00078
9	94000	0.78308	+/- 0.00087

TABLE 5.8: Numeric characters removed dataset accuracy vs Number of features

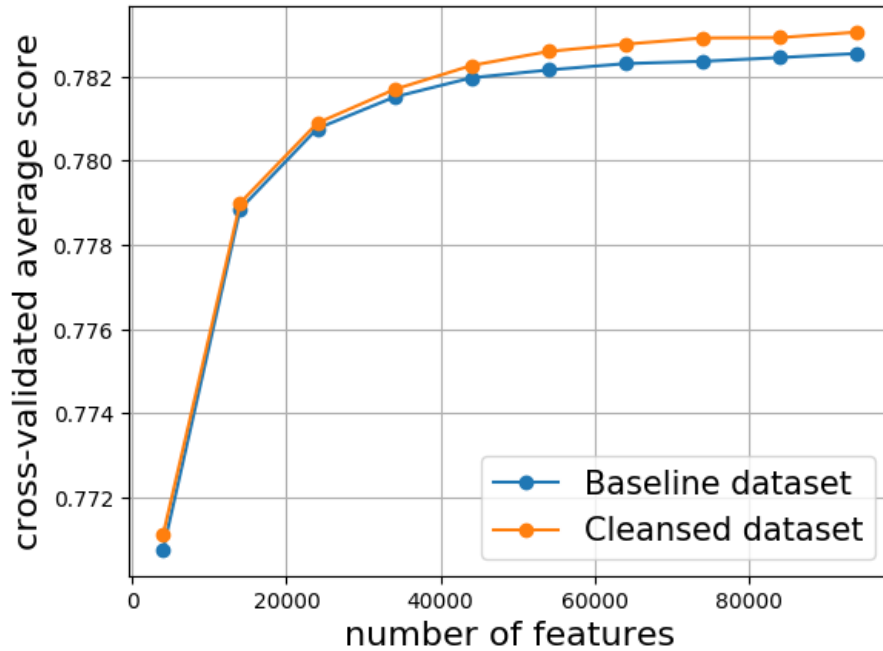


FIGURE 5.6: Baseline dataset accuracy vs Numeric characters removed dataset accuracy

5.4 Data Pre Processing and Evaluation- Small Dataset

Define Baseline

The goal of this evaluation is to find a suitable data processing functions that should perform for a small dataset (tweets of 230). For the evaluation, I have randomly selected ten datasets from the base dataset.

For each data preprocessing or data cleaning function, the cross-validated average score will be calculated against the number of features. For the randomly selected eight datasets the calculation will be performed eight times against each dataset and the average score will be calculated as the score for small datasets. Classifier performance will be evaluated by using k-fold cross-validation with K equals to 10.

Appendix B.1 table B.1 contains each cross-validated score for each dataset and average score value.

Removing Stop Words

The removing stop words function is evaluated against all eight control datasets. Appendix B.2 Table B.2 contains the cross-validated accuracy score for each dataset. When comparing the results with the average accuracy of the control datasets, we can identify that removing stop words decrease the accuracy of the classification. This behaviour is already observed under the large dataset also. Classification with the large dataset and also small dataset both decrease the accuracy when stop word removing approach is used.

The summary of the result is plotted in Figure 5.7

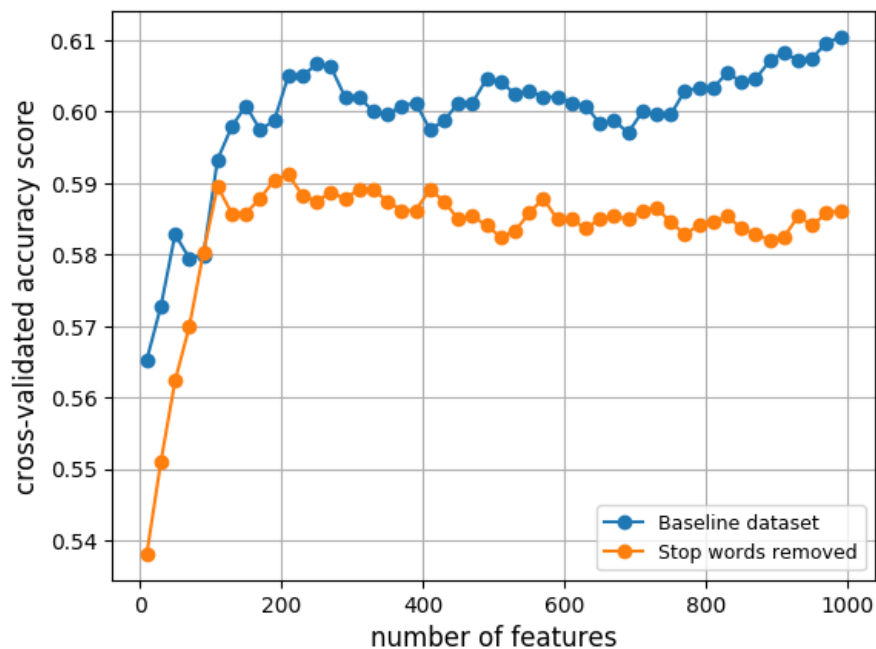


FIGURE 5.7: Baseline datasets average accuracy vs Stop words removed datasets average accuracy

Expanding Contractions

Expanding contractions function is also evaluated for all eight random datasets. The result is listed in Appendix B.3 Table B.3. When analyzing the retrieved results, It is clear that expanding contractions approach has significantly increased the accuracy. This is the opposite behaviour when comparing to the expanding contractions in the large dataset. In the large dataset, expanding contractions approach decreased the classifier accuracy.

Figure 5.8 is a comparison between baseline datasets average accuracy vs contractions expanded datasets average accuracy

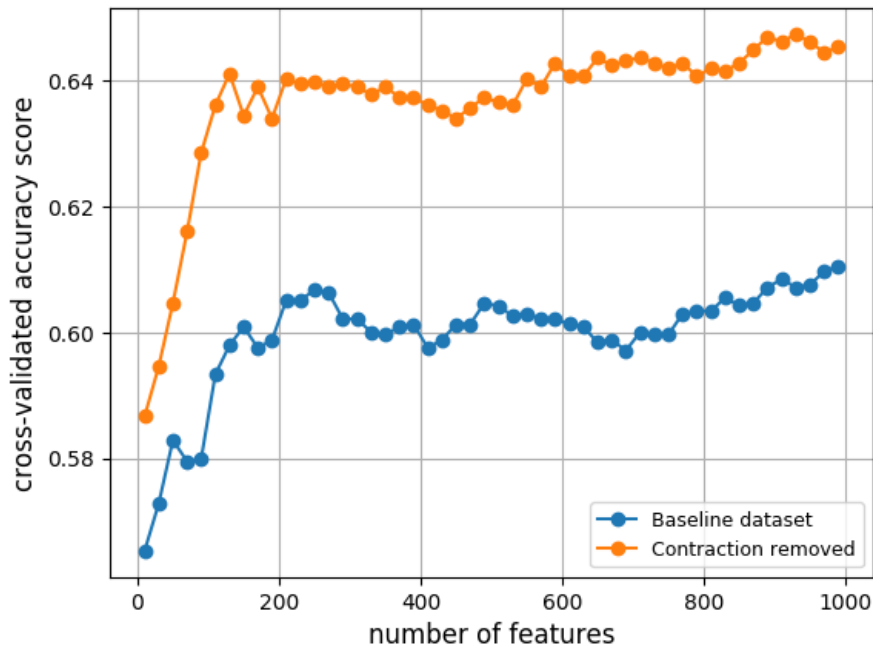


FIGURE 5.8: Baseline datasets average accuracy vs Contractions expanded datasets average accuracy

Converting Emoji in to Text

The Figure 5.9 is a comparison between the baseline dataset average score vs Emoji converted dataset average accuracy. According to the Figure 5.9 , converting emoji into text has increased the accuracy of the classification significantly when used for small datasets.

Converting Emoji approach has a similar behaviour when applied in both large and small datasets. In both datasets converting Emoji to text increases the classification accuracy. But in

the small dataset, accuracy was increased more than in the large dataset. The classifier evaluation results score for each baseline dataset is listed under Appendix B.4 Table B.4.

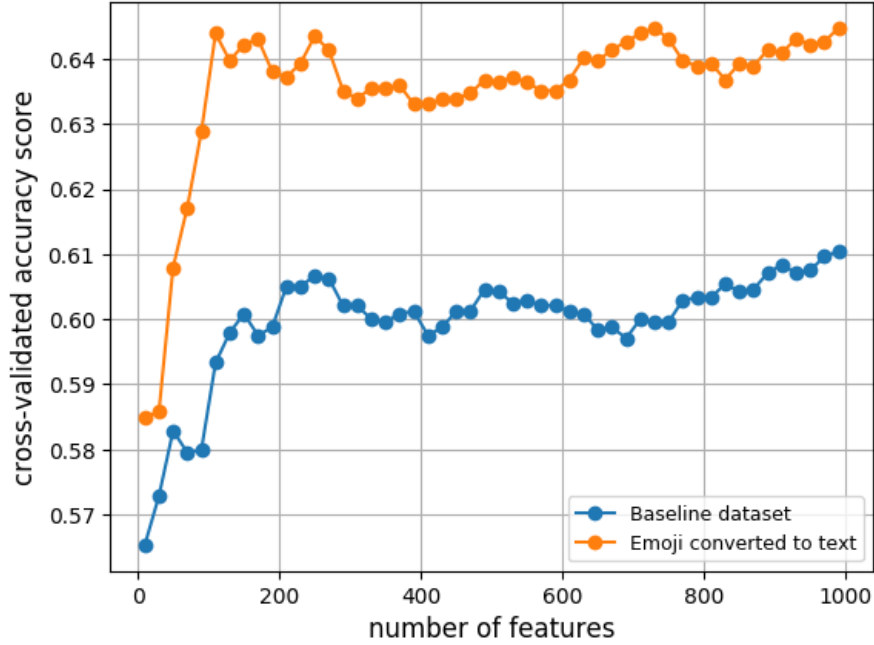


FIGURE 5.9: Baseline datasets average accuracy vs Emoji converted dataset average accuracy

Converting Emoticons into Text

Table B.5 under Appendix B.5 list down the accuracy score of each small dataset. When comparing the average of each dataset to the baseline datasets average we can identify an increase in accuracy. Figure 5.10 summarize the baseline datasets average accuracy and Emoticons converted dataset average accuracy.

When considering the large dataset, converting emoticons into text approaches' outcomes decrease the performance. In the large dataset, there was a slight decrease in performance. But when considering a small amount of dataset with a small number of features the outcome completely changed. The accuracy of the classification increased considerably with the Emoticon conversion approach.

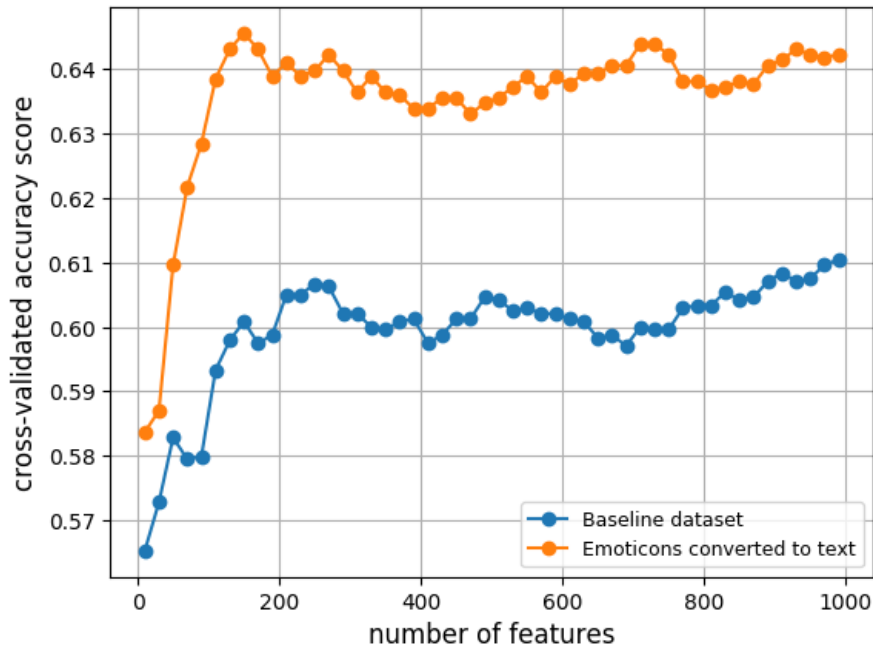


FIGURE 5.10: Baseline datasets average accuracy vs Emoticons converted dataset average accuracy

lemmatize

Table B.6 under Appendix B.6 list down the accuracy score of each small dataset when Lemmatizing approach is used. When analyzing the results, we can detect that the approach has increased the classification accuracy. This result also shows the opposite behaviour when comparing to the large dataset classification scores.

Figure 5.11 is the comparison between the baseline dataset average accuracy and Lemmatized dataset average accuracy. When considering the number of features, in general, Lemmatizing has increased the performance of the dataset.

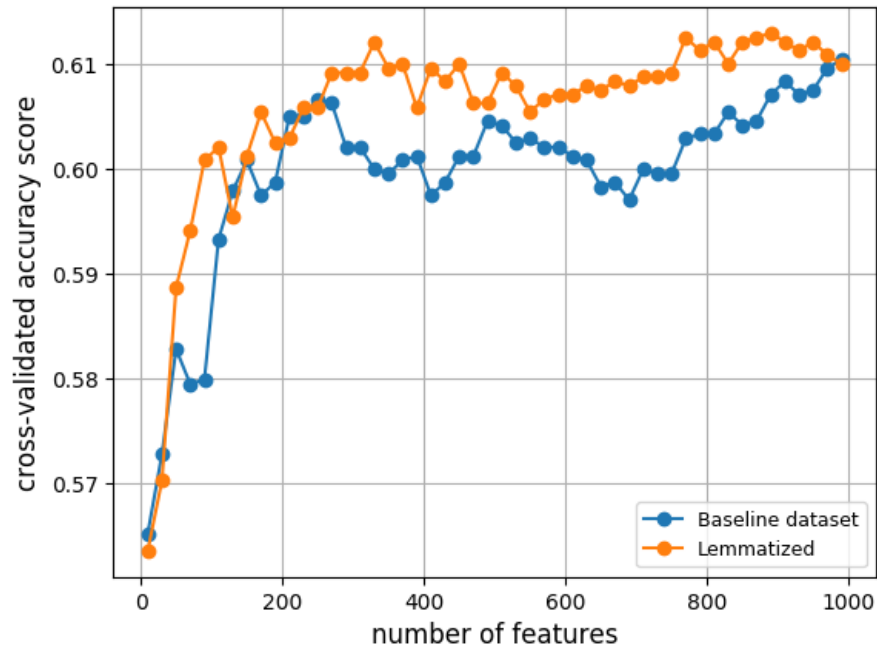


FIGURE 5.11: Baseline dataset average accuracy vs Lemmatized dataset average accuracy

Remove Numeric Characters

The Table B.7 under Appendix B.7 contains score for each small dataset and the average of the score when numeric characters are removed from the dataset. When considering the results, it is hard to say whether removing numeric characters gives a clear increase or a decrease in the performance of the classifier since the accuracy varies with the number of features used.

A comparison between baseline dataset average accuracy vs numeric characters removed dataset average accuracy is illustrated in Figure 5.12. Considering the large dataset, removing numeric characters increased the performance of classification.

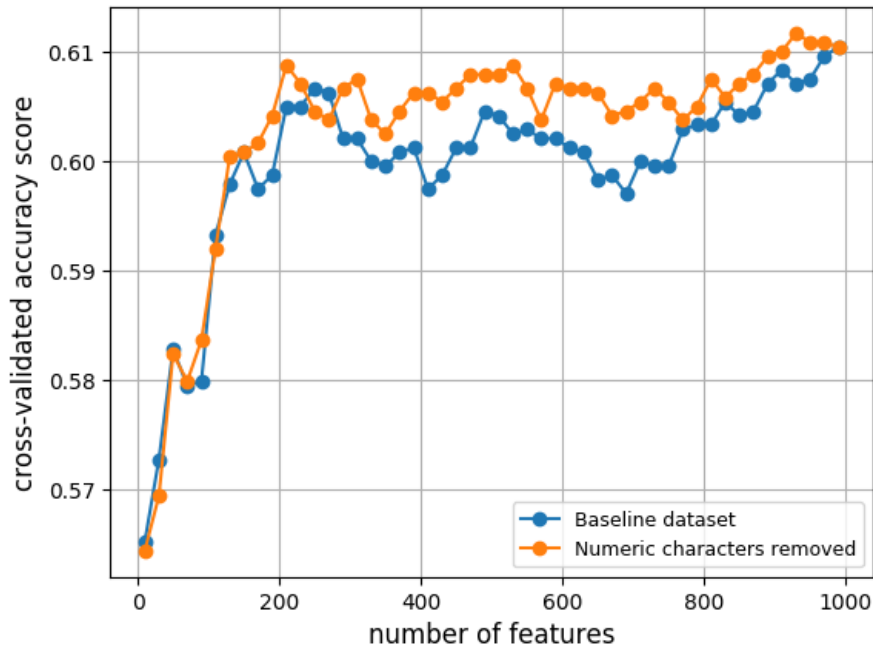


FIGURE 5.12: Baseline dataset average accuracy vs Numeric characters removed dataset average accuracy

When considering the small datasets, the stop word removing approach has decreased the accuracy of the classification performance. Therefore this approach was not selected as data cleaning approach in the implementation.

The effect of removing the numeric characters approach does not provide a clear positive or a negative impact on the classification. But when considering the literature in natural language processing and with the evaluation done against the large dataset, it is good to remove unnecessary numeric characters from the data. Therefore developed implementation uses Numeric character removing approach in the data cleaning.

Emoji converting, Emoticon converting and Lemmatizing approaches are included in the data cleaning process as those approaches increase the classification accuracy in the small datasets.

5.5 Comparison of Classifiers

I have used Naïve Bayes classification for the evaluation of data cleaning and preprocessing functions. In this section, I will compare the accuracy of multiple classifiers to select the best

classification for a small dataset. The baseline for the evaluation is the final cleaned dataset. According to the statistics I have derived in data preprocessing for a small dataset, only the removing stop words have a negative effect on the accuracy of the classification. Thus, the final dataset was cleaned and processed using, Expanding Contractions, Converting Emoji and Emoticons into the text, Lemmatizing and Numeric characters removing approaches. For the evaluation, 14 datasets were extracted from the base dataset. Each of these datasets contained 959 tweets. Each data set was classified using multiple classifiers. The average accuracy for each classifier was calculated using above 14 sub-datasets. K-fold cross-validation was used to calculate the accuracy of the classifiers. The selected K value was 10. Table 5.9 contains the result of average accuracy.

	MultinomialNB	BernoulliNB	SVC	LinearSVC l2	LogisticRegression l2	LinearSVC l1	LogisticRegression l1
Dataset 01	0.66155	0.63605	0.50526	0.61243	0.64187	0.61164	0.66687
Dataset 02	0.59605	0.56327	0.55746	0.57412	0.60023	0.57942	0.59635
Dataset 03	0.57357	0.58804	0.51053	0.57348	0.58351	0.5269	0.5983
Dataset 04	0.62898	0.54687	0.53635	0.62421	0.62392	0.5136	0.57208
Dataset 05	0.60506	0.56792	0.56792	0.59906	0.64617	0.62512	0.66167
Dataset 06	0.64079	0.54132	0.53105	0.60947	0.64053	0.64079	0.63474
Dataset 07	0.66673	0.54637	0.54164	0.67278	0.69804	0.67673	0.63512
Dataset 08	0.59319	0.54713	0.53635	0.57322	0.59845	0.54629	0.60368
Dataset 09	0.65661	0.63997	0.50526	0.65468	0.65082	0.61667	0.62442
Dataset 10	0.61406	0.55193	0.54693	0.62687	0.61295	0.59909	0.60713
Dataset 11	0.61026	0.56842	0.52105	0.58868	0.61447	0.54816	0.54263
Dataset 12	0.64143	0.56687	0.53635	0.61515	0.64149	0.60482	0.61874
Dataset 13	0.64877	0.63377	0.50526	0.68132	0.70237	0.70737	0.68626
Dataset 14	0.59038	0.56795	0.56269	0.54269	0.55933	0.58351	0.5743
Average Accuracy	0.62339	0.57613	0.53315	0.61058	0.62958	0.59858	0.61588

TABLE 5.9: Comparison of classifiers accuracy

According to the comparison, The Logistic regression classifier with penalty l2 has performed well in small datasets on average. The comparison is done between MultinomialNB, BernoulliNB, SVC, LinearSVC with l1 and l2 penalty, Logistic regression with l1 and l2 penalty. Figure 5.13 illustrates a summary of each of classification accuracy.

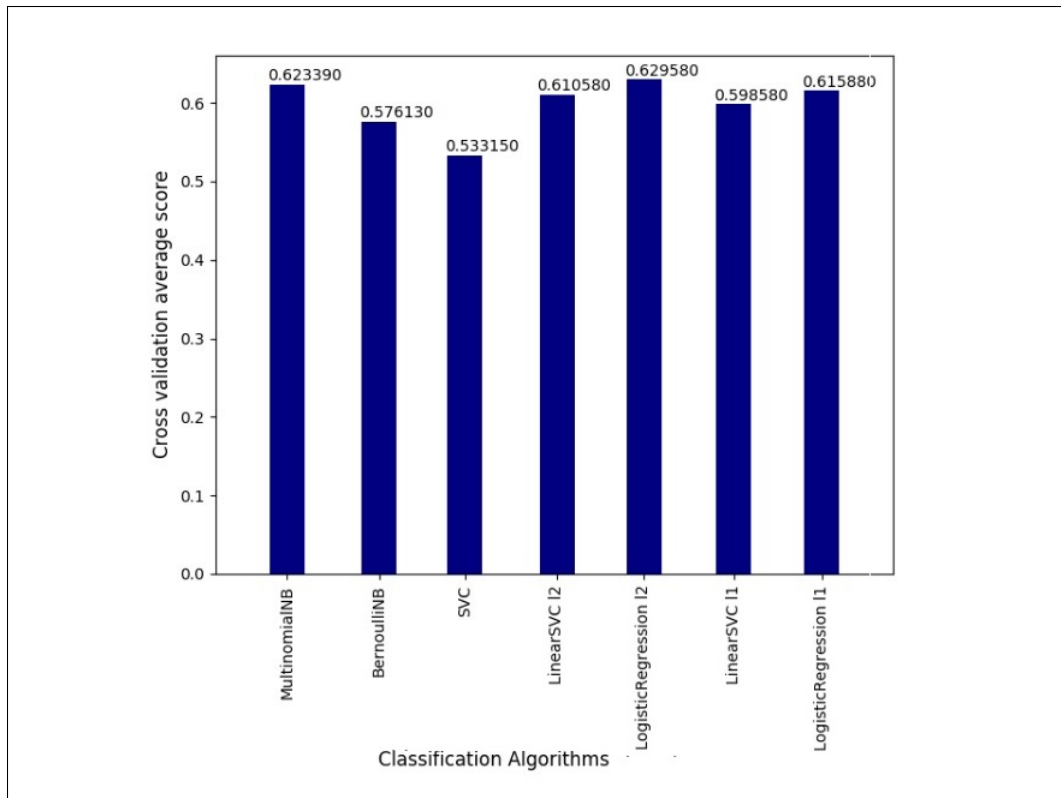


FIGURE 5.13: Classification accuracy summary

5.6 Improving Classifier Accuracy using TFIDF vectorization

Currently, the text using count vectorizer to convert text into the numeric form. TFIDF is another way of converting text to numeric form. I have selected the classifier as Logistic regression classifier and evaluated the results when TFIDF vectorization is used.

Eight datasets were randomly selected from the base to evaluate the impact of TFIDF vectorization approach. Each of those datasets was converted to numerical format using count vectorizer and TFIDF vectorizer and the results were plotted against the number of features used.

The Table B.8 on Appendix B.8 section containst the average classification accuracy score of the each dataset with using TFIDF vectorizer. The Table B.9 on Appendix B.9 section, list down the average accuracy of the classification with using only count vectorizer.

When comparing the average cross-validated accuracy, the figure shows using the TFIDF vectorizer provides a clear increase of accuracy. The Figure 5.14 contains a summary of the comparison.

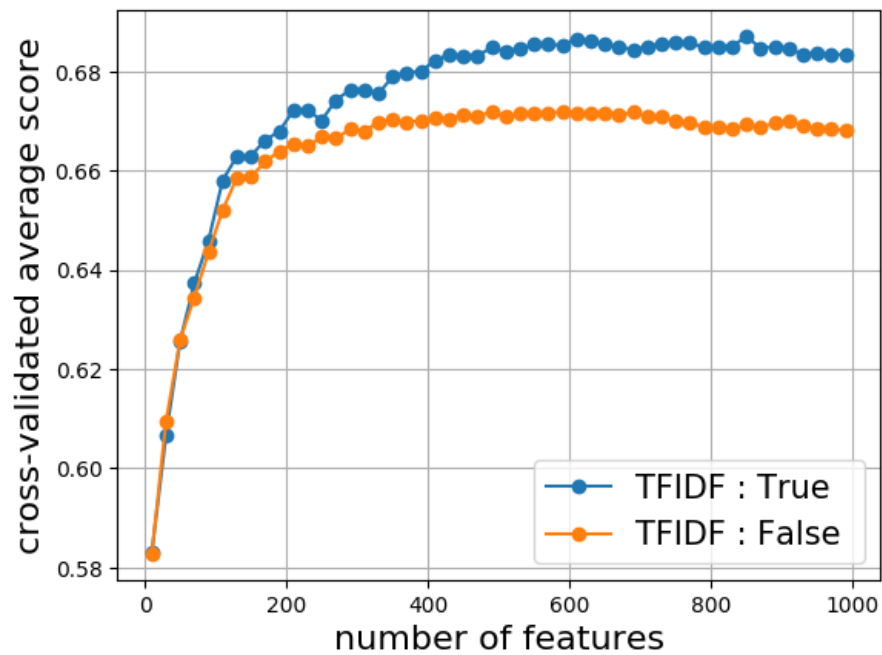


FIGURE 5.14: TFIDF vectorization average accuracy vs Count Vectorizer average accuracy

5.7 Selecting N-Gram Value

Selecting the best n-gram value is also crucial for the accuracy of the classification. Four random datasets containing 959 tweets were selected to evaluate the best N-gram value. The calculated average score for the n-gram value (1,1) (1,2) and (1,3) are listed in Appendix section B.10, Table B.10, B.11, B.12

The Figure 5.15 gives a summary of average accuracy. When comparing the three average scores, the figure clearly depicts that the highest performance was given by the n-gram value (1,1).

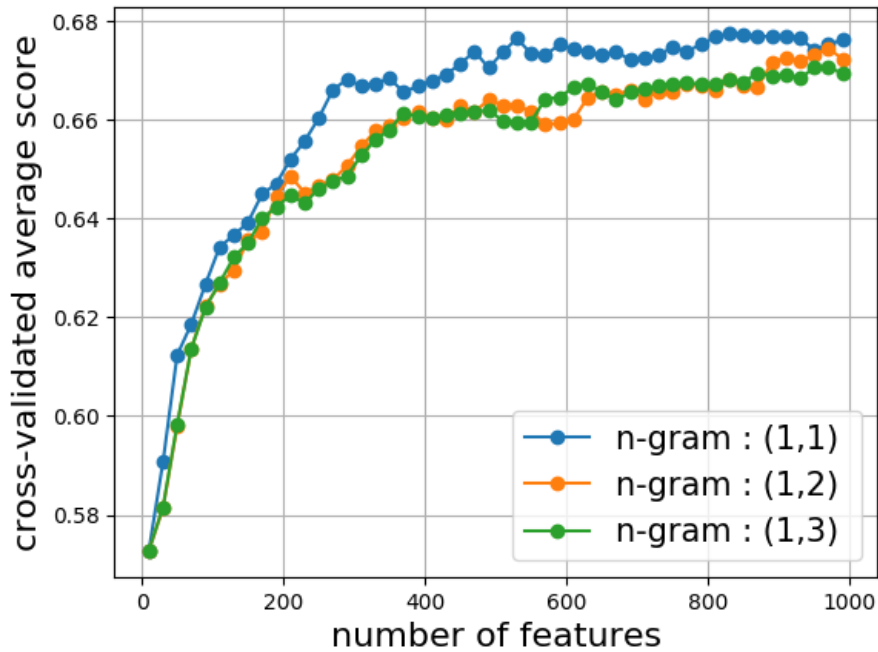
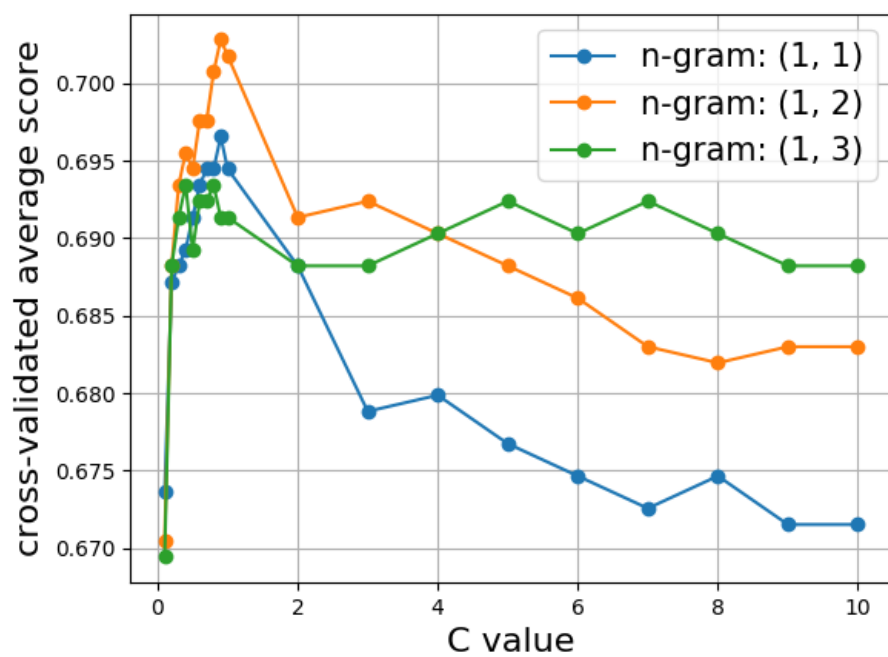


FIGURE 5.15: N-Gram average accuracy

5.8 Parameter Tuning in Classifier

The logistic regression classifier provides a parameter C that specifies stronger regularization. The parameter tuning is also related to the n-gram values because the regularization depends on the feature representation. To find the best fitting C value, a randomly selected dataset was evaluated.

When analysing the summary of the table B.13 listed in Appendix section B.11 displayed in Figure 5.16, it is clear that n-gram 2 with C value of 1.9 gives the highest accuracy for the classification.

FIGURE 5.16: Classifier accuracy vs C value

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The objective of this research was to build a computerized system to provide preferred tweets for twitter users. Twitter does not extract preferences of the Twitter users for the tweets populated in users feed. “Intelligent Twitter Agent” captures the user’s feedback for tweets using a mobile application. Then captured tweets are processed using data cleansing. Processed tweets were analyzed using classifiers to identify the preferences. Twitter dataset containing 1.6 million tweets were used in the analysis of the research as the base dataset. Six data preprocessing approaches were used to identify the data preprocessing approaches that increase the accuracy of classification results. Lemmatization, converting Emoji and Emoticons, Removing numeric and Expanding contractions approaches were identified as accuracy increasing approaches and used in the implementation of “Intelligent Twitter Agent”. To convert tweet texts into numerical representations, TFIDF was used as the vectorization method. The cleaned dataset was analyzed using different classifiers to discover the classification with the highest accuracy. From seven classifiers used, Logistic regression achieved approximate of 63% average classification accuracy when used 14 datasets with 959 tweets in each.

After the implementation is completed, “Intelligent Twitter Agent” achieved approximate 70% accuracy in identifying the user preference in small Twitter datasets containing around 900 tweets, when logistic regression classifier used with n-gram (1, 2) combination.

The outcome of this research will enhance the Twitter social media experience of Twitter users as they will receive preferred tweets in their feeds. Therefore unfollowing rates due to unwanted tweets of a Tweeter will be reduced.

6.2 Future Work

The developed computerized system is only support of analyzing the preferences of tweets from the English Language. But considering the Twitter users from many countries who post tweets from many languages, the research can be developed to identify the user preference for any language.

Twitter users use Images, Videos and GIF media content in tweets. In this development process, the content of the images was not taken into classification. The content of the images and the sentimental analysis based on the images can be also integrated into the system for a better performance in the classification. And it can be expanded into video processing as well.

Tweets contents third-party URLs, These URL contains metadata that can be extracted to gather information from those URLs. Those data can be used for classification in future and increase the classification accuracy.

The mobile application that was developed to gather the user preference only consider the "Thumbs-up" and "Thumbs-down inputs. The user preference gathering can be optimized by considering user behaviour also. Reading a tweet is also can be recognized as a preference and also replying for a tweet. These responses can be also ranked and be considered to evaluate the user preference.

The IFIDF implementation was used in the system was the basic TFIDF implementation. There are various kinds of TFIDF implementations that were not evaluated in the system implementation. That can be considered as a future work, Evaluating TFIDF implementations that is suitable for tweets and a small number of a dataset.

This system was developed based on a single user preference. It can be easily modified to evaluate multiple user preferences for a topic. For an example, we can select a topic such as elections, events and identify the user positive or the negative towards those kinds of topics.

Bibliography

- [1] *Character counting — twitter developers*, <https://dev.twitter.com/basics/counting-characters>, (Accessed on 07/16/2017).
- [2] *How did twitter start? what is the history behind twitter?*, http://profilerehab.com/twitter-help/history_of_twitter, (Accessed on 07/15/2017).
- [3] *Twitter: number of active users 2010-2017 | statista*, <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>, (Accessed on 07/15/2017).
- [4] S. Alhabash and M. Ma, “A tale of four platforms: motivations and uses of facebook, twitter, instagram, and snapchat among college students?”, *Social Media + Society*, vol. 3, no. 1, p. 2056305117691544, 2017. doi: 10.1177/2056305117691544. eprint: <https://doi.org/10.1177/2056305117691544>. [Online]. Available: <https://doi.org/10.1177/2056305117691544>.
- [5] *About different types of tweets*, <https://help.twitter.com/en/using-twitter/types-of-tweets>, (Accessed on 03/23/2018).
- [6] *About your twitter timeline*, <https://help.twitter.com/en/using-twitter/twitter-timeline>, (Accessed on 03/23/2018).
- [7] S. B. T. Statistics, *Natgeo twitter stats summary profile (social blade twitter statistics) - socialblade.com*, <https://socialblade.com/twitter/user/natgeo>, (Accessed on 03/23/2018).
- [8] ———, *Forbestech twitter stats summary profile (social blade twitter statistics) - socialblade.com*, <https://socialblade.com/twitter/user/forbestech>, (Accessed on 03/23/2018).
- [9] W. N. LOCKE, *Machine translation*, <http://mt-archive.info/Locke-1975.pdf>, (Accessed on 03/25/2018).

- [10] *Zipf's law: modeling the distribution of terms*, <https://nlp.stanford.edu/IR-book/html/htmledition/zipfs-law-modeling-the-distribution-of-terms-1.html>, (Accessed on 03/23/2018).
- [11] *Stemming and lemmatization*, <https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>, (Accessed on 03/20/2018).
- [12] *The stanford natural language processing group*, <https://nlp.stanford.edu/software/tagger.shtml>, (Accessed on 03/23/2018).
- [13] *Stanford corenlp*, <http://nlp.stanford.edu:8080/corenlp/process>, (Accessed on 03/25/2018).
- [14] J. H. M. Daniel Jurafsky, *Speech and language processing*, <http://www.cs.colorado.edu/~martin/SLP/Updates/1.pdf>, 2017.
- [15] *Using hashtags on twitter | twitter help center*, <https://support.twitter.com/articles/49309>, (Accessed on 07/15/2017).
- [16] Z. Ma, W. Dou, X. Wang, and S. Akella, "Tag-latent dirichlet allocation: understanding hashtags and their relationships", in *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, vol. 1, 2013, pp. 260–267. doi: 10.1109/WI-IAT.2013.38.
- [17] *Game of thrones (tv series 2011–) - imdb*, <http://www.imdb.com/title/tt0944947/>, (Accessed on 07/16/2017).
- [18] *9/11 attacks - facts & summary - history.com*, <http://www.history.com/topics/9-11-attacks>, (Accessed on 07/15/2017).
- [19] *How the emoticon was invented - business insider*, <http://www.businessinsider.com/how-the-emoticon-was-invented-2015-9>, (Accessed on 07/15/2017).
- [20] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak, "Exploiting emoticons in sentiment analysis", in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, ser. SAC '13, Coimbra, Portugal: ACM, 2013, pp. 703–710, ISBN: 978-1-4503-1656-9. doi: 10.1145/2480362.2480498. [Online]. Available: <http://doi.acm.org/10.1145/2480362.2480498>.

- [21] *Ellen degeneres on twitter: "if only bradley's arm was longer. best photo ever. #oscars* <http://t.co/c9u5notgap>", <https://twitter.com/theellenshow/status/440322224407314432?lang=en>, (Accessed on 07/06/2017).
- [22] *Ellen degeneres - imdb*, <http://www.imdb.com/name/nm0001122/>, (Accessed on 07/16/2017).
- [23] J. S. Hare, S. Samangoeei, D. P. Dupplaw, and P. H. Lewis, "Twitter's visual pulse", in *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, ser. ICMR '13, Dallas, Texas, USA: ACM, 2013, pp. 297–298, ISBN: 978-1-4503-2033-7. DOI: 10.1145/2461466.2461514. [Online]. Available: <http://doi.acm.org/10.1145/2461466.2461514>.
- [24] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web", in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10, Firenze, Italy: ACM, 2010, pp. 715–718, ISBN: 978-1-60558-933-6. DOI: 10.1145/1873951.1874060. [Online]. Available: <http://doi.acm.org/10.1145/1873951.1874060>.
- [25] Q. You, "Sentiment and emotion analysis for social multimedia: methodologies and applications", in *Proceedings of the 2016 ACM on Multimedia Conference*, ser. MM '16, Amsterdam, The Netherlands: ACM, 2016, pp. 1445–1449, ISBN: 978-1-4503-3603-1. DOI: 10.1145/2964284.2971475. [Online]. Available: <http://doi.acm.org/10.1145/2964284.2971475>.
- [26] J. Dumoulin, D. Affi, E. Mugellini, O. Abou Khaled, M. Bertini, and A. Del Bimbo, "Affect recognition in a realistic movie dataset using a hierarchical approach", in *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, ser. ASM '15, Brisbane, Australia: ACM, 2015, pp. 15–20, ISBN: 978-1-4503-3750-2. DOI: 10.1145/2813524.2813526. [Online]. Available: <http://doi.acm.org/10.1145/2813524.2813526>.
- [27] *Spam - definition of spam in english | oxford dictionaries*, <https://en.oxforddictionaries.com/definition/spam>, (Accessed on 07/16/2017).

- [28] raghavj, *Fighting spam with botmaker*, https://blog.twitter.com/engineering/en_us/a/2014/fighting-spam-with-botmaker.html, (Accessed on 08/26/2017), 2014.
- [29] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, “Design and evaluation of a real-time url spam filtering service”, in *Proceedings of the 2011 IEEE Symposium on Security and Privacy*, ser. SP ’11, Washington, DC, USA: IEEE Computer Society, 2011, pp. 447–462, ISBN: 978-0-7695-4402-1. DOI: 10.1109/SP.2011.25. [Online]. Available: <http://dx.doi.org/10.1109/SP.2011.25>.
- [30] Anita, D. Gupta, and A. Kumar, “Spam and sentiment analysis model for twitter data using statistical learning”, in *Proceedings of the Third International Symposium on Computer Vision and the Internet*, ser. VisionNet’16, Jaipur, India: ACM, 2016, pp. 54–58, ISBN: 978-1-4503-4301-5. DOI: 10.1145/2983402.2983404. [Online]. Available: <http://doi.acm.org/10.1145/2983402.2983404>.
- [31] T. Wu, S. Liu, J. Zhang, and Y. Xiang, “Twitter spam detection based on deep learning”, in *Proceedings of the Australasian Computer Science Week Multiconference*, ser. ACSW ’17, Geelong, Australia: ACM, 2017, 3:1–3:8, ISBN: 978-1-4503-4768-6. DOI: 10.1145/3014812.3014815. [Online]. Available: <http://doi.acm.org/10.1145/3014812.3014815>.
- [32] K. Lee, J. Caverlee, and S. Webb, “Uncovering social spammers: social honeypots + machine learning”, in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’10, Geneva, Switzerland: ACM, 2010, pp. 435–442, ISBN: 978-1-4503-0153-4. DOI: 10.1145/1835449.1835522. [Online]. Available: <http://doi.acm.org/10.1145/1835449.1835522>.
- [33] L. A. de Freitas, A. A. Vanin, D. N. Hogetop, M. N. Bochernitsan, and R. Vieira, “Pathways for irony detection in tweets”, in *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, ser. SAC ’14, Gyeongju, Republic of Korea: ACM, 2014, pp. 628–633, ISBN: 978-1-4503-2469-4. DOI: 10.1145/2554850.2555048. [Online]. Available: <http://doi.acm.org/10.1145/2554850.2555048>.

-
- [34] *Overview — twitter developers*, <https://developer.twitter.com/en/docs/tweets/timelines/overview>, (Accessed on 10/08/2017).
 - [35] *Spring social twitter*, <http://projects.spring.io/spring-social-twitter/>, (Accessed on 10/12/2017).
 - [36] *Twitter4j - a java library for the twitter api*, <http://twitter4j.org/en/>, (Accessed on 10/12/2017).
 - [37] *Rate limiting — twitter developers*, <https://developer.twitter.com/en/docs/basics/rate-limiting>, (Accessed on 10/08/2017).
 - [38] *Natural language toolkit — nltk 3.2.5 documentation*, <https://www.nltk.org/>, (Accessed on 03/17/2018).
 - [39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: machine learning in Python”, *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
 - [40] *3.4. model persistence — scikit-learn 0.19.1 documentation*, http://scikit-learn.org/stable/modules/model_persistence.html, (Accessed on 03/26/2018).
 - [41] *Emoji 0.4.5 : python package index*, <https://pypi.python.org/pypi/emoji/>, (Accessed on 03/20/2018).
 - [42] *Full emoji list, v11.0*, <http://www.unicode.org/emoji/charts/full-emoji-list.html>, (Accessed on 03/20/2018).
 - [43] *Emoji cheat sheet for github, basecamp and other services*, <https://www.webpagefx.com/tools/emoji-cheat-sheet/>, (Accessed on 03/20/2018).
 - [44] *20.1. html — hypertext markup language support python 3.6.5rc1 documentation*, <https://docs.python.org/3/library/html.html#html.unescape>, (Accessed on 03/20/2018).
 - [45] *Beautiful soup documentation — beautiful soup 4.4.0 documentation*, <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>, (Accessed on 03/20/2018).

-
- [46] *Contractions - english grammar today - cambridge dictionary*, <https://dictionary.cambridge.org/grammar/british-grammar/writing/contractions>, (Accessed on 03/20/2018).
 - [47] *Nltk.stem package — nltk 3.2.5 documentation*, <http://www.nltk.org/api/nltk.stem.html>, (Accessed on 03/20/2018).
 - [48] *5. categorizing and tagging words*, <http://www.nltk.org/book/ch05.html>, (Accessed on 03/20/2018).
 - [49] B. O'Connor, *Tweetmotif*, <https://github.com/brendano/tweetmotif>, 2010.
 - [50] *2. accessing text corpora and lexical resources*, <http://www.nltk.org/book/ch02.html>, (Accessed on 03/22/2018).
 - [51] L. H. Alec Go Richa Bhayani, "Twitter sentiment classification using distant supervision", in *Proceedings of the Third International Symposium on Computer Vision and the Internet*. [Online]. Available: <https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf>.
 - [52] *Choosing the right estimator — scikit-learn 0.19.1 documentation*, http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html, (Accessed on 03/26/2018).

Appendix A

Code Samples

A.1 Twitter API Timeline JSON Response

```
1  [  
2    {  
3      "coordinates": null ,  
4      "truncated": false ,  
5      "created_at": "Tue Aug 28 21:16:23 +0000 2012",  
6      "favorited": false ,  
7      "id_str": "240558470661799936",  
8      "in_reply_to_user_id_str": null ,  
9      "entities": {  
10       "urls": [  
11  
12       ],  
13       "hashtags": [  
14  
15       ],  
16       "user_mentions": [  
17  
18       ]  
19     },  
20     "text": "just another test",  
21     "contributors": null ,  
22     "id": 240558470661799936,  
23     "retweet_count": 0,  
24     "in_reply_to_status_id_str": null ,  
25     "geo": null ,  
26     "retweeted": false ,
```

```
27     "in_reply_to_user_id": null ,
28     "place": null ,
29     "source": "OAuth Dancer Reborn" ,
30     "user": {
31         "name": "OAuth Dancer" ,
32         "profile_sidebar_fill_color": "DDEEF6" ,
33         "profile_background_tile": true ,
34         "profile_sidebar_border_color": "CODEED" ,
35         "profile_image_url": "http://a0.twimg.com/profile_images/730275945/
oauth-dancer-normal.jpg" ,
36         "created_at": "Wed Mar 03 19:37:35 +0000 2010" ,
37         "location": "San Francisco , CA" ,
38         "follow_request_sent": false ,
39         "id_str": "119476949" ,
40         "is_translator": false ,
41         "profile_link_color": "0084B4" ,
42         "entities": {
43             "url": {
44                 "urls": [
45                     {
46                         "expanded_url": null ,
47                         "url": "http://bit.ly/oauth-dancer" ,
48                         "indices": [
49                             0,
50                             26
51                         ],
52                         "display_url": null
53                     }
54                 ]
55             },
56             "description": null
57         },
58         "default_profile": false ,
59         "url": "http://bit.ly/oauth-dancer" ,
60         "contributors_enabled": false ,
61         "favourites_count": 7,
62         "utc_offset": null ,
```

```
63     "profile_image_url_https": "https://si0.twimg.com/profile_images
64     /730275945/oauth-dancer_normal.jpg",
65     "id": 119476949,
66     "listed_count": 1,
67     "profile_use_background_image": true,
68     "profile_text_color": "333333",
69     "followers_count": 28,
70     "lang": "en",
71     "protected": false,
72     "geo_enabled": true,
73     "notifications": false,
74     "description": "",
75     "profile_background_color": "CODEED",
76     "verified": false,
77     "time_zone": null,
78     "profile_background_image_url_https": "https://si0.twimg.com/
79     profile_background_images/80151733/oauth-dance.png",
80     "statuses_count": 166,
81     "profile_background_image_url": "http://a0.twimg.com/
82     profile_background_images/80151733/oauth-dance.png",
83     "default_profile_image": false,
84     "friends_count": 14,
85     "following": false,
86     "show_all_inline_media": false,
87     "screen_name": "oauth_dancer"
88 },
89 "in_reply_to_screen_name": null,
90 "in_reply_to_status_id": null
91 }
```


Appendix B

Data Cleanning and pre-processing Evaluation

B.1 Accuracy on control datasets

Please refer Table B.1

B.2 Stopwords removed datasets accuracy vs Number of features

Please refer Table B.2

B.3 Contractions expanded datasets accuracy vs Number of features

Please refer Table B.3

B.4 Emoji converted datasets accuracy vs Number of features

Please refer Table B.4

B.5 Emoticons converted datasets accuracy vs Number of features

Please refer Table B.5

B.6 Lemmatized dataset accuracy vs Number of features

Please refer Table B.6

B.7 Numeric characters removed dataset accuracy vs Number of features

Please refer Table B.7

B.8 TFIDF vectorizer classification accuracy

Please refer Table B.8

B.9 Count vectorizer classification accuracy

Please refer Table B.9

B.10 N-Gram accuracy calculation

Please refer Tables B.10, B.11, B.12

B.11 Parameter tuning

Please refer Table B.13

No of Features	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Dataset 7	Dataset 8	Dataset 9	Dataset 10	Average Score
10	0.58159	0.57322	0.57741	0.59414	0.47699	0.55649	0.59833	0.56904	0.57322	0.55230	0.56527
30	0.60251	0.59414	0.56067	0.58159	0.51883	0.56067	0.61925	0.60669	0.52720	0.55649	0.57280
50	0.61925	0.55230	0.56904	0.59833	0.58159	0.54812	0.62343	0.60669	0.53975	0.58996	0.58285
70	0.61088	0.56067	0.60669	0.55230	0.53975	0.54393	0.63180	0.61506	0.56067	0.57322	0.57950
90	0.58996	0.56067	0.61088	0.57741	0.54393	0.53975	0.60669	0.62762	0.57322	0.56904	0.57992
110	0.58996	0.59414	0.62343	0.60251	0.55230	0.53138	0.62762	0.66109	0.56904	0.58159	0.59331
130	0.58159	0.59414	0.61088	0.60669	0.56904	0.54812	0.63598	0.66527	0.56904	0.59833	0.59791
150	0.58577	0.59833	0.62762	0.60669	0.57322	0.54393	0.64017	0.67364	0.57741	0.58159	0.60084
170	0.56067	0.58577	0.64017	0.60251	0.56067	0.56067	0.64854	0.66109	0.58996	0.56485	0.59749
190	0.56067	0.58159	0.64017	0.60251	0.57322	0.56067	0.62343	0.66109	0.60669	0.57741	0.59875
210	0.57322	0.58996	0.63180	0.61506	0.57741	0.56067	0.63598	0.66527	0.62343	0.57741	0.60502
230	0.57322	0.58159	0.62762	0.60669	0.58159	0.55649	0.66527	0.66527	0.62343	0.56904	0.60502
250	0.56067	0.60669	0.63180	0.61088	0.57741	0.56485	0.66527	0.66527	0.61506	0.56904	0.60669
270	0.56904	0.60251	0.64017	0.59414	0.58159	0.56485	0.65690	0.66109	0.61506	0.57741	0.60628
290	0.56904	0.58996	0.64017	0.59414	0.58996	0.55649	0.64854	0.64435	0.62343	0.56485	0.60209
310	0.58577	0.58577	0.63598	0.60251	0.59414	0.55649	0.63598	0.65272	0.62343	0.54812	0.60209
330	0.58159	0.58996	0.62343	0.60251	0.58996	0.56904	0.63180	0.64854	0.61088	0.55230	0.60000
350	0.58996	0.58159	0.61088	0.59833	0.59414	0.57322	0.62343	0.64854	0.61925	0.55649	0.59958
370	0.59414	0.58159	0.62343	0.58996	0.58996	0.58159	0.61925	0.65690	0.61088	0.56067	0.60084
390	0.58996	0.58996	0.62343	0.58159	0.58996	0.58159	0.61925	0.65690	0.61506	0.56485	0.60125
410	0.58577	0.58577	0.62762	0.57741	0.58577	0.57741	0.60669	0.65272	0.61088	0.56485	0.59749
430	0.58577	0.58996	0.63180	0.57741	0.58996	0.57741	0.60669	0.64854	0.61506	0.56485	0.59874
450	0.58577	0.58577	0.63598	0.58159	0.58577	0.58996	0.61506	0.65690	0.61506	0.56067	0.60125
470	0.58577	0.58577	0.63598	0.58577	0.58996	0.59414	0.61506	0.65272	0.60251	0.56485	0.60125
490	0.58577	0.59414	0.63598	0.58577	0.58577	0.60251	0.62343	0.65690	0.60251	0.57322	0.60460
510	0.58159	0.58577	0.64017	0.58996	0.58577	0.61506	0.61925	0.64854	0.61088	0.56485	0.60418
530	0.57741	0.58159	0.63180	0.58159	0.58996	0.61088	0.62343	0.66109	0.60669	0.56067	0.60251
550	0.57741	0.58159	0.62762	0.57741	0.59833	0.61088	0.61925	0.65690	0.62343	0.55649	0.60293
570	0.57741	0.57741	0.62762	0.56904	0.59414	0.62343	0.61925	0.65690	0.61925	0.55649	0.60209
590	0.57322	0.58159	0.63180	0.57322	0.58996	0.62343	0.62762	0.65690	0.61088	0.55230	0.60209
610	0.57741	0.58159	0.62343	0.56904	0.59414	0.62762	0.62343	0.66527	0.60251	0.54812	0.60126
630	0.57741	0.58159	0.61506	0.57322	0.58159	0.62762	0.63180	0.66527	0.60669	0.54812	0.60084
650	0.57741	0.58159	0.61088	0.56904	0.58577	0.62343	0.62762	0.66527	0.60669	0.53556	0.59833
670	0.57741	0.58159	0.61925	0.57322	0.57741	0.61925	0.62762	0.65690	0.61506	0.53975	0.59875
690	0.57741	0.58159	0.62343	0.56904	0.58577	0.62343	0.62343	0.63598	0.61506	0.53556	0.59707
710	0.57741	0.58577	0.62343	0.57322	0.59414	0.62343	0.62762	0.63598	0.61506	0.54393	0.60000
730	0.57741	0.58159	0.62762	0.56067	0.59833	0.62343	0.62343	0.64017	0.61506	0.54812	0.59958
750	0.57741	0.58159	0.62343	0.56485	0.59833	0.62762	0.61925	0.64017	0.61925	0.54393	0.59958
770	0.57741	0.58577	0.62343	0.56904	0.59833	0.62762	0.63180	0.63598	0.62762	0.55230	0.60293
790	0.58159	0.58159	0.61925	0.56904	0.59414	0.62762	0.63180	0.64017	0.62762	0.56067	0.60335
810	0.58996	0.57741	0.61925	0.57741	0.59833	0.62343	0.62762	0.63598	0.62762	0.55649	0.60335
830	0.59414	0.58577	0.61925	0.57741	0.60251	0.62343	0.62343	0.63180	0.63180	0.56485	0.60544
850	0.58996	0.58159	0.62343	0.57741	0.59833	0.61088	0.63180	0.62762	0.62762	0.57322	0.60419
870	0.58577	0.58577	0.62343	0.57741	0.60251	0.61506	0.63180	0.62343	0.63180	0.56904	0.60460
890	0.58577	0.60251	0.61925	0.57741	0.60669	0.61506	0.63180	0.62343	0.63180	0.57741	0.60711
910	0.58159	0.60251	0.62343	0.57741	0.61088	0.61506	0.63180	0.62343	0.63598	0.58159	0.60837
930	0.58159	0.60669	0.62343	0.57741	0.60251	0.61088	0.62762	0.62343	0.64017	0.57741	0.60711
950	0.58159	0.60669	0.61925	0.57322	0.60669	0.62343	0.62762	0.61506	0.64435	0.57741	0.60753
970	0.58159	0.61088	0.61925	0.57741	0.60251	0.63180	0.62343	0.61925	0.64854	0.58159	0.60963
990	0.57741	0.61506	0.62343	0.57741	0.60669	0.64017	0.62343	0.61506	0.64854	0.57741	0.61046

TABLE B.1: Accuracy on control datasets

No of Features	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Dataset 7	Dataset 8	Dataset 9	Dataset 10	Average Score
10	0.54393	0.53556	0.54393	0.48117	0.54393	0.54812	0.56485	0.52720	0.52720	0.56485	0.53807
30	0.51883	0.57741	0.51464	0.54393	0.51883	0.54393	0.53138	0.59833	0.58159	0.58159	0.55105
50	0.53138	0.56067	0.57741	0.58159	0.54393	0.51883	0.53975	0.62762	0.56904	0.57322	0.56234
70	0.55649	0.57322	0.56485	0.55649	0.55230	0.53556	0.54812	0.64017	0.59833	0.57322	0.56987
90	0.55230	0.54393	0.59414	0.57322	0.58159	0.55649	0.55649	0.65690	0.60669	0.58159	0.58033
110	0.56067	0.57322	0.59414	0.58159	0.60251	0.57322	0.55230	0.65690	0.62343	0.57741	0.58954
130	0.55649	0.56485	0.59414	0.58159	0.59414	0.58159	0.55230	0.65690	0.61506	0.56067	0.58577
150	0.56067	0.55230	0.60251	0.59414	0.58996	0.58159	0.54812	0.64017	0.62343	0.56485	0.58577
170	0.55230	0.57322	0.58996	0.60251	0.61088	0.57322	0.54812	0.63180	0.62343	0.57322	0.58787
190	0.55649	0.58159	0.58996	0.60251	0.60669	0.56904	0.55230	0.63598	0.63180	0.57741	0.59038
210	0.54812	0.58577	0.58577	0.60251	0.61925	0.56904	0.55649	0.62762	0.63598	0.58159	0.59121
230	0.56904	0.59414	0.56485	0.59833	0.62343	0.56904	0.53556	0.62762	0.63180	0.56904	0.58829
250	0.56904	0.59414	0.57322	0.58996	0.62343	0.56067	0.53975	0.62762	0.63180	0.56485	0.58745
270	0.56904	0.59833	0.57322	0.60251	0.61925	0.56904	0.53975	0.61506	0.62762	0.57322	0.58870
290	0.56904	0.60251	0.57322	0.59414	0.62762	0.56485	0.53556	0.60669	0.63180	0.57322	0.58786
310	0.56904	0.59414	0.57741	0.59833	0.62343	0.57741	0.53556	0.60251	0.64017	0.57322	0.58912
330	0.56485	0.58159	0.58577	0.59833	0.62343	0.58159	0.53556	0.61088	0.63598	0.57322	0.58912
350	0.56485	0.58159	0.58996	0.59833	0.62343	0.58159	0.52720	0.61088	0.63180	0.56485	0.58745
370	0.56904	0.58159	0.58577	0.59833	0.61925	0.58159	0.53138	0.60251	0.62762	0.56485	0.58619
390	0.57322	0.58159	0.58577	0.59833	0.61925	0.57741	0.53975	0.59414	0.62762	0.56485	0.58619
410	0.57322	0.58159	0.59833	0.58996	0.62762	0.58159	0.54812	0.59414	0.62762	0.56904	0.58912
430	0.56067	0.57741	0.60251	0.58996	0.61925	0.58577	0.53975	0.59414	0.62762	0.57741	0.58745
450	0.55649	0.58159	0.58996	0.58159	0.61506	0.58159	0.53138	0.59833	0.63180	0.58159	0.58494
470	0.56485	0.58159	0.58996	0.58159	0.60669	0.58159	0.53975	0.59414	0.63180	0.58159	0.58536
490	0.55649	0.57741	0.58996	0.57322	0.61088	0.58577	0.53556	0.59414	0.63180	0.58577	0.58410
510	0.54812	0.58577	0.59414	0.57322	0.60251	0.58577	0.53138	0.59414	0.63180	0.57741	0.58243
530	0.54393	0.57741	0.59833	0.56904	0.61088	0.58996	0.53975	0.59833	0.63180	0.57322	0.58327
550	0.54812	0.57741	0.60669	0.56485	0.61088	0.59414	0.54812	0.60251	0.62762	0.57741	0.58578
570	0.54812	0.57741	0.60669	0.56904	0.62343	0.58577	0.56067	0.60251	0.62762	0.57741	0.58787
590	0.54812	0.57322	0.60251	0.57322	0.61506	0.58996	0.55230	0.60251	0.62762	0.56485	0.58494
610	0.53975	0.56904	0.59833	0.56904	0.61925	0.58577	0.55230	0.60669	0.63598	0.57322	0.58494
630	0.53975	0.57322	0.58996	0.56904	0.62343	0.58577	0.55230	0.59833	0.63180	0.57322	0.58368
650	0.54393	0.57322	0.59833	0.56904	0.62762	0.58996	0.55230	0.58996	0.63180	0.57322	0.58494
670	0.53975	0.57741	0.60251	0.56485	0.61925	0.59414	0.55649	0.58996	0.63180	0.57741	0.58536
690	0.52301	0.58577	0.60669	0.56485	0.61925	0.59414	0.55230	0.58996	0.63180	0.58159	0.58494
710	0.52301	0.58159	0.60669	0.56485	0.61925	0.59833	0.55230	0.58996	0.64017	0.58577	0.58619
730	0.52301	0.57741	0.60669	0.56904	0.61925	0.59833	0.55649	0.58996	0.64017	0.58577	0.58661
750	0.51464	0.57741	0.60251	0.57322	0.61925	0.58996	0.55230	0.58996	0.64435	0.58159	0.58452
770	0.51046	0.58159	0.60251	0.56904	0.62343	0.58577	0.55649	0.57322	0.64017	0.58577	0.58285
790	0.49791	0.58159	0.60251	0.58159	0.61925	0.58996	0.56067	0.58159	0.64017	0.58577	0.58410
810	0.49791	0.58996	0.60251	0.58996	0.61506	0.59414	0.55649	0.57741	0.63598	0.58577	0.58452
830	0.49791	0.58996	0.61506	0.58577	0.61088	0.59833	0.55230	0.57741	0.64017	0.58577	0.58536
850	0.49372	0.58996	0.61925	0.58159	0.61088	0.59414	0.54812	0.56904	0.64017	0.58996	0.58368
870	0.48954	0.59833	0.61925	0.56485	0.61506	0.59414	0.54812	0.58159	0.63180	0.58577	0.58285
890	0.48954	0.58996	0.61506	0.56067	0.61088	0.59414	0.55230	0.58577	0.63180	0.58996	0.58201
910	0.48954	0.58577	0.62762	0.55230	0.61925	0.59414	0.55649	0.57741	0.63180	0.58996	0.58243
930	0.49791	0.58996	0.62343	0.55649	0.61925	0.59833	0.56904	0.57322	0.63180	0.59414	0.58536
950	0.50209	0.59414	0.61088	0.56067	0.61088	0.59833	0.56904	0.57322	0.61925	0.60251	0.58410
970	0.50628	0.60251	0.62343	0.55649	0.60669	0.59833	0.56904	0.56904	0.62762	0.59833	0.58578
990	0.51464	0.59833	0.62762	0.55649	0.60251	0.60251	0.57741	0.56904	0.61506	0.59833	0.58619

TABLE B.2: Stopwords removed datasets accuracy vs Number of features

No of Features	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Dataset 7	Dataset 8	Dataset 9	Dataset 10	Average Score
10	0.62762	0.58577	0.56485	0.56485	0.56485	0.60669	0.61088	0.58996	0.61925	0.53138	0.58661
30	0.66109	0.57322	0.55649	0.52720	0.63598	0.59833	0.57322	0.59833	0.57322	0.64854	0.59456
50	0.66946	0.58577	0.61925	0.53138	0.61088	0.62762	0.58159	0.63180	0.56485	0.62343	0.60460
70	0.68619	0.58996	0.64017	0.55230	0.61506	0.64435	0.56904	0.64435	0.59833	0.61925	0.61590
90	0.71130	0.59833	0.64435	0.58996	0.62762	0.65690	0.59833	0.64017	0.59414	0.62343	0.62845
110	0.70293	0.61088	0.64854	0.59414	0.62762	0.63598	0.60669	0.66946	0.62762	0.63598	0.63598
130	0.71130	0.61088	0.64854	0.60251	0.66527	0.63598	0.58577	0.65272	0.62762	0.66946	0.64101
150	0.72385	0.61088	0.64017	0.58577	0.66527	0.61925	0.56485	0.64435	0.62762	0.66109	0.63431
170	0.72803	0.62343	0.63180	0.58577	0.64017	0.62343	0.56485	0.64854	0.65690	0.68619	0.63891
190	0.71548	0.62343	0.63180	0.57741	0.63180	0.61925	0.55649	0.64435	0.65272	0.68619	0.63389
210	0.71548	0.63598	0.63180	0.58996	0.63598	0.60251	0.57741	0.65272	0.67364	0.68619	0.64017
230	0.70293	0.63598	0.62762	0.58159	0.64854	0.61088	0.57741	0.64854	0.66527	0.69456	0.63933
250	0.70293	0.64017	0.62762	0.58159	0.65272	0.61088	0.58577	0.64435	0.66109	0.69038	0.63975
270	0.69456	0.64017	0.62762	0.57322	0.65272	0.60251	0.59833	0.64854	0.66109	0.69038	0.63891
290	0.69456	0.64017	0.63598	0.57741	0.64435	0.60669	0.58577	0.66109	0.65272	0.69456	0.63933
310	0.71130	0.62762	0.62343	0.58159	0.64435	0.60251	0.57741	0.65690	0.66527	0.69874	0.63891
330	0.71130	0.61925	0.61506	0.58577	0.64854	0.59833	0.58577	0.65272	0.66527	0.69456	0.63766
350	0.70711	0.61925	0.61925	0.57322	0.64854	0.60251	0.58577	0.66527	0.66946	0.69874	0.63891
370	0.70293	0.61925	0.61506	0.57322	0.65690	0.60669	0.58577	0.66109	0.66946	0.68201	0.63724
390	0.71130	0.61925	0.61088	0.57741	0.66109	0.59414	0.58996	0.65272	0.66527	0.69038	0.63724
410	0.71130	0.61925	0.61088	0.57322	0.66527	0.59414	0.58996	0.64854	0.66527	0.68201	0.63598
430	0.71130	0.61506	0.61088	0.57322	0.66946	0.59414	0.58577	0.64017	0.66527	0.68619	0.63515
450	0.71548	0.61925	0.61506	0.56904	0.66109	0.59414	0.58577	0.63598	0.66527	0.67782	0.63389
470	0.71548	0.61925	0.61506	0.56904	0.66109	0.59414	0.58577	0.64017	0.66946	0.68619	0.63556
490	0.71130	0.61925	0.61925	0.57322	0.66109	0.60251	0.58996	0.63180	0.66946	0.69456	0.63724
510	0.71130	0.61925	0.61088	0.58159	0.65690	0.59833	0.58996	0.62762	0.66946	0.69874	0.63640
530	0.71548	0.62762	0.61088	0.57322	0.65272	0.59833	0.58577	0.62762	0.66946	0.69874	0.63598
550	0.71548	0.63598	0.61506	0.57741	0.66109	0.60251	0.58577	0.63180	0.66946	0.70711	0.64017
570	0.71548	0.62762	0.61925	0.57741	0.66109	0.60669	0.58996	0.62762	0.66946	0.69456	0.63891
590	0.71548	0.63180	0.62762	0.58159	0.66527	0.61925	0.58996	0.63598	0.66946	0.69038	0.64268
610	0.71967	0.62343	0.62343	0.57741	0.66109	0.61925	0.58996	0.63180	0.66946	0.69038	0.64059
630	0.71130	0.63180	0.61925	0.57322	0.66527	0.62343	0.59414	0.63598	0.66109	0.69038	0.64059
650	0.71967	0.62762	0.61925	0.57741	0.66527	0.63180	0.60251	0.63180	0.66527	0.69456	0.64352
670	0.71967	0.62762	0.61506	0.56904	0.66109	0.63180	0.60669	0.62762	0.66946	0.69456	0.64226
690	0.71967	0.62343	0.61925	0.56485	0.66527	0.63180	0.60669	0.63180	0.66946	0.69874	0.64310
710	0.71967	0.61925	0.61506	0.56485	0.66109	0.63598	0.60669	0.64017	0.67364	0.69874	0.64351
730	0.71548	0.62343	0.61088	0.56485	0.66946	0.62343	0.60669	0.64435	0.67364	0.69456	0.64268
750	0.71130	0.62762	0.60669	0.57322	0.66527	0.62343	0.60669	0.63598	0.67364	0.69456	0.64184
770	0.71548	0.62762	0.61506	0.58159	0.66109	0.62343	0.60669	0.62343	0.67364	0.69874	0.64268
790	0.71548	0.62762	0.61088	0.56904	0.66527	0.62762	0.59833	0.61925	0.67782	0.69456	0.64059
810	0.71548	0.62343	0.61088	0.57741	0.66527	0.62762	0.59833	0.61925	0.68201	0.69874	0.64184
830	0.71548	0.62762	0.61088	0.58159	0.66527	0.62762	0.59414	0.61925	0.67782	0.69456	0.64142
850	0.71548	0.62762	0.61088	0.59414	0.66527	0.62343	0.59414	0.62343	0.67782	0.69456	0.64268
870	0.70711	0.62343	0.61925	0.59833	0.66109	0.63598	0.59833	0.62762	0.67782	0.69874	0.64477
890	0.71130	0.62343	0.61506	0.60251	0.66946	0.62762	0.60251	0.63180	0.68619	0.69874	0.64686
910	0.70711	0.62343	0.61925	0.59833	0.66946	0.62343	0.60251	0.63598	0.68619	0.69456	0.64602
930	0.70293	0.62343	0.61925	0.60669	0.66946	0.62762	0.60251	0.63598	0.68619	0.69874	0.64728
950	0.70293	0.61925	0.61925	0.60251	0.66527	0.62762	0.61088	0.63598	0.68201	0.69456	0.64603
970	0.69874	0.61925	0.61925	0.59833	0.66527	0.62762	0.60251	0.63598	0.68201	0.69456	0.64435
990	0.70293	0.61088	0.61506	0.60251	0.66109	0.62762	0.61088	0.63598	0.69038	0.69456	0.64519

TABLE B.3: Contractions expanded datasets accuracy vs Number of features

No of Features	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Dataset 7	Dataset 8	Dataset 9	Dataset 10	Average Score
10	0.61088	0.57322	0.56485	0.54393	0.56485	0.59414	0.60251	0.62343	0.60669	0.56485	0.58493
30	0.66527	0.61088	0.56485	0.51464	0.59833	0.55230	0.55230	0.57322	0.58577	0.64017	0.58577
50	0.64854	0.61088	0.58996	0.52301	0.60669	0.62343	0.62343	0.63180	0.58577	0.63598	0.60795
70	0.68619	0.61088	0.61925	0.53975	0.58577	0.61506	0.61925	0.66527	0.60251	0.62762	0.61715
90	0.69038	0.60251	0.62343	0.58577	0.57741	0.63180	0.63180	0.67782	0.61506	0.65272	0.62887
110	0.71548	0.63180	0.64017	0.59833	0.61925	0.61925	0.62343	0.67364	0.64435	0.67364	0.64393
130	0.70711	0.61925	0.64435	0.59414	0.61088	0.63180	0.60669	0.67782	0.62762	0.67782	0.63975
150	0.69874	0.64435	0.64854	0.58159	0.61506	0.62762	0.60251	0.67782	0.64854	0.67782	0.64226
170	0.71130	0.64854	0.64435	0.56904	0.61506	0.62343	0.59833	0.67364	0.66109	0.68619	0.64310
190	0.71130	0.63180	0.63598	0.56485	0.61506	0.62343	0.57741	0.67364	0.65690	0.69038	0.63808
210	0.70293	0.63598	0.63180	0.57741	0.61506	0.60669	0.58996	0.66946	0.66109	0.68201	0.63724
230	0.70711	0.63180	0.62343	0.57741	0.61925	0.60669	0.59833	0.67364	0.67364	0.68201	0.63933
250	0.70293	0.63598	0.63180	0.57741	0.63180	0.61506	0.60251	0.68619	0.67364	0.67782	0.64351
270	0.71130	0.63598	0.64017	0.57322	0.62343	0.60251	0.60251	0.69038	0.65272	0.68201	0.64142
290	0.71130	0.62762	0.63598	0.56904	0.62343	0.60251	0.57741	0.68619	0.64854	0.66946	0.63515
310	0.70711	0.62343	0.62762	0.57322	0.63598	0.60251	0.58996	0.66527	0.64435	0.66946	0.63389
330	0.70293	0.62762	0.63598	0.57741	0.63180	0.59414	0.60669	0.66527	0.64854	0.66527	0.63557
350	0.70293	0.62343	0.64435	0.58577	0.63180	0.59414	0.59833	0.66109	0.64854	0.66527	0.63556
370	0.70711	0.61925	0.64017	0.58996	0.63180	0.59414	0.60251	0.66109	0.65272	0.66109	0.63598
390	0.69874	0.62343	0.64017	0.58577	0.64017	0.57322	0.59833	0.65690	0.64854	0.66527	0.63305
410	0.70293	0.61925	0.64435	0.56904	0.64854	0.57741	0.59414	0.66109	0.64854	0.66527	0.63306
430	0.71548	0.62343	0.64017	0.56904	0.65272	0.57741	0.59414	0.64854	0.65272	0.66527	0.63389
450	0.71548	0.62762	0.64017	0.56485	0.64854	0.58577	0.59414	0.64854	0.64854	0.66527	0.63389
470	0.71548	0.61925	0.63598	0.57322	0.64435	0.59414	0.59414	0.65690	0.64854	0.66527	0.63473
490	0.71548	0.61925	0.63180	0.57741	0.64854	0.60251	0.58996	0.66109	0.64854	0.67364	0.63682
510	0.71130	0.61925	0.63598	0.57322	0.64854	0.60251	0.58996	0.65690	0.64854	0.67782	0.63640
530	0.71130	0.61506	0.63598	0.57322	0.64854	0.61088	0.58996	0.66109	0.64854	0.67782	0.63724
550	0.70711	0.61925	0.64017	0.56904	0.64854	0.61088	0.58159	0.66527	0.65272	0.66946	0.63640
570	0.71130	0.61925	0.64017	0.56904	0.64435	0.60251	0.58159	0.66109	0.65690	0.66527	0.63515
590	0.71130	0.61088	0.64435	0.57741	0.63598	0.60669	0.58159	0.66527	0.65690	0.66109	0.63515
610	0.71130	0.61088	0.64435	0.58159	0.63598	0.61506	0.58577	0.66527	0.65690	0.66109	0.63682
630	0.71548	0.61506	0.65272	0.58577	0.64017	0.61506	0.58996	0.66527	0.65690	0.66527	0.64017
650	0.72803	0.60251	0.64854	0.58577	0.64017	0.61088	0.58996	0.66527	0.65690	0.66946	0.63975
670	0.72803	0.60251	0.64435	0.58996	0.64017	0.61506	0.59414	0.66109	0.66527	0.67364	0.64142
690	0.72803	0.60669	0.64435	0.58577	0.63598	0.62343	0.60251	0.66109	0.66109	0.67782	0.64268
710	0.72803	0.60669	0.64435	0.58577	0.64854	0.62343	0.60251	0.65690	0.66109	0.68201	0.64393
730	0.72385	0.61088	0.64854	0.58996	0.64435	0.62762	0.60251	0.66109	0.66109	0.67782	0.64477
750	0.71967	0.61506	0.64435	0.58159	0.64854	0.62343	0.59833	0.65690	0.66109	0.68201	0.64310
770	0.71548	0.61506	0.64017	0.58996	0.64854	0.61088	0.59414	0.64017	0.66109	0.68201	0.63975
790	0.71130	0.61506	0.64017	0.58577	0.64854	0.61506	0.59414	0.63598	0.65690	0.68619	0.63891
810	0.71130	0.61925	0.63598	0.58996	0.64854	0.61506	0.59414	0.63598	0.65690	0.68619	0.63933
830	0.71130	0.61506	0.63598	0.58159	0.64854	0.61506	0.58159	0.63598	0.65272	0.69038	0.63682
850	0.71548	0.61506	0.63598	0.59414	0.64854	0.62343	0.59414	0.63180	0.64435	0.69038	0.63933
870	0.71967	0.61506	0.63598	0.59414	0.64017	0.62762	0.59414	0.63598	0.64435	0.68201	0.63891
890	0.71548	0.61506	0.63180	0.59414	0.65272	0.62762	0.58996	0.64017	0.65690	0.69038	0.64142
910	0.71548	0.61506	0.63180	0.59833	0.65690	0.62343	0.58996	0.64017	0.65690	0.68201	0.64100
930	0.71130	0.61506	0.62762	0.61088	0.65690	0.63180	0.59414	0.64017	0.66109	0.68201	0.64310
950	0.70711	0.61088	0.63180	0.61506	0.65690	0.62762	0.59414	0.64017	0.66109	0.67782	0.64226
970	0.71130	0.61088	0.63180	0.61088	0.65690	0.63180	0.58996	0.64017	0.66527	0.67782	0.64268
990	0.70711	0.61088	0.63180	0.61088	0.65690	0.63598	0.59833	0.64854	0.66946	0.67782	0.64477

TABLE B.4: Emoj converted datasets accuracy vs Number of features

No of Features	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Dataset 7	Dataset 8	Dataset 9	Dataset 10	Average Score
10	0.61506	0.57322	0.56904	0.54393	0.56485	0.59414	0.60251	0.60251	0.60669	0.56485	0.58368
30	0.66527	0.60251	0.56485	0.51883	0.60251	0.55649	0.55230	0.57741	0.59414	0.63598	0.58703
50	0.65272	0.60669	0.59414	0.52301	0.61088	0.62343	0.62343	0.64435	0.58159	0.63598	0.60962
70	0.67782	0.62343	0.62343	0.56067	0.58159	0.61925	0.61925	0.66946	0.61088	0.63180	0.62176
90	0.69038	0.61506	0.63180	0.56485	0.58159	0.63180	0.63180	0.67782	0.62343	0.63598	0.62845
110	0.70711	0.62343	0.64435	0.58577	0.60251	0.61506	0.62343	0.66946	0.64017	0.67364	0.63849
130	0.70293	0.63598	0.64854	0.59833	0.61088	0.63180	0.60669	0.68201	0.64017	0.67364	0.64310
150	0.69874	0.63180	0.64017	0.60251	0.61506	0.63180	0.60251	0.67782	0.67782	0.67782	0.64560
170	0.70711	0.63180	0.64017	0.59414	0.61088	0.63598	0.59833	0.67364	0.66109	0.67782	0.64310
190	0.71548	0.63180	0.63180	0.57322	0.61506	0.61925	0.57741	0.67364	0.66527	0.68619	0.63891
210	0.70711	0.64435	0.63180	0.58159	0.62762	0.61925	0.58996	0.67364	0.64854	0.68619	0.64100
230	0.70711	0.65272	0.63598	0.56904	0.60669	0.60251	0.59833	0.67364	0.64854	0.69456	0.63891
250	0.69874	0.65272	0.63598	0.56485	0.61506	0.60251	0.60251	0.67782	0.64435	0.70293	0.63975
270	0.71548	0.64435	0.65272	0.56904	0.61506	0.61088	0.60251	0.67364	0.64435	0.69456	0.64226
290	0.71130	0.65690	0.64854	0.57741	0.61506	0.61925	0.57741	0.66527	0.64435	0.68201	0.63975
310	0.70293	0.64435	0.62762	0.57741	0.63598	0.60669	0.58996	0.66946	0.64435	0.66527	0.63640
330	0.70711	0.63180	0.63598	0.57741	0.63180	0.61088	0.60669	0.67364	0.64854	0.66527	0.63891
350	0.69874	0.62343	0.64435	0.58577	0.63180	0.60669	0.59833	0.66109	0.64435	0.66946	0.63640
370	0.70711	0.61925	0.64017	0.58577	0.63180	0.59833	0.60251	0.66946	0.64017	0.66527	0.63598
390	0.70293	0.62343	0.63598	0.58159	0.64435	0.58159	0.59833	0.66527	0.64854	0.65690	0.63389
410	0.70711	0.62343	0.64017	0.57741	0.64854	0.57741	0.59414	0.66109	0.65272	0.65690	0.63389
430	0.71548	0.61925	0.63598	0.58159	0.65272	0.58577	0.59414	0.65690	0.65272	0.66109	0.63556
450	0.71548	0.62343	0.63598	0.57322	0.65272	0.59833	0.59414	0.64854	0.64854	0.66527	0.63556
470	0.71548	0.61506	0.62762	0.56485	0.65272	0.60251	0.59414	0.64854	0.64435	0.66527	0.63305
490	0.71548	0.61925	0.62762	0.56904	0.65272	0.60251	0.58996	0.65690	0.64435	0.66946	0.63473
510	0.71130	0.61925	0.63180	0.56904	0.65272	0.60251	0.58996	0.65690	0.64435	0.67782	0.63557
530	0.71130	0.61925	0.63598	0.57741	0.65272	0.60251	0.58996	0.65690	0.64854	0.67782	0.63724
550	0.71548	0.61925	0.64017	0.57322	0.65690	0.61506	0.58159	0.66109	0.64854	0.67782	0.63891
570	0.71548	0.61088	0.63598	0.56904	0.64854	0.61088	0.58159	0.66109	0.65272	0.67782	0.63640
590	0.71548	0.61506	0.63598	0.58159	0.64854	0.61925	0.58159	0.66109	0.65272	0.67782	0.63891
610	0.71967	0.61088	0.63598	0.58159	0.64017	0.60669	0.58577	0.66527	0.65690	0.67364	0.63766
630	0.71548	0.61506	0.62762	0.58577	0.64435	0.61088	0.58996	0.66946	0.65690	0.67782	0.63933
650	0.72803	0.60251	0.62762	0.58996	0.64017	0.61088	0.58996	0.66946	0.65690	0.67782	0.63933
670	0.72803	0.60251	0.62762	0.59414	0.64017	0.61925	0.59414	0.66527	0.66109	0.67364	0.64059
690	0.72803	0.60251	0.62762	0.58577	0.63598	0.62343	0.60251	0.66527	0.66109	0.67364	0.64058
710	0.72803	0.60669	0.64017	0.58996	0.64854	0.62343	0.60251	0.66109	0.65690	0.68201	0.64393
730	0.72385	0.61088	0.63598	0.58996	0.64435	0.62762	0.60251	0.66527	0.65690	0.68201	0.64393
750	0.72385	0.61088	0.63180	0.58577	0.64435	0.62343	0.59833	0.66527	0.65272	0.68619	0.64226
770	0.71967	0.61088	0.63180	0.58996	0.64435	0.61088	0.59414	0.64854	0.65272	0.67782	0.63808
790	0.71548	0.61088	0.63598	0.58996	0.64435	0.61506	0.59414	0.64017	0.65272	0.68201	0.63808
810	0.71130	0.61506	0.63180	0.58577	0.64435	0.61506	0.59414	0.64017	0.64435	0.68619	0.63682
830	0.71548	0.61088	0.63180	0.58577	0.64435	0.62343	0.58159	0.64435	0.64435	0.69038	0.63724
850	0.71967	0.61088	0.63180	0.59833	0.64435	0.62343	0.59414	0.63598	0.63180	0.69038	0.63808
870	0.71548	0.61088	0.63180	0.59414	0.64017	0.62762	0.59414	0.64017	0.64017	0.68201	0.63766
890	0.71548	0.61088	0.63180	0.59833	0.64854	0.62762	0.58996	0.64435	0.65272	0.68619	0.64059
910	0.71967	0.61506	0.63180	0.60251	0.65272	0.62343	0.58996	0.64435	0.65272	0.68201	0.64142
930	0.71130	0.61088	0.63180	0.61506	0.65272	0.63180	0.59414	0.64435	0.65690	0.68201	0.64310
950	0.70711	0.60669	0.63598	0.61925	0.65272	0.62762	0.59414	0.64435	0.65690	0.67782	0.64226
970	0.71130	0.60669	0.63180	0.61506	0.65272	0.63180	0.58996	0.64435	0.66109	0.67364	0.64184
990	0.70293	0.60669	0.63180	0.61506	0.65272	0.63598	0.59833	0.64017	0.66109	0.67782	0.64226

TABLE B.5: Emoticons converted datasets accuracy vs Number of features

No of Features	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Dataset 7	Dataset 8	Dataset 9	Dataset 10	Average Score
10	0.58996	0.53975	0.57741	0.57741	0.45607	0.54812	0.60669	0.57741	0.58577	0.57741	0.56360
30	0.61925	0.55649	0.56067	0.60669	0.51883	0.56067	0.62343	0.58159	0.53138	0.54393	0.57029
50	0.61925	0.58159	0.56904	0.61088	0.59833	0.56904	0.63180	0.59414	0.55649	0.55649	0.58871
70	0.61925	0.58577	0.64435	0.56904	0.58159	0.54812	0.61925	0.62343	0.57322	0.57741	0.59414
90	0.61506	0.57322	0.62762	0.58577	0.58996	0.56067	0.61925	0.65272	0.58577	0.59833	0.60084
110	0.62343	0.58996	0.64435	0.56904	0.56485	0.55230	0.62343	0.67782	0.58996	0.58577	0.60209
130	0.57741	0.58159	0.62762	0.57322	0.56904	0.56067	0.61506	0.67782	0.58577	0.58577	0.59540
150	0.58996	0.58159	0.64017	0.57741	0.55230	0.58159	0.62343	0.69038	0.58159	0.59414	0.60126
170	0.60251	0.58996	0.65272	0.57741	0.56485	0.59833	0.61925	0.67782	0.58577	0.58577	0.60544
190	0.59414	0.58159	0.64017	0.56904	0.57322	0.61925	0.60669	0.66109	0.58577	0.59414	0.60251
210	0.58577	0.58577	0.64017	0.57322	0.58996	0.60669	0.61925	0.64854	0.59414	0.58577	0.60293
230	0.58996	0.58159	0.64854	0.57322	0.59414	0.59833	0.63598	0.64435	0.61088	0.58159	0.60586
250	0.58996	0.56904	0.64435	0.58159	0.61088	0.59414	0.63598	0.65272	0.60251	0.57741	0.60586
270	0.61506	0.57741	0.64854	0.59833	0.60251	0.58577	0.62762	0.64435	0.60669	0.58577	0.60920
290	0.61506	0.58577	0.65272	0.59833	0.61088	0.58996	0.60669	0.64854	0.60669	0.57741	0.60920
310	0.61088	0.58159	0.64435	0.60251	0.62762	0.57322	0.61925	0.64435	0.61088	0.57741	0.60921
330	0.60669	0.58159	0.65272	0.59833	0.62343	0.57741	0.62343	0.65272	0.63180	0.57322	0.61213
350	0.60669	0.57322	0.65690	0.58996	0.62343	0.57741	0.61506	0.66109	0.62343	0.56904	0.60962
370	0.60669	0.57322	0.65272	0.57741	0.62762	0.57741	0.60669	0.66109	0.63598	0.58159	0.61004
390	0.59833	0.56904	0.64435	0.56904	0.62343	0.58159	0.61088	0.66109	0.62343	0.57741	0.60586
410	0.60251	0.58159	0.64854	0.57741	0.62762	0.58577	0.61088	0.66109	0.61925	0.58159	0.60963
430	0.59414	0.57741	0.64854	0.57741	0.63180	0.58996	0.61506	0.65690	0.61925	0.57322	0.60837
450	0.59414	0.57741	0.65690	0.57741	0.63180	0.60251	0.61506	0.66109	0.62343	0.56067	0.61004
470	0.59414	0.57322	0.65690	0.57741	0.62343	0.59414	0.61506	0.64854	0.61925	0.56067	0.60628
490	0.59414	0.57322	0.65272	0.57322	0.62343	0.59833	0.61925	0.65272	0.61506	0.56067	0.60628
510	0.59414	0.56904	0.65690	0.58577	0.64435	0.60251	0.61506	0.65690	0.61925	0.54812	0.60920
530	0.59833	0.56904	0.65690	0.57741	0.64017	0.60251	0.61088	0.64854	0.63180	0.54393	0.60795
550	0.58577	0.57322	0.65690	0.56904	0.63598	0.59833	0.61088	0.65272	0.62343	0.54812	0.60544
570	0.58577	0.57322	0.66109	0.57322	0.64017	0.59833	0.61506	0.64854	0.61925	0.55230	0.60669
590	0.58996	0.56904	0.66109	0.56485	0.64435	0.61088	0.61506	0.64017	0.62343	0.55230	0.60711
610	0.58577	0.57322	0.65690	0.56067	0.64854	0.61925	0.62343	0.63180	0.61925	0.55230	0.60711
630	0.58996	0.56067	0.65690	0.56485	0.64017	0.62762	0.62343	0.64017	0.61925	0.55649	0.60795
650	0.58996	0.56067	0.64435	0.56485	0.64435	0.62343	0.62343	0.64435	0.62343	0.55649	0.60753
670	0.59414	0.56904	0.65272	0.56485	0.64017	0.63180	0.62343	0.63598	0.61925	0.55230	0.60837
690	0.58996	0.57322	0.65690	0.56485	0.64435	0.62343	0.61506	0.64017	0.61088	0.56067	0.60795
710	0.58996	0.57322	0.65690	0.56485	0.64017	0.62762	0.61506	0.63598	0.61088	0.57322	0.60879
730	0.59414	0.56904	0.64854	0.57322	0.63598	0.63598	0.61088	0.63598	0.61506	0.56904	0.60879
750	0.59414	0.56485	0.64435	0.57741	0.63598	0.64017	0.61088	0.63598	0.61506	0.57322	0.60920
770	0.59414	0.56067	0.64435	0.59414	0.63598	0.63598	0.62343	0.63598	0.62762	0.57322	0.61255
790	0.59833	0.56485	0.63598	0.59414	0.63180	0.62762	0.62343	0.63180	0.62762	0.57741	0.61130
810	0.58996	0.57741	0.63598	0.59414	0.63180	0.62343	0.62343	0.63598	0.63180	0.57741	0.61213
830	0.58159	0.57741	0.64017	0.59414	0.62343	0.62343	0.62343	0.63598	0.61925	0.58159	0.61004
850	0.58159	0.58159	0.64435	0.59414	0.62343	0.62343	0.61925	0.63598	0.63598	0.58159	0.61213
870	0.57741	0.58577	0.64435	0.59414	0.62343	0.62343	0.61925	0.63598	0.63598	0.58577	0.61255
890	0.57741	0.58577	0.64435	0.58577	0.62343	0.62762	0.62343	0.63598	0.64017	0.58577	0.61297
910	0.57322	0.58159	0.64854	0.58159	0.62343	0.63180	0.61925	0.63598	0.63598	0.58996	0.61213
930	0.57741	0.58577	0.64854	0.57741	0.62343	0.61925	0.61506	0.63180	0.64435	0.58996	0.61130
950	0.57741	0.58577	0.65690	0.57741	0.61925	0.62343	0.61506	0.63598	0.64435	0.58577	0.61213
970	0.57322	0.58577	0.66109	0.58577	0.61506	0.61925	0.61506	0.62343	0.64017	0.58996	0.61088
990	0.57322	0.58996	0.66109	0.58996	0.61088	0.62343	0.61506	0.61088	0.64435	0.58159	0.61004

TABLE B.6: Lemmatized datasets accuracy vs Number of features

No of Features	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Dataset 7	Dataset 8	Dataset 9	Dataset 10	Average Score
10	0.58159	0.57322	0.57322	0.58577	0.46862	0.56067	0.59833	0.56904	0.57322	0.56067	0.56444
30	0.60251	0.56904	0.56904	0.58159	0.53138	0.56067	0.61506	0.58159	0.52720	0.55649	0.56946
50	0.61088	0.56485	0.56485	0.59833	0.57741	0.55649	0.61925	0.60669	0.54812	0.57741	0.58243
70	0.61506	0.55649	0.60669	0.54812	0.56067	0.53556	0.61506	0.60669	0.57741	0.57741	0.57992
90	0.60251	0.56485	0.61506	0.58159	0.56904	0.51464	0.59833	0.62762	0.59833	0.56485	0.58368
110	0.59833	0.59414	0.61925	0.62343	0.54812	0.55230	0.61506	0.62343	0.57322	0.57322	0.59205
130	0.60669	0.58996	0.61925	0.61506	0.58159	0.54393	0.62343	0.66109	0.57322	0.58996	0.60042
150	0.58577	0.59414	0.64017	0.59833	0.57741	0.54393	0.64435	0.65690	0.58159	0.58577	0.60084
170	0.58577	0.60669	0.65690	0.60251	0.58159	0.55230	0.61506	0.66109	0.56904	0.58577	0.60167
190	0.58996	0.60251	0.66109	0.60251	0.58996	0.55649	0.62343	0.65272	0.58996	0.57322	0.60418
210	0.59414	0.61925	0.64435	0.60251	0.60251	0.54812	0.63180	0.64435	0.62343	0.57741	0.60879
230	0.58159	0.60251	0.64854	0.59833	0.59414	0.55649	0.64017	0.64017	0.61925	0.58996	0.60712
250	0.58159	0.59833	0.64854	0.60251	0.58159	0.56904	0.62343	0.64017	0.61925	0.58159	0.60460
270	0.58159	0.60251	0.63180	0.60251	0.58577	0.55649	0.63598	0.64854	0.61088	0.58159	0.60377
290	0.58577	0.61088	0.63180	0.60669	0.58159	0.56485	0.64017	0.66109	0.61088	0.57322	0.60669
310	0.59833	0.60669	0.63180	0.60669	0.58996	0.56904	0.64435	0.65272	0.61506	0.56067	0.60753
330	0.59833	0.59414	0.63180	0.59833	0.58577	0.57741	0.63180	0.64435	0.60669	0.56904	0.60377
350	0.59833	0.59833	0.63598	0.59414	0.58577	0.57322	0.62762	0.64854	0.59833	0.56485	0.60251
370	0.60251	0.59414	0.63598	0.59833	0.58577	0.58996	0.62762	0.64854	0.60251	0.56067	0.60460
390	0.60669	0.58996	0.64017	0.59414	0.58577	0.60251	0.62343	0.65272	0.60251	0.56485	0.60628
410	0.61088	0.58996	0.64435	0.59414	0.58577	0.58996	0.62343	0.64854	0.61506	0.56067	0.60628
430	0.61088	0.58577	0.64854	0.58996	0.57741	0.58577	0.62343	0.65690	0.61506	0.56067	0.60544
450	0.61506	0.58996	0.64435	0.58577	0.58577	0.59414	0.61925	0.66109	0.61088	0.56067	0.60669
470	0.61506	0.58996	0.66109	0.58577	0.58577	0.59414	0.61088	0.67364	0.60251	0.56067	0.60795
490	0.61506	0.58996	0.64435	0.59414	0.58577	0.59833	0.60669	0.67364	0.60251	0.56904	0.60795
510	0.61506	0.59414	0.64017	0.59414	0.58577	0.59414	0.60669	0.67364	0.60669	0.56904	0.60795
530	0.61925	0.58577	0.64854	0.58577	0.58996	0.59414	0.61088	0.66946	0.61506	0.56904	0.60879
550	0.61506	0.57741	0.63180	0.58577	0.58996	0.60251	0.60669	0.66527	0.62762	0.56485	0.60669
570	0.60669	0.57741	0.62762	0.58577	0.58159	0.60251	0.60669	0.66946	0.62343	0.55649	0.60377
590	0.61088	0.58159	0.63180	0.58996	0.58577	0.60669	0.60669	0.67364	0.62762	0.55649	0.60711
610	0.60669	0.58159	0.62762	0.57741	0.58996	0.61506	0.61506	0.66946	0.62762	0.55649	0.60670
630	0.61088	0.58996	0.63180	0.57322	0.58996	0.61088	0.61925	0.66527	0.61925	0.55649	0.60670
650	0.60251	0.58577	0.63180	0.57322	0.58577	0.61506	0.62762	0.66527	0.61506	0.56067	0.60627
670	0.59833	0.58996	0.61506	0.57322	0.58577	0.62762	0.62762	0.65272	0.61925	0.55230	0.60418
690	0.60251	0.58577	0.61506	0.57741	0.59833	0.62762	0.63180	0.64435	0.61506	0.54812	0.60460
710	0.60251	0.58577	0.61506	0.57322	0.60251	0.62762	0.63180	0.64435	0.62343	0.54812	0.60544
730	0.59833	0.59414	0.61506	0.57322	0.60251	0.62762	0.64017	0.64854	0.61925	0.54812	0.60670
750	0.59414	0.58996	0.61506	0.57741	0.60251	0.63598	0.63598	0.64435	0.61506	0.54393	0.60544
770	0.59833	0.58577	0.61088	0.58159	0.59833	0.62762	0.62762	0.64017	0.61088	0.55649	0.60377
790	0.60251	0.57741	0.60669	0.58159	0.60251	0.63180	0.62762	0.63180	0.62343	0.56485	0.60502
810	0.60251	0.58159	0.61925	0.58159	0.59833	0.62343	0.63598	0.63598	0.62343	0.57322	0.60753
830	0.59414	0.56904	0.61925	0.58577	0.59833	0.61925	0.63598	0.63598	0.62762	0.57322	0.60586
850	0.59414	0.57322	0.61506	0.58577	0.61088	0.61506	0.63598	0.63180	0.63180	0.57741	0.60711
870	0.58996	0.58159	0.62343	0.58159	0.61506	0.61506	0.63598	0.63180	0.63180	0.57322	0.60795
890	0.58996	0.58159	0.63180	0.58159	0.61506	0.61088	0.63598	0.63180	0.63598	0.58159	0.60962
910	0.58159	0.58577	0.63598	0.58159	0.62343	0.61088	0.63598	0.63180	0.63598	0.57741	0.61004
930	0.58159	0.58159	0.63180	0.58577	0.62762	0.63180	0.62762	0.63180	0.63180	0.58577	0.61172
950	0.58159	0.58577	0.62762	0.59414	0.62343	0.62762	0.62762	0.62343	0.64017	0.57741	0.61088
970	0.58159	0.58996	0.62343	0.58996	0.61506	0.63180	0.62762	0.62762	0.64017	0.58159	0.61088
990	0.58159	0.59414	0.62762	0.58996	0.61506	0.62343	0.62762	0.61925	0.64435	0.58159	0.61046

TABLE B.7: Numeric characters removed dataset accuracy vs Number of features

No of Features	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Dataset 7	Dataset 8	Average Score
10	0.5683	0.6048	0.57247	0.5756	0.55579	0.58498	0.58916	0.61314	0.58303
30	0.58394	0.63087	0.59333	0.61418	0.58498	0.60375	0.61418	0.62774	0.60662
50	0.63608	0.64546	0.60584	0.6194	0.62044	0.61522	0.61627	0.64651	0.62565
70	0.63191	0.65693	0.61835	0.64859	0.64964	0.62148	0.62774	0.64338	0.63725
90	0.64129	0.6757	0.63608	0.64964	0.63191	0.63608	0.64129	0.65485	0.64586
110	0.65798	0.6903	0.64859	0.66528	0.65068	0.64442	0.65798	0.64859	0.65798
130	0.65902	0.68196	0.65068	0.66111	0.65902	0.66111	0.65485	0.67466	0.6628
150	0.65068	0.68822	0.65798	0.66423	0.66319	0.65798	0.64755	0.67362	0.66293
170	0.65276	0.683	0.65798	0.6684	0.66528	0.66423	0.65485	0.68196	0.66606
190	0.64964	0.6757	0.66528	0.66736	0.67779	0.66006	0.66528	0.68196	0.66788
210	0.65068	0.68822	0.67362	0.6757	0.68405	0.66111	0.66945	0.67362	0.67206
230	0.65276	0.69239	0.66945	0.66945	0.67987	0.65798	0.6757	0.67883	0.67205
250	0.64651	0.69135	0.6684	0.66736	0.68509	0.65485	0.68196	0.66528	0.6701
270	0.64651	0.69343	0.66945	0.6684	0.68926	0.66736	0.6903	0.6684	0.67414
290	0.64859	0.69239	0.67258	0.67466	0.68613	0.66736	0.68822	0.67987	0.67623
310	0.64546	0.6903	0.67258	0.66632	0.68613	0.67675	0.69656	0.6757	0.67623
330	0.64234	0.69447	0.67258	0.67153	0.6903	0.6757	0.68717	0.66945	0.67544
350	0.64546	0.69864	0.67883	0.67258	0.69447	0.67258	0.69343	0.67466	0.67883
370	0.64964	0.69552	0.6757	0.67362	0.6976	0.67362	0.69447	0.67675	0.67962
390	0.65381	0.6903	0.67675	0.67258	0.69864	0.67675	0.69447	0.67675	0.68001
410	0.65902	0.69135	0.68717	0.6757	0.69656	0.67779	0.69552	0.67362	0.68209
430	0.66111	0.69656	0.6903	0.67466	0.70282	0.67779	0.68717	0.67675	0.6834
450	0.66423	0.69343	0.68926	0.66945	0.69969	0.68092	0.68926	0.67883	0.68313
470	0.66945	0.69552	0.68613	0.67362	0.69656	0.68092	0.6903	0.67258	0.68314
490	0.67362	0.69343	0.69135	0.67258	0.70282	0.68092	0.68717	0.6757	0.6847
510	0.67258	0.69135	0.69343	0.6684	0.70386	0.68092	0.68613	0.6757	0.68405
530	0.66945	0.68926	0.69343	0.67466	0.70177	0.68405	0.68717	0.67675	0.68457
550	0.6684	0.69447	0.6976	0.67258	0.70073	0.68926	0.68509	0.67466	0.68535
570	0.66736	0.69343	0.69135	0.67258	0.70282	0.68926	0.68926	0.67675	0.68535
590	0.67466	0.68822	0.69135	0.6684	0.69447	0.69239	0.69447	0.67779	0.68522
610	0.67258	0.68926	0.69447	0.67153	0.69969	0.69135	0.69447	0.67883	0.68652
630	0.67362	0.68717	0.6903	0.67258	0.69656	0.69343	0.69239	0.68196	0.686
650	0.67153	0.68405	0.69447	0.6684	0.69969	0.69447	0.68822	0.68196	0.68535
670	0.67466	0.68405	0.68926	0.66945	0.70177	0.69239	0.68926	0.67883	0.68496
690	0.6757	0.68405	0.68926	0.66945	0.69656	0.69343	0.68822	0.67779	0.68431
710	0.67883	0.68613	0.69239	0.66736	0.69343	0.6976	0.68613	0.67779	0.68496
730	0.67779	0.6903	0.68926	0.66632	0.6976	0.69552	0.68822	0.67987	0.68561
750	0.67779	0.68822	0.6903	0.66945	0.69864	0.69552	0.68822	0.67883	0.68587
770	0.67362	0.68717	0.69135	0.67258	0.6976	0.69552	0.68822	0.67987	0.68574
790	0.67049	0.68717	0.69135	0.66945	0.69656	0.69447	0.6903	0.67779	0.6847
810	0.67153	0.68717	0.68717	0.67153	0.6976	0.69656	0.68717	0.68092	0.68496
830	0.67362	0.68717	0.68613	0.67049	0.69969	0.69656	0.68509	0.68092	0.68496
850	0.67675	0.68926	0.68926	0.67049	0.69969	0.69656	0.6903	0.68405	0.68705
870	0.67258	0.68926	0.68196	0.66736	0.70177	0.69656	0.68717	0.67987	0.68457
890	0.67362	0.6903	0.68405	0.67049	0.70177	0.69656	0.68717	0.6757	0.68496
910	0.67153	0.68926	0.683	0.66945	0.69969	0.69656	0.6903	0.6757	0.68444
930	0.66945	0.68717	0.68196	0.66945	0.70282	0.69447	0.68717	0.67466	0.68339
950	0.6684	0.68822	0.68405	0.67258	0.70177	0.69447	0.68613	0.67362	0.68366
970	0.66736	0.68405	0.68196	0.67153	0.70386	0.69343	0.68717	0.67675	0.68326
990	0.66736	0.68509	0.683	0.67153	0.70177	0.69656	0.68822	0.67362	0.68339

TABLE B.8: TFIDF vectorizer classification accuracy

No of Features	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Dataset 7	Dataset 8	Average Score
10	0.55892	0.61522	0.56726	0.58707	0.55892	0.57873	0.58394	0.6121	0.58277
30	0.59228	0.61835	0.59124	0.62044	0.5902	0.60688	0.62148	0.63504	0.60949
50	0.62982	0.63608	0.60167	0.63295	0.62148	0.61522	0.62461	0.64546	0.62591
70	0.62669	0.64651	0.6121	0.64442	0.63921	0.62148	0.63504	0.64964	0.63439
90	0.63712	0.6684	0.62982	0.65589	0.63816	0.63191	0.63504	0.65172	0.64351
110	0.65798	0.67675	0.63712	0.66528	0.64234	0.63295	0.64859	0.65485	0.65198
130	0.66423	0.67466	0.64338	0.67258	0.65068	0.64651	0.65172	0.66423	0.6585
150	0.64964	0.67779	0.63712	0.67049	0.66111	0.65485	0.64964	0.6684	0.65863
170	0.65381	0.67362	0.64546	0.67049	0.66319	0.65589	0.65693	0.67466	0.66176
190	0.65276	0.67049	0.65172	0.67049	0.66632	0.65693	0.66736	0.67466	0.66384
210	0.65485	0.6757	0.65798	0.66736	0.67362	0.65589	0.66215	0.6757	0.66541
230	0.65589	0.68092	0.65381	0.66319	0.6684	0.65172	0.67153	0.67362	0.66489
250	0.65276	0.68613	0.66006	0.66215	0.67362	0.65485	0.66736	0.67675	0.66671
270	0.65068	0.68613	0.65798	0.66006	0.6684	0.66111	0.66736	0.67987	0.66645
290	0.65276	0.68613	0.65902	0.66215	0.67258	0.66632	0.66736	0.67987	0.66827
310	0.64964	0.6903	0.65798	0.65902	0.67258	0.66736	0.67258	0.67362	0.66789
330	0.65172	0.69239	0.65798	0.66528	0.6757	0.66736	0.6757	0.67153	0.66971
350	0.65068	0.69552	0.65902	0.66736	0.6757	0.66736	0.67258	0.67362	0.67023
370	0.64964	0.6903	0.65276	0.67049	0.6757	0.66945	0.67258	0.67675	0.66971
390	0.65589	0.68926	0.65276	0.67153	0.6757	0.6684	0.67362	0.67362	0.6701
410	0.65589	0.68926	0.65589	0.6684	0.67362	0.66945	0.67258	0.67883	0.67049
430	0.65381	0.6903	0.65381	0.66945	0.67466	0.67153	0.67258	0.67675	0.67036
450	0.65589	0.68717	0.65589	0.66945	0.67258	0.67362	0.67362	0.68196	0.67127
470	0.65798	0.6903	0.65381	0.66736	0.67153	0.67049	0.67258	0.68196	0.67075
490	0.65798	0.6903	0.65693	0.67049	0.67153	0.67258	0.67466	0.68092	0.67192
510	0.65902	0.68613	0.65798	0.67153	0.67258	0.67153	0.67049	0.67779	0.67088
530	0.66006	0.68509	0.66319	0.6684	0.67675	0.67258	0.6684	0.67675	0.6714
550	0.65902	0.68405	0.66319	0.6684	0.67675	0.67362	0.66945	0.67675	0.6714
570	0.65902	0.68717	0.66319	0.66736	0.67675	0.67362	0.6684	0.6757	0.6714
590	0.66319	0.68613	0.65902	0.67153	0.6757	0.67466	0.66945	0.6757	0.67192
610	0.66319	0.68717	0.65798	0.67049	0.66945	0.67258	0.67362	0.67675	0.6714
630	0.66319	0.68613	0.65693	0.66736	0.67153	0.67466	0.67258	0.67883	0.6714
650	0.66423	0.68822	0.65589	0.66632	0.67049	0.67466	0.67153	0.68092	0.67153
670	0.66423	0.68717	0.65589	0.66632	0.67049	0.67258	0.67153	0.68092	0.67114
690	0.66423	0.68613	0.65902	0.66736	0.67153	0.67466	0.67049	0.67987	0.67166
710	0.66423	0.68613	0.65589	0.6684	0.66945	0.67466	0.66945	0.67883	0.67088
730	0.66319	0.68717	0.65381	0.6684	0.67258	0.67466	0.66736	0.67883	0.67075
750	0.66423	0.68509	0.65276	0.66945	0.66945	0.67466	0.66528	0.67883	0.66997
770	0.66111	0.68613	0.65381	0.66632	0.66945	0.67362	0.66736	0.67987	0.66971
790	0.65902	0.68926	0.65381	0.66632	0.67153	0.67049	0.66319	0.6757	0.66867
810	0.65902	0.68717	0.65068	0.66632	0.67362	0.67258	0.66423	0.6757	0.66867
830	0.66006	0.68717	0.65068	0.66215	0.67362	0.67153	0.66736	0.67466	0.6684
850	0.66111	0.68717	0.65172	0.66632	0.67153	0.6757	0.66319	0.67675	0.66919
870	0.66215	0.68613	0.64964	0.6684	0.67049	0.67362	0.66319	0.6757	0.66867
890	0.66215	0.68613	0.65068	0.6684	0.67258	0.67466	0.66632	0.6757	0.66958
910	0.66215	0.68717	0.65068	0.66736	0.67362	0.67466	0.6684	0.6757	0.66997
930	0.66111	0.68613	0.65276	0.66423	0.67049	0.6757	0.66528	0.6757	0.66893
950	0.66006	0.683	0.64964	0.66423	0.67362	0.67675	0.66528	0.6757	0.66854
970	0.66006	0.683	0.64859	0.66528	0.67258	0.67675	0.66528	0.67466	0.66828
990	0.65902	0.68405	0.64859	0.66528	0.67362	0.67675	0.66528	0.67258	0.66815

TABLE B.9: Count vectorizer classification accuracy

No of features	Score 1	Score 2	Score 3	Score 4	Average Score
10	0.57351	0.54327	0.59958	0.57456	0.57273
30	0.60271	0.56934	0.59750	0.59333	0.59072
50	0.61522	0.58290	0.63816	0.61314	0.61236
70	0.62252	0.58916	0.64546	0.61731	0.61861
90	0.62461	0.61835	0.64234	0.62148	0.62669
110	0.62878	0.63295	0.64234	0.63295	0.63425
130	0.63504	0.64129	0.64129	0.62878	0.63660
150	0.64442	0.63608	0.64338	0.63295	0.63921
170	0.65693	0.63816	0.63816	0.64651	0.64494
190	0.65172	0.64129	0.64964	0.64546	0.64703
210	0.65693	0.63921	0.65589	0.65589	0.65198
230	0.65693	0.64755	0.65798	0.66006	0.65563
250	0.65798	0.64859	0.66736	0.66736	0.66032
270	0.66632	0.65589	0.66945	0.67258	0.66606
290	0.67362	0.65485	0.66840	0.67570	0.66814
310	0.68092	0.64546	0.67049	0.67049	0.66684
330	0.68509	0.64755	0.66632	0.67049	0.66736
350	0.68509	0.65276	0.66840	0.66736	0.66840
370	0.68405	0.64859	0.66215	0.66736	0.66554
390	0.68509	0.64859	0.66632	0.66736	0.66684
410	0.68509	0.64859	0.66736	0.67049	0.66788
430	0.68300	0.65068	0.67362	0.66945	0.66919
450	0.68405	0.64964	0.67883	0.67258	0.67127
470	0.69239	0.65589	0.67987	0.66736	0.67388
490	0.68822	0.64755	0.68092	0.66528	0.67049
510	0.68613	0.65485	0.68300	0.67049	0.67362
530	0.68822	0.66423	0.68509	0.66840	0.67649
550	0.68822	0.65172	0.68300	0.67049	0.67336
570	0.68822	0.65589	0.67883	0.66945	0.67310
590	0.69030	0.65798	0.68509	0.66736	0.67518
610	0.68926	0.65902	0.67987	0.66945	0.67440
630	0.69239	0.65693	0.67883	0.66632	0.67362
650	0.69030	0.65798	0.68092	0.66319	0.67310
670	0.68822	0.66006	0.67675	0.67049	0.67388
690	0.68300	0.66006	0.67779	0.66736	0.67205
710	0.68509	0.65798	0.67987	0.66736	0.67258
730	0.68405	0.65902	0.68092	0.66840	0.67310
750	0.68092	0.66528	0.67675	0.67570	0.67466
770	0.67987	0.66319	0.68092	0.67153	0.67388
790	0.68092	0.66006	0.68196	0.67779	0.67518
810	0.68092	0.66215	0.68509	0.67883	0.67675
830	0.68509	0.66736	0.68092	0.67675	0.67753
850	0.68509	0.66945	0.67987	0.67466	0.67727
870	0.68196	0.66632	0.68092	0.67883	0.67701
890	0.68300	0.66945	0.67987	0.67466	0.67675
910	0.68300	0.66945	0.67779	0.67779	0.67701
930	0.68405	0.66528	0.67779	0.67883	0.67649
950	0.68196	0.66528	0.67466	0.67466	0.67414
970	0.68300	0.66632	0.67675	0.67466	0.67518
990	0.68196	0.66945	0.67675	0.67675	0.67623

TABLE B.10: n-gram (1,1)

No of features	Score 1	Score 2	Score 3	Score 4	Average Score
10	0.57351	0.54327	0.59958	0.57456	0.57273
30	0.58603	0.55787	0.60063	0.58081	0.58133
50	0.58498	0.57456	0.62044	0.61210	0.59802
70	0.61001	0.58603	0.63504	0.62357	0.61366
90	0.61940	0.60584	0.64129	0.62252	0.62226
110	0.61314	0.62982	0.64338	0.62044	0.62669
130	0.62565	0.63087	0.64651	0.61522	0.62956
150	0.63921	0.63608	0.63608	0.63087	0.63556
170	0.63921	0.64129	0.64025	0.62774	0.63712
190	0.64651	0.64546	0.64651	0.63921	0.64442
210	0.64859	0.65172	0.65693	0.63712	0.64859
230	0.64755	0.63921	0.65381	0.64025	0.64520
250	0.64546	0.64755	0.65485	0.63816	0.64651
270	0.64546	0.64755	0.65485	0.64338	0.64781
290	0.64859	0.64859	0.65485	0.65068	0.65068
310	0.65068	0.64338	0.66319	0.66111	0.65459
330	0.65276	0.64546	0.66632	0.66632	0.65772
350	0.65902	0.64964	0.66111	0.66528	0.65876
370	0.65902	0.64546	0.66632	0.67049	0.66032
390	0.66528	0.65276	0.66423	0.66423	0.66163
410	0.65693	0.65693	0.66736	0.66006	0.66032
430	0.65902	0.65485	0.67153	0.65485	0.66006
450	0.66632	0.65381	0.66840	0.66319	0.66293
470	0.66840	0.64964	0.66736	0.66111	0.66163
490	0.67466	0.65172	0.67258	0.65693	0.66397
510	0.67362	0.64755	0.67153	0.65902	0.66293
530	0.67570	0.65381	0.66632	0.65589	0.66293
550	0.66945	0.64964	0.66528	0.66215	0.66163
570	0.67466	0.65172	0.65589	0.65381	0.65902
590	0.67675	0.64755	0.66111	0.65172	0.65928
610	0.67570	0.64859	0.66006	0.65589	0.66006
630	0.67883	0.65485	0.66111	0.66319	0.66449
650	0.67675	0.65798	0.66528	0.66319	0.66580
670	0.68092	0.65589	0.66006	0.66319	0.66502
690	0.67883	0.65798	0.66423	0.66319	0.66606
710	0.67466	0.65693	0.66215	0.66319	0.66423
730	0.67258	0.65693	0.66945	0.66423	0.66580
750	0.67153	0.65693	0.66632	0.66840	0.66580
770	0.67675	0.65798	0.66423	0.67049	0.66736
790	0.67987	0.65693	0.66319	0.66736	0.66684
810	0.67779	0.65485	0.66423	0.66736	0.66606
830	0.68092	0.65485	0.66632	0.67049	0.66814
850	0.68092	0.65693	0.66111	0.66840	0.66684
870	0.67570	0.65693	0.66215	0.67153	0.66658
890	0.68509	0.66006	0.66736	0.67362	0.67153
910	0.68405	0.66215	0.66840	0.67570	0.67258
930	0.67987	0.66423	0.66840	0.67466	0.67179
950	0.68092	0.66423	0.67049	0.67675	0.67310
970	0.68717	0.66319	0.66632	0.68092	0.67440
990	0.68196	0.66423	0.66423	0.67883	0.67231

TABLE B.11: n-gram (1,2)

No of features	Score 1	Score 2	Score 3	Score 4	Average Score
10	0.57351	0.54327	0.59958	0.57456	0.57273
30	0.58603	0.55683	0.60167	0.58186	0.58160
50	0.58916	0.57247	0.62044	0.61105	0.59828
70	0.61418	0.58498	0.63504	0.62044	0.61366
90	0.62148	0.60271	0.64234	0.62148	0.62200
110	0.61835	0.63087	0.64234	0.61627	0.62696
130	0.62774	0.63399	0.64755	0.61940	0.63217
150	0.63608	0.63399	0.63921	0.63087	0.63504
170	0.64234	0.64546	0.64129	0.63087	0.63999
190	0.64651	0.64234	0.64129	0.63921	0.64234
210	0.64859	0.64025	0.65381	0.63608	0.64468
230	0.64338	0.64129	0.64755	0.64025	0.64312
250	0.64234	0.64964	0.65172	0.64025	0.64599
270	0.63921	0.64859	0.65798	0.64442	0.64755
290	0.64442	0.64651	0.65485	0.64859	0.64859
310	0.65381	0.64546	0.65902	0.65276	0.65276
330	0.65485	0.64546	0.66006	0.66319	0.65589
350	0.65798	0.64964	0.66215	0.66215	0.65798
370	0.65798	0.65381	0.66423	0.66945	0.66137
390	0.65798	0.65381	0.66736	0.66319	0.66058
410	0.66632	0.64859	0.66632	0.66006	0.66032
430	0.66632	0.65068	0.67153	0.65589	0.66111
450	0.67362	0.65068	0.66840	0.65276	0.66137
470	0.67258	0.64859	0.66840	0.65693	0.66163
490	0.67049	0.64964	0.66945	0.65798	0.66189
510	0.67258	0.64234	0.66945	0.65485	0.65980
530	0.67049	0.64755	0.66632	0.65381	0.65954
550	0.67466	0.64859	0.66423	0.65068	0.65954
570	0.67987	0.64964	0.66840	0.65902	0.66423
590	0.68092	0.65172	0.66319	0.66215	0.66449
610	0.68196	0.65693	0.66423	0.66319	0.66658
630	0.68092	0.65798	0.66319	0.66736	0.66736
650	0.67675	0.65589	0.66111	0.66840	0.66554
670	0.67258	0.65798	0.65902	0.66632	0.66397
690	0.67258	0.65902	0.66423	0.66632	0.66554
710	0.67570	0.65798	0.66528	0.66632	0.66632
730	0.67258	0.65589	0.66840	0.67049	0.66684
750	0.67362	0.65693	0.66840	0.67049	0.66736
770	0.67570	0.65589	0.66840	0.67049	0.66762
790	0.67258	0.65485	0.66632	0.67466	0.66710
810	0.67049	0.65172	0.67362	0.67362	0.66736
830	0.66840	0.65798	0.67258	0.67362	0.66814
850	0.66736	0.65276	0.67466	0.67570	0.66762
870	0.67258	0.65798	0.67049	0.67675	0.66945
890	0.67258	0.65693	0.66945	0.67675	0.66893
910	0.67675	0.65589	0.66840	0.67570	0.66919
930	0.67987	0.65485	0.66423	0.67466	0.66840
950	0.68613	0.65276	0.66840	0.67466	0.67049
970	0.68613	0.65276	0.66945	0.67362	0.67049
990	0.68092	0.65798	0.66840	0.67049	0.66945

TABLE B.12: n-gram (1,3)

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	2	3	4	5	6	7	8	9	10
"n-gram (11)"	0.67361835	0.68717414	0.68821689	0.68925965	0.69134515	0.69343066	0.69447341	0.69447341	0.69655892	0.69447341	0.68821689	0.67883212	0.67987487	0.67674661	0.67466111	0.6725756	0.67466111	0.67153285	0.67153285
"n-gram (12)"	0.67049009	0.68821689	0.69343066	0.69551616	0.69447341	0.69760167	0.69760167	0.70072993	0.70281543	0.70177268	0.69134515	0.6923879	0.6903024	0.68821689	0.68613139	0.68300313	0.68196038	0.68300313	0.68300313
"n-gram (13)"	0.66944734	0.68821689	0.69134515	0.68925965	0.68925965	0.6923879	0.6923879	0.69343066	0.69134515	0.69134515	0.68821689	0.68821689	0.6903024	0.6923879	0.6903024	0.6923879	0.6903024	0.68821689	0.68821689

TABLE B.13: c value vs n-gram