Machine Learning Approach to Opinion Mining

By

WONG KOH SING

A project to be submitted to the Department of Internet Engineering and Computer Science,

Faculty of Engineering and Science,

UniversitiTunku Abdul Rahman,

In partial fulfillment of the requirements for the degree of

Master of Information Systems

April 2015

Table of Contents

ABSTRACT v
ACKNOWLEDGEMENT viii
APPROVAL SHEET ix
SUBMISSION OF PROJECT x
DECLARATION xi
Chapter 1: Introduction 1
1.0 Background of study 2
1.1 Problem statement 3
1.2 Objectives 5
1.3 Research Motivation 7
1.4 Project Scope 7
1.5 Research outline 9
Chapter 2: Literature Review 10
2.0 Opinion Mining 11
2.1 Application 12
2.2 Emotion mining 14
2.3 Classification techniques 15
2.3.1 Feature-based opining mining 15
2.3.2 Sentimental classification 16
2.3.3 comparative sentence and relation extraction 17
2.3.4 Other mining task approaches 18
2.4 Opinion mining process 19
2.4.1 Pre-Processing 21
2.4.2 Part-Of-Speech tagging 22
2.4.3 Bookstrapping 24
2.5 Machine learning 26
2.5.1 Naïve Bayes 26
2.5.2 Support vector machine 27
2.6 Challenge 28

2.6.1 Noise text problem 28
2.6.2 Spam opinion problem 29
2.6.3 Idiom problem 30
2.6.4 Document level problem 30
2.6.5 Sentiment classifying problem 31
2.6.6 Domain specific problem 32
2.7 Feature selection 33
2.7.1 Unigram 33
2.7.2 Bigram 34
2.7.3 Trigram 34
2.7.4 N-gram 35
2.7.5 Lemmas/stems 35
2.7.6 Negation 36
2.8 Standard evaluation measures 37
2.9 Summary chapter 38
Chapter 3: Research Methodology 39
3.0 Introduction 40
3.1 Datasets 40
3.1.2 Mejaj dataset HS+ 40
3.2 Sentiment sentences labeling 41
3.3 Baseline tools 42
3.3.1 Python 43
3.3.1.1 Python for tokenization process 44
3.3.1.2 Python for stopwords process 45
3.3.1.3 Python for feature extraction 46
3.3.1.4 Python with bigram collocation 47
3.3.2 R project for statistical computing 47
3.4 Supervised learning 48
3.5 Unit Testing 50
3.6 Summary Chapter 52
Chapter 4: Experiment and Result 53

4.0 Introduction 54
4.1 Experimental setup 54
4.1.1 Experimental data 54
4.1.2 Experiment process steps 55
4.1.3 Pre-processing step 56
4.1.3.1 Tokenization 56
4.1.3.2 Stopwords 57
4.1.4 Classification tool –Naïve bayes classification 58
4.2 Experiment result 58
4.2.1 Performance analysis on 1000 samples with naïve bayes classifier 59
4.2.2 Performance analysis on 2000 samples with naïve bayes classifier 60
4.2.3 Performance analysis on 6000 samples with naïve bayes classifier 61
4.2.4 Performance analysis on 1000 samples with support vector linear classifier 62
4.2.5 Performance analysis on 2000 samples with support vector linear classifier 63
4.2.6 Performance analysis on 6000 samples with support vector linear classifier 64
4.3 Result discussion 65
4.4 Comparison with sentiment analysis with AFINN wordlist 67
4.4.1 Performance analysis by using AFINN wordlist system 68
4.4.2 Performance analysis by using both naïve bayes and support vector linear 69
4.5 Summary Chapter 70
Chapter 5: Conclusion & future work 71
5.0 Introduction 72
5.1 Research Summary 72
5.2 Main Contributions 73
5.3 Limitation of research 74
5.4 Future works 74
References 76
Appendix A 79

Test with naïve bayes 79
Test with support vector linear 81

ABSTRACT

The incessant growth of the use of social media has increased the need for powerful and robust machinery in analysingimmense amount of data from different realms of industries. We are always eager to know abouthow do other people think and how do they perceive things. Thus, opinion mining is important in learning human behaviours and different personality traits by extracting information from all possible instances such as hidden emotions, unidentified representative, big texts, and eventhe use of urban language. The use of it is not limited to detecting meaningful terms and particular technical word used in different aspects of fields such as politics, economics, and sociology. The fastest growing social media forum such as the Facebook allows all its users tocontribute information without borders; sharing opinions about current world issues, and even their attitude and views towards life. Thus, opinion mining is one of most well-known and important fields of study nowadays.

This dissertation presents sentimental analysis of users' opinions. Explaining the work of machine used to classify and identify important words in a sentence that are used to gain better understanding andto have greater sentimental impact onother users. For example, any "netizens" like to use Facebook as a medium to express their thoughts on current political topics such as the upcoming presidential election in the United States.

This study uses a structural modelling and Bayesian inference machine to identify opinions polarity and subsequently classifies positive and negative oriented opinions

Keywords: Opinion mining, sentiment analysis, text mining, sentiment mining, text analysis

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my project supervisor, Dr.Tay Yong Haur for his continuous guidance and counselling in my preparation of this MIS project, much thanks is needed to be expressed for his constant support and encouragement until the completion of this project. I would also like to convey my heartfelt appreciation to Mr. Tan for contributing his useful ideas and knowledge in this field. I have learned so much from Mr Tan on his great passion, wide scope of knowledge, and accute insight to the research. I am grateful to him not only for his leadership, but also for his continuous support and encouragement and his willingness to spend his precious time on guiding my research.

Another warmest thank is extended to Dr.Tay for sharing his resources, opinion, knowledge, good experience in programming skill and development. Last but not least, I would also like to personally thank my family, friends and my fellow course-mates with no exception. I offer my sincere thanks for giving clear guidelines and material-wise support given by Mr. Teck. My project is made better because of your assistance. Once again, to those who sharedyour valuable advice and helped me to improve the quality of my work, I offer my sincere gratitude.

APPROVAL SHEET

This project is entitled "**Machine Learning Approach to Opinion Mining**", prepared by Wong Koh Sing and submitted as partial fulfillment of the requirements for the degree of Master of Information Systems at UniversitiTunku Abdul Rahman.

Approved by:

_____

(Dr. Tay Yong Haur) Date:……..

Supervisor

Department of Internet Engineering and Computer Science

Faculty of Engineering and Science

Head of Programme (Master of Information Systems)

Chairperson, Centre for Computing and Intelligent Systems (CCIS)

Universiti Tunku Abdul Rahman

FACULTY OF ENGINEERING AND SCIENCE

UNIVERSITI TUNKU ABDUL RAHMAN

Date: _____

SUBMISSION OF PROJECT

It is hereby certified that WONG KOH SING (ID No: **11UEM06213**) has completed this project entitled "Machine Learning Approach to Opinion Machine" under the supervision of DrTay Yong Haur (Supervisor) from the Department of Internet Engineering and Computer Science, Faculty of Engineering and Science.

I hereby give permission to the University to upload softcopy of my project in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,

(WONG KOH SING)

DECLARATION

I hereby declare that thisproject is my original work except for quotations and citations which have been duly acknowledged and credited. I also declare that it has not been previously or concurrently submitted for any other degree programme at UTAR or other institutions.

_____

(WONG KOH SING)

Date _____

Chapter 1: Introduction

- Background of study
  - ✓ Problem statement
  - ✓ Objective
- Research motivation
- Project scope
- Result outline

1.0 Background of study

Social networking such as Facebook, Twitter, Instagram, Google +, LinkedIn and Blogs are directing their businesses with the same vision that tries to attract more users to develop social relations and virtual environment human interaction in societies around the world. They conceived the trait that human beings have the tendency to share information regarding their opinion, state of mind, lifestyle, emotions, and sentimental feelings via different avenues. Hence, social media provides a platform for the sharing and witnessing of interactive information. They come with attractive features that allow users to manage their own posts and provide the convenience to save link for sharing or to read later. With these facilities, the shared information is at the reach of fingertips of every one who has access to that particular social media, at any time, and in any place.

Many scholars apply machine learning approach and implemented text mining tool in social networking websites. Opinion mining/sentiment analysis has been an enduring and fascinating idea. Many researches have beendone in the past few years, such as natural language processing (NLP), web mining, data mining, text mining, and so on. All were done with the same purpose - to gather and analyze disparate words of thoughts and opinions about different topics.

By applying sentiment analysis, information can be accurately classified and human thoughts can be analyzed as well. This is significantly important to refine a NLP tool in order to perform mining task efficiently. There were

numerous attempts aimed to enhance the work of NLP, yet it is still confined by stark challenges.

1.1 Problem statement

Comprehensive analysis and in-depth conclusion can only be reached at when the existing challenges identified and solved.

First of all, we should ask a few simple questions: What is an opinion mining? What is the definition of opinion mining? First part talks about what counts as an “opinion”; second part explains what does opinion-mining do when it deals with opinions. Thereafter follows theconclusion onthe importance of opinion summarization. These are to be done by referring to comments, blogs, and forums, and hence, review statements can thus be justified as positive or as negative.

A studyhas shown that "an opinion is a quintuple form. It consists of 5 components in *ej*, *ajk*, *soijkl*, *hi*, and *tl*. Where *ej* is an opinion target entity, *ajk*is a feature of the entity,*soijkl* is the sentiment value of opinion from the opinion holder on feature of entity at time(Liu, 2010). *hi*, is an opinion holder who holds the opinion, and *tl*is the time when the opinion is expressed. This 5 components are essential elements and they must correspond to each other in the particular set of determined number of entities on an opinion. For example, “regular opinion normally is expressed only one entity at a time whereas comparative opinion can express more than one entity at a time”(Liu, 2010). Other than that, *Lm* is a

location where the opinion is expressed. Location is one of the significant components to indicate where the actual place of opinion holder expresses the opinion. It is very useful for detecting the location of accident, inspection, bogus information, atmosphere, daily activities, and so on. But somehow it might beauncertain tool to evaluate author's feeling because different conditions could possible influence a person's view, opinion, perspective, emotion, and attitude effectively.

Furthermore, we cannot easily define a sentiment because it is subject to how different readers may feel on an opinion. For example, I am happy that SPM result is released today and I get 10As. This sentence is thoroughly subjective to person'sperception of reality. For others, they might not be happy the results they have gotten that day. In fact, sentiment classification is hard and there is no precise answer that can be given. But with the advancement of technology, the precision of classification is improving and becoming more accurate.

There is a stark increment in opinions given online, hence summarization task has to be done. In this case, feature-based summary is the most suitable approach to summarize all opinions according to the frequency of feature appearance result.

1.2 Objectives

The need of value information for decision making process causes most of the user to seek for reliable opinions and experiences from those in large volume of adept consumers and professor users. Therefore, it is very important to point out the pros and consfrom individual perspectives. The expression of positive or negative opinions is a primitive issue to be determined beforegiving a conclusion on whether the user's comments aredesirable or undesirable opinions.

The objectives of this research project are to classify words, phrases, or sentences into positive and negative orientation. This has been splitted into following 4 phases:

1. Information retrieval –to follow the area of study concerned with searching information by using search engine to send a series of queries to trigger procedures via internet, which contain thousands of comments, blogs, and forums. The tools are commonly used to crawl through websites are those web crawler, web spider, and etc. However, Self-feeding data is chosen to operate in this research purpose instead of extract data from the web. A set of trained data will be used in the system, which analyzes the input sentence from end users.

2. Mining out the object features from the reviewer - The aim of extraction process is to automatically extract multiple documents or various data containing positive or negative opinions. The objective of using feature

extraction is to extract adjectives from opinion reviews, if there is no adjective located, adverbs will be chosen as an alternative. It is very useful as pre-processing step of opinion mining method in extracting object and subject features.

3. Identifying opinion sentences whether it is positive or negative - the aim of classification is to classify the sets of adjectives obtained from the previous phase. It can be either used in supervised learning method or unsupervised method to classify or discriminate sources of positive opinions and negative opinions. In this case, supervised learning method has been selected as the preferred method in classifying any specific sentences, phrases, or words.

4. Summarizing the result –to summarize the whole polarity of positive or negative opinions into 2 categories according to the opinion sentence orientation. The measurement of frequency according to the feature appearance in the reviews.

1.3 Research Motivation

What other people think and how do people perceive about on various topics is very valuable information and knowledge on social media. For example, Facebook contains lots of sentiment semantic that motivates scholars to mine out valuable information as much as possible and study the knowledge-based sentiment analysis. The objective of opinion mining is to determinewhat is the opinion polarity of people about specific topics such as political issue at any country.

1.4 Project Scope

The incessant growth of the use of social media has increased the need for powerful and robust machinery in analysingimmense amount of data from different realms of industries. We are always eager to know abouthow do other people think and how do they perceive things. Thus, opinion mining is important in learning human behaviours and different personality traits by extracting information from all possible instances such as hidden emotions, unidentified representative, big texts, and eventhe use of urban language. The use of it is not limited to detecting meaningful terms and particular technical word used in different aspects of fields such as politics, economics, and sociology. The fastest growing social media forum such as the Facebook allows all its users tocontribute information without borders; sharing opinions about current world issues, and even

their attitude and views towards life. Thus, opinion mining is one of most well-known and important fields of study nowadays.

Sentiment analysis uses machine learning approach for classifying sentiments based on information gatheredform survey method, conferences, discourses, and opinion polls. Therefore, sentiment classification plays a role in identifying positive and negative sentiment polarity of user expression toward their discussion topics.

At the end of project, the summarization of the result is generated by supervised learning method. It is a set of boundaries in the project scope that unable to be explained in detail. I will onlyprovide my personal insight on explanation about how actually the machine learning work and specify explicitly the method that I used on a later stage. In order to provide better understanding, detailed discussion is available for further areas of different skills and techniques that are used to extract and classify the overall polarity of positive and negative opinions. Discussion about the calculation for standard evaluation measures of precision, recall, and score, is also included.

1.5 Research outline

This dissertation consists of 5 chapters. The first chapter is the introduction that briefly discusses the background of study followed by problem statement, objective, motivation, and scope of project. Chapter 2 is literature review, which consists of literature review on opinion mining, social networking, text mining, sentiment mining, text analysis, sentiment classification, and sentiment analysis of movie reviews. Chapter 3 is on research methodology including dataset, sentiment sentence labeling,baseline tools, and supervised learning. Chapter 4 is about experimental setup, experiment result, performance analysis, and comparative discussion. Chapter 5, the final chapter, contains research summary, main contributions, limitation of research, and possibilities offuture works.

Chapter 2: Literature Review

Opinion Mining

Application

Emotion mining

Classification techniques

Opinion mining process

Bookstrapping

Machine learning

Challenge

Feature selection

Standard evaluation measures

2.0 Opinion Mining

Opinion mining/sentiment analysis is a computational study of tracking sentiment and opinions from public about political issues or to a particular product. It generally utilized machine learning technique and text mining approach to make computer understanding the sentimental expression. "Opinion mining can be also defined as sub-discipline of text classification which is concerned not with the topic a document about, but with the opinion it expresses" (Juling Ding, 2009). Sentiment can be defined as attitude, emotion, view, appraisal and discrimination subtly by feeling part of specific notion. In common senses, opinion mining may also refers to sentiment analysis, text analysis, sentiment mining, subjectivity/objectivity analysis, and etc. They are doing the same task to collect and mine out the opinions from society.

It cannot be denied that everyone is willing to know "what other people think?" because of humanity curiosity then cause people try to mine out opinions and sentiments as many as possible. Opinions are key factor of our behaviors. It is consisted of believes and perceptions of feelingson how a person see the world. Under these circumstances, where can we scout for opinions? How to obtain information through society? There are two ways in getting it done. One is through acquired opinions by individually from family and friends. Another is to collect opinions through organizations by using various survey methods, conferences, discourses, and opinion polls.

2.1 Application

A study stated in October 2007 found that more than three-quarters of internet users are significantly influenced by online consumer reviews when making a purchase.The study also revealed that consumers were willing to pay at least 20% more for services receiving an"excellent" (5-star) rating rather than for the same service receiving only a"good" (4-star) rating" (Levene, 2010). Hence, marketing analysis makes use of customer feedbacks or would invite them to share their opinions about products that they have purchased. Unfortunately, reading through all customer reviews is a difficult and cumbersome task because there can have thousands of reviews posted by customersevery day. This will produce a significant problem to a new customer who would not possibly be able to read everything before making a good decision. Therefore, opinion mining is used to classify and summarize reviews from opinions on the products at e-commerce websites such as Ebay and Amazon.com for comparison purpose, suggestion and be an efficient advisor for rational purchases.

Moreover, it is suitable for those travelers who travel abroad in selecting which airline is with value added services. It can also be used to look for preferable accommodationand to look up for famous restaurants that serve delicious local or traditional food. For banking purposes, everyone may want to know which bank offers the best transaction processing and provides the best long-term saving plan.

On the other hand, for monitoring social phenomenon purposes, companies and large firms can be informed on the impact of people on their

contribution of huge of volume opinions of which the expression on them offers a direct, unbiased, and global feedback. Especially to those uncontrolled expression of opinion which are anti-social, instigative and those contain negative message could possibly get spread around the world. Opinion mining helps to prevent potentially dangerous situations such as terrorism, watch against symptoms of rebellion simply to determine the general sentiments of blog users.

The potential use of sentiment analysis or opinion mining is to track and gather political views. “Some work has focused on understanding what voters are, whereas other projects have as a long term goal the clarification of politician positions, such as what public figures support or oppose, to enhance the quality of information that voters have access to” (Lee, 2008). It was recently discovered that political views of voters could be influenced or predicted by discussion thread. It can also be used to detect consistency and inconsistency between substantial information and action from voters. President make use this distributed channel to gather useful information and up-to-date political views and understanding of what are the elector trends.

2.2 Emotion mining

Emotion is a strong feeling deriving from frame of mind, mood, or relationship with others. Ekman has categorized emotions in "6 elements: happiness, sadness, anger, fear, disgust and surprise" (Ekman, 1992). However, other researchers have categorized emotions into 2 metrics: positive or negative emotions, and the power of excitement. Mike Thelwall believes that "women are more likely to express their feelings than men in public social website" (Mike Thelwall, 2010). A study has shown that "emotion mining can be divided into 3dimensions. First dimension is identifies positive or negative text. Second dimension evaluates the emotions is rich or not. Third dimension indicates the power of excitement" (Kim, 2011).

Emotions classified from text can be done by classification. It typically refers to the process of dividing set of group data into 2 parts which could be ordered item or unordered item. Also, "classification is very important to formulate the existing problems by applying classification, regression, and ranking to a particular text or phrase given" (Lee, 2008). However, there is a lot of classification techniques that can be applied in opinion mining approach. In the study of markLeveneit was stated that there are 3 essential important mining tasks on texts that express opinions (Levene, 2010).

2.3 Classification techniques

2.3.1 Feature-based opining mining

Feature-based opinion mining evaluate sentences-level to reveal various aspects or features of an object. In general, “it clearly tells about what opinions author expressed is like or didn’t like” (liu, 2006). A positive oriented sentence on an object does not mean that the writer has positive opinions to like everything on the object. However, a negative oriented sentence does not mean that the writer dislike everything one the object. There are divided into 3 phases. First of all, mining machine mine out the specific object features in the customer reviews using natural language processing like part-of-speech-tagging to perform identifying of noun, noun phrase, adjective, and verb. For instance, “That sports car looks awesome”. The object feature is “awesome” and the object is “sports car”. Secondly, the aim of identified object in the reviews is to determine whether opinion on the object feature is positive, negative, or neutral. Thus, the adjectives will be extracted from the sentences or phrases as it effectively expressed the opinions ofits positive or negative orientation. For example, “that sports car looks awesome” and the word “awesome” is positive.

2.3.2 Sentimental classification

Sentimental classification is usually determined by classifying document-level sentenceinto 3 classes: positive, negative, and neutral. “But no details is discovered what people like and didn’t like.” (liu, 2006) For example, in a book and movie review, the system classifies whether it is positive oriented review or negative oriented review. Likewise, sentiment classification also mines opinion words and phrases from the review, at least two words in each sentence will be extracted as the mutual information. The manner of sentiment classification is totally different from topic-based classification. In topic-based classification, topic related words are important whereas in sentiment classification, topic related words are unimportant. The semantic orientation of opinion word that indicate positive or negative opinions are important. For example, positive reference words are “great”, “amazing”, and “perfect” and negative reference words are “bad”, “worst”, and “weak”. Sentiment classification has obviously used 3 supervised learning methods to classify the reviews such as naïve bayes, support vector machine, and maximum entropy. By using naïve bayes method, it would use the positive and negative opinion reviews as training data for selecting the highest probabilities in which the sentence is positive or negative oriented. But they need a lot of training data for the classification to achieve the final result. The document-level sentence can contain one or more opinion reviews, if the opinion review feature is more than positive level, then the review is considered as

positive oriented. However, if the opinion review feature is neutral, then the previous sentence can be taken as a measure to positive, negative or neutral.

2.3.3 comparative sentence and relation extraction

Comparative sentence and Relation Extraction allows one or more objects to be compared with each other. It compares the object with some other similar or different object characteristic. The comparison of object is usually done by using comparative form such as adjective and adverb as a comparative norm. Typically, opinion sentence can be such as “Car A is a great car.” However comparative expressed sentence is expressed in this way “car A is faster than car B whereas car A is slower than car C.” A comparative sentence is indicated as a relationship between two objects and to be extracted from the particular sentences. Generally, one object must be identified and classified into different types or classes, and extract another object within a sentence use to compare with the first object. The example of detection words are “faster”, “slower”, “taller”, and “shorter.”

2.3.4 Other mining task approaches

- The lexicon-based approach/Dictionary-based approach is proposed to "use opinion bearing words or simply opinion words to determine the orientation of an opinion feature" (Xiaowen Ding, 2008). WordNet (http://wordnet.princeton.edu) is used to distinguish synonyms and antonyms of each adjective and group it into frequency sets of synonyms or antonyms to be orientation of the opinion on a feature. Therefore, 2 sets of positive or negative words will be well prepared by iterative searching for their synonyms and antonyms in Wordnet. However, "the shortcoming of dictionary-based approach is unable to find opinion words with domain specific orientation" (Liu, 2010).
- Corpus-based approach rely on "syntactic or co-occurrence patterns and also a seed list of opinion words to find other opinion words in a large corpus" (Liu, 2010). Conjunction (AND) is conjoined adjectives usually have the same orientation. For example, in the sentence, "*This car is beautiful and spacious*," if "beautiful" is known to be positive, it can be inferred that "spacious" is also positive (Liu, 2010). Others have similar constraints such as OR, BUT, EITHER-OR, and NEITHER-NOR.
- There is a subjective classification expressing personal's felling, views, emotions, and behaviors, it is only classifying objective or

subjective sentence thereafter producing positive or negative opinions. “Classifying a sentence or a clause of the sentence as subjective or objective, and for a subjective sentence or clause classifying it as expressing a positive, negative or neutral opinion” (Liu, 2010).

2.4 Opinion mining process

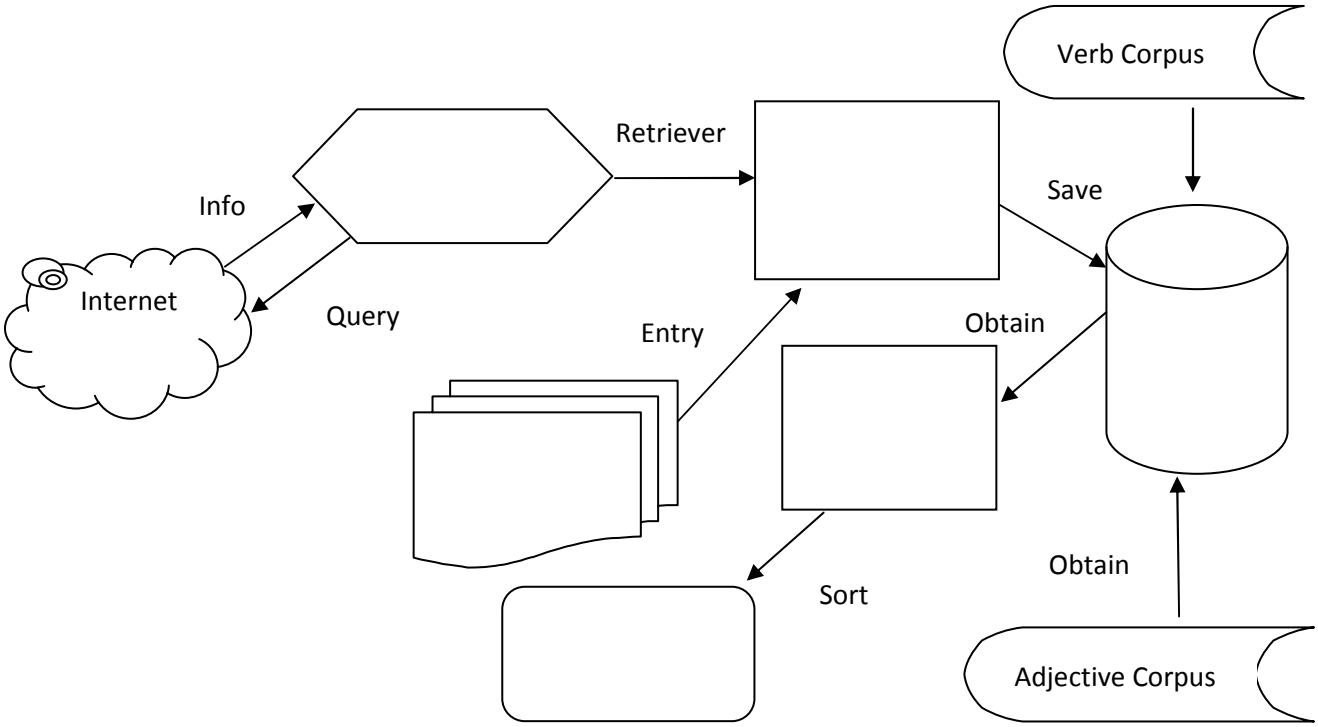


Figure 1.1 opinion mining process

In general, there are 3 processes of opinion mining that must be performed such as extraction, understanding, classification, and summarizing. The methods that are needed for feature extraction, sentiment classification, and opinion

summarization have already been targets of research in other areas such as "document classification and text summarization" (Dongjoo Lee, 2008).

In figure 1.1, search engine send out a query to execute information retrieval from global network which contains thousands of comments, blogs, and forums. Likewise, sending sentences or feeding data manually into extraction phrase is an alternative way to operate it. There are typically 2 opinion search queries. "One is search for opinions on a particular object or feature of an object and second is search for opinions of a person or organization." (liu, 2006) The goal of extraction process is to automatically extract multiple documents or data containing positive or negative opinions. Then, these multiple documents and data are stored into data mining system,within it we can find tons of adjectives, nouns, and verbs, corpus. The object usually refers to noun whereas the object's feature refers to adjective. The goal of classification is to classify the sets of adjectives obtained from the previous phase. They could either use supervised learning method or unsupervised method to classify or discriminate sources of positive opinions and negative opinions.

In order to allowfirst-time-users to have a better understanding of how it works. We can demonstrate it by using web crawler tool to gather raw data from forums or blogs. Then, we choose the approach of data mining techniques in SAS to retrieve the raw data. After that, we develop clusters after using SAS filters and Parsing Procedures to eliminated unwanted words. "It will show the percentage, frequency, and the descriptive terms of the words in the clusters after completing summarization result" (Stamper, 2008).

2.4.1 Pre-Processing

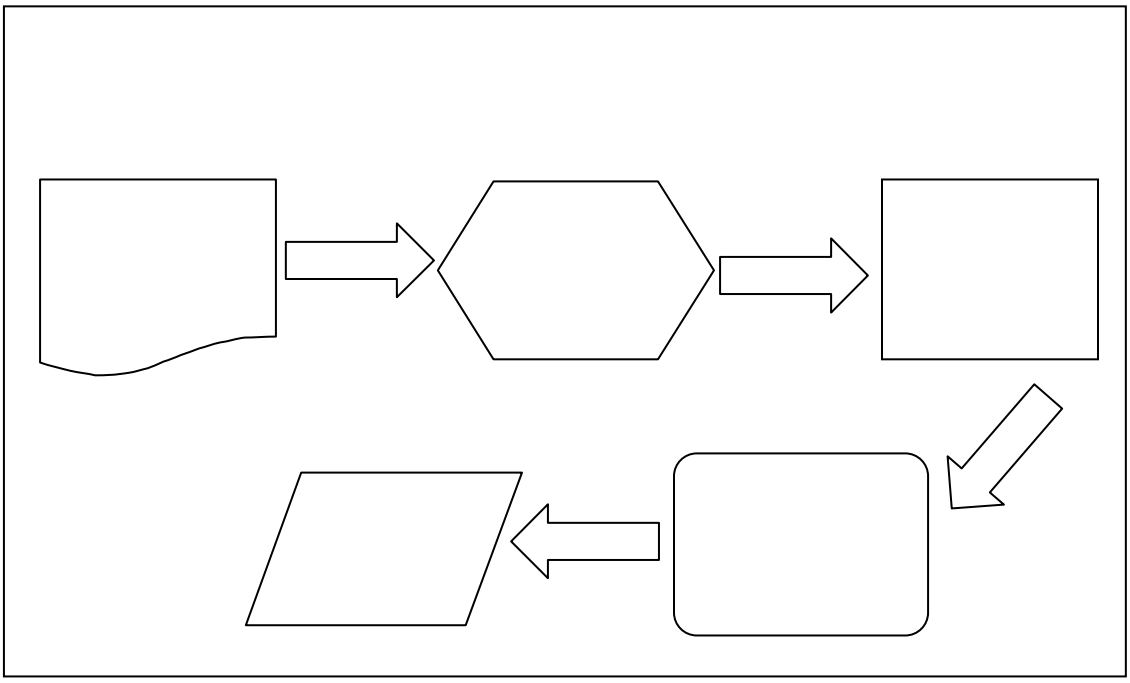


Figure 1.2 Opinion feature extraction

I. Opinion sentiment was obtained through blogsphere, forums, commercial websites, or other social network

II. By using POS tools such as Stanford-POStagger, TreeTagging, and Taggers to perform dividing and categorizing problem into noun, adjective, verb, conjunction, pronoun, and interjection.

III. Applying extraction algorithm to extract opinion and feature.

IV. Applying association rules mining to discard or prune unrelated opinion feature

V. Extract those opinions and features into sentiment classification and analysis the polarity of expression opinions.

2.4.2 Part-Of-Speech tagging

Part-of-speech tagging is also called grammatical tagging or open class words. It is a process of distributing a word in the particular part-of-speech in a corpus such as verbs, adjectives, nouns, and etc. “The adjective have been employed because it is discovered that high correlation between presence of adjectives and sentence subjectivity” (Lee, 2008). In fact, “somehow adjective is a good indicator of sentiment detection and has been used as indicator feature selection for sentiment classification” (Lee, 2008). POS tagging plays an important role for information retrieval and word sense disambiguation. Normally, the extract of opinion is noun or noun phrase from review sentence and opinion feature is adjective word. By the following example given, I denoted which POS trees will be extracted.

“Shotgun is fantastic”

“When I was learning how to shoot, I hoped that I won’t get there next time”

The primary method of POS is to extract adjectives from opinion reviews, if there is no adjective located, adverb will be the alternative way. In fact, “there are two types of manners of expressing opinion, typically implicit sentences and explicit sentences” (Ryu, 2009). The adjectives can be easily located when it is explicitly expressed opinions as the first sentence “fantastic” that has positive oriented and indicated author desirable with the opinionate. However, it is difficult to denote whether second sentence is positive or negative oriented reviews. But, we know that how the author feels because ofthe frustrated and

depressed emotions expressed during his learning period. While constructing parse tree, we need a tool to build up the "tree' and extract each opinion and opinion feature for classifying purpose and draw us into better understanding. If the sentences without adjectives, then this sentence will discarded and skip to the next sentence. There are a lot of available tools like Stanford-POStagger, TreeTagging, Taggers and etc. We could find many of free POS tools in merchant websites but some of them are trial versions for 30 days use only.

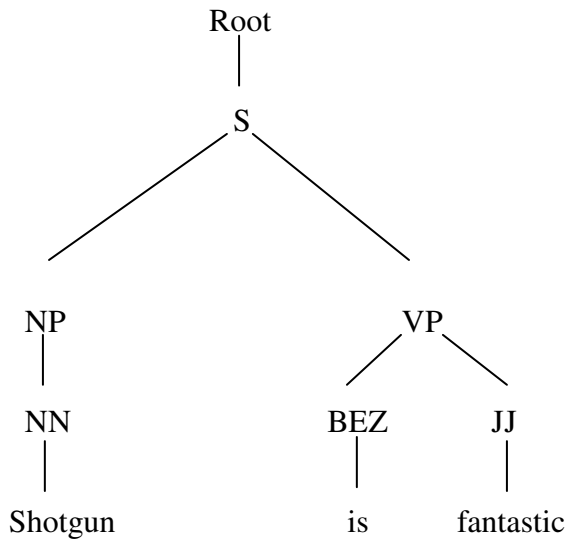


Figure 1.4 Parse structure tree

In figure 1.4, we can refer to these POS tags and definition by brown corpus for further reference. This is how the POS parser extracts product and product's feature from opinion review. The representation of algorithm dedicated in dividing noun, adjective, verb, conjunction, pronoun, and interjection in order

2.4.3Bookstrapping

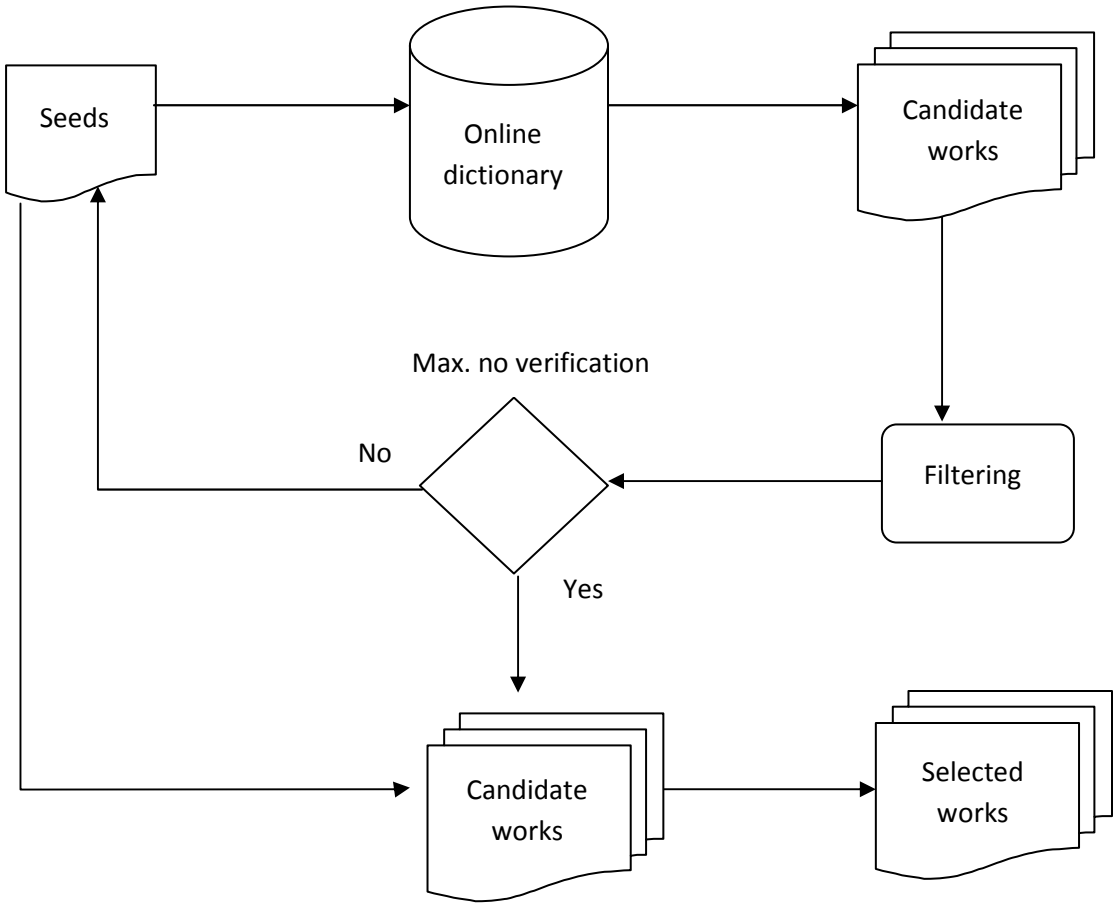


Figure 1.3 Bookstrapping Process

Bookstrapping method is a simple and effective process to obtain relevant data from holistic lexicon and Wordnet for a lexicon database inclusive synonyms and antonyms. Based on the figure 1.3 (Carmen Banea, 2008), we need to locate a set of seed words sampled from nouns, adjective, adverb, and verb. Seed list is used to extract the features and opinion words. It could be easy to acquire seed sets for any other international language. Dictionary possess a function of open-

class words contained many of definition such as holistic lexicon database. The process starts with a query sent for requiring dictionary to select appropriate related words and added it into the list of candidate words if obtained words have similarity with the definition of dictionary. However, if they do not have a similar tool with the dictionary, then the candidate words likely to compare with the original seed words continuously in the process of iteration until the maximum number of iteration is reached means that there are no similarity of definition found in the dictionary. The aim of filtering is to remove noise text from the set of seed words, and also performs filtering of similarity between original seed words and candidate words.

2.5 Machine learning

Machine learning is a study of computer algorithm that gets computer to act and improve automatically by learning from data. Such algorithm is being used to construct a model based on inputs given and leveraging statistical analysis to make prediction and decision without being explicitly rule-based programmed.

Many researchers used machine learning technique for identifying attributes and objects to generate summary results. Dave used "machine learning approach to classified opinionated document from web" (Kushal Dave, 2003). Pang used machine learning technique for sentiment classification (Bo Pang, 2002). There are 3 essential important machine learning algorithms such as support vector machine, Naïve Bayes, and maximum entropy.

2.5.1 Naïve Bayes

Naïve Bayes classifier is a "simple probabilistic classifier model based on the Bayes rule with a strong independence assumption between the features" (Vivek Narayanan, 2013). It classifies a class either in positive or negative, which means that the words are formed as independent feature model. One of the advantages of the Naïve Bayes classifies is that it requires less training data to estimate the best option of parameter for classification in opinion mining. Naïve Bayes classifier contains both high speed and optimal feature selection through data that have been put into the system.

2.5.2 Support vector machine

Support vector machine (SVM) is a "supervised classification technique which is based on maximum margin linear discriminants" (Banitaan, 2010). The SVM uses a "kernel function approach to map an input feature space into a new space where the classes are linearly separable" (Banitaan, 2010).

Table 8: Experimental results on opinion orientation prediction.

Feature set	Technique	Precision	Recall	F-score
$F8$	CRF	0.950	0.232	0.373
$F4$	SVM	0.976	0.488	0.650
$F4$	Naive Bayes	0.608	0.378	0.466
$F4$	Bayesian Net	0.535	0.458	0.494
$F3$	Random Forests	0.955	0.506	0.662

Figure 1.8 Comparison between 5 supervised learning machines

Svm give the best classification result should you have sufficient balancing training data. In figure 1.8, a study has shown that "SVM achieved the highest precision compared to others" (Banitaan, 2010).

2.6 Challenge

2.6.1 Noise text problem

Noise in text can be defined as informally written text and disparate between the surface forms or extended mean it is an unstructured text data constituted by informal setting. “Noisy text data typically comprises spelling errors, ad-hoc abbreviations and improper casing, incorrect punctuation and malformed sentences” (Lipika Dey, 2008). The main reason of causing noise text emerge in the form of chat transcripts, email, blogs, forums, and customer reviews is because of faster typing practice or semantic affection especially trying to minimize message length during seminar talks, discourses, or conference meetings. Many of the online chat users prefer to use short forms or abbreviation techniques such as “you” written as “U” and “friend” describe as “frens”. Some even uses numeric method for representing a word “too” is “2” and “night” resembles “9” because they have same syllable pronounces. Likewise, symbol can be used to represent certain meaning too. For instance, love is “❤”, grieve is “ =(“, happy symbol with “ =) ” , and so forth.

2.6.2 Spam opinion problem

Opinion spam is normally described as “non-behave action” especially to who is purposely try to misguide or mislead readers digress from the main topic of discussion. The spammers deliver some false or bogus sentences and words containing undeserved positive opinion in order to bring out negative opinion and cause damage to personal reputation by fulfilling their greed. These mostly happened to forums or blogs where are likely to post reviews or comment on merchant sites or political issues. “Such contents contributed by Web users is collectively called the *user-generated content* (as opposed to the content provided by Web site owners)” (Jindal, 2008). There are 3 types of opinion spam:

- Untruthful information, which allows untrustworthy information to be delivered to mislead readers in order to generate vicious opinions to be brainwashed and damage their reputation.
- Review on brand only, which treat brand product of priority of “Good”, “Excellent”, or “Nice” product can lead misunderstanding of customer reviews because they are not clearly targeted at a specific product and this may cause biased result.
- Non-reviews, there are 2 types of non-reviews spam which are advertisement and irrelevant reviews not containing opinion. For example, random text, question, and answer.

2.6.3 Idiom problem

Idiom is an expression of words or phrases that is comprehended in figurative language, containing metaphor or ironic sentences to strengthen the explanation of message delivering. It can also be identified with positive, negative or neutral idioms. For example, positive idiom such as “Content is better than riches” and negative idiom such as “Misfortune never come alone”. In fact, most idioms express strong opinions. For example “cost (somebody) an arm and a leg” (Xiaowen Ding, 2008). There are more than 25000 idiomatic expressions annotated in English language.

Non-opinion phrases containing opinion word are some sentences or phrases that have neutral opinions but they contain adjective words. For example, “pretty large”- where “pretty” is a positive opinion word, but the whole phrase has no opinion (Xiaowen Ding, 2008).

2.6.4 Document level problem

Document level can be defined as having a function to evaluate an aspect or feature of an object. In fact, it does not clearly reveal what opinions author expressed is desirable or undesirable. A positive oriented document on an object does not mean that the author has personal positive opinions on the object. On the same token, negative oriented document does not mean that the author dislikes the object. Document-level usually contains one or more than one expression

opinions. Hence, it is difficult to determine only by the first or second sentence and immediately conclude that this sentence is positive or negative opinions.

2.6.5 Sentiment classifying problem

"Sentiment classification is classifying an opinionated document as expression a positive or negative opinion."(Liu, 2010). Hence, sentiment classification is to determine whether a sentence is positive or negative oriented. A study of sentiment-analysis problem given a review segment on iphone as example (Hsinchun Chen, 2010):

1) I bought an iphone two days ago.
2) It was such a nice phone.
3) The touch screen was really cool.
4) The voice quality was clear too
5) However, my mother was mad with me as I did not tell her before I bought it
6) She also though the phone was too expensive, and wanted me to return it to the shop….

The first thing we may notice is that it is consisted of 6 opinions or reviews sentences. First sentence is a neutral opinion, author just describes what object he bought and how great it is. The second, third, and forth sentences are positive opinions. From here, we can find that 3 of these sentences are using positive adjective words such as "nice", "cool", and "clear". However, fifth and

sixth sentences expressed negative opinions or emotions. The chief of objects has been targeted to author himself and his mother in fifth sentence. We can also find that the negative adjective word of “expensive” is shown in sixth sentence and the main object displaces once again to phone.

2.6.6 Domain specific problem

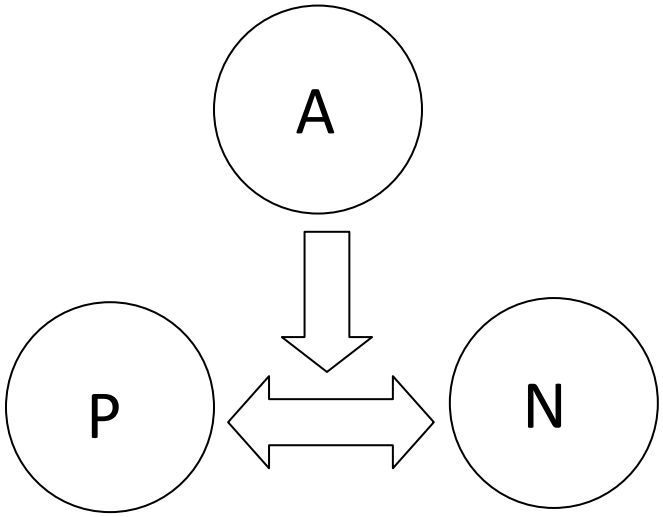


Figure 1.5 Domain specific words

A=certain adjectives (long, short, thick, thin) P=Positive, N=Negative

Domain is a specific word or phrase that can show complexity and yet unspoken. “One reason is that the same phrase can indicate different sentiment in different domains: recall the Bob Bland example mentioned earlier, where ‘go read the book’ most likely indicates positive sentiment for book reviews, but negative sentiment for movie reviews” (Lee, 2008). Another major problem is

related to some adjectives such as “simple”, “easy”, “long”, and etc. For instance, “The battery life can be remained for a long time” is a positive sentiment whereas “I’ve been waiting for the lift so long” expressed a negative sentiment. A sophisticated technique of detection must be employed to capture the specific term of adjectives in order to distinguish specific domains that could provide different meaning.

2.7 Feature selection

In computational linguistic it is emphasized about how to make an important decision for classification based on a document. The selection of determination opinion words is either positive or negative oriented. Feature vector is a characteristic of a specific object. But one or more features are commonly stored into feature vector before it starts to evaluate an object.

2.7.1 Unigram

In general, “This is the classic approach to feature selection, in which each document is represented as a feature vector, where the elements indicate the presence (or frequency) of a word in the document. In other words, the document is represented by its keywords.” (Erik Boiy, 2007) It normally refers to one letter or one word. This means that the features in the document is represented a single word.

2.7.2 Bigram

Bigram refers to pair of words or letters. It could be 2-gram sequence if characters such as “AB”, “CD”, or “EF” are used to represent, whereas in computational linguistic manner it looks like “to be”, “not yet”, “not be”, and so on. All these are neutral oriented words do not provide any opinions. There are also positive oriented bigramsuch as “well done”, “pretty good”, and “good job”. However, the terms of negative oriented bigram have the examples of “bad luck”, “no values”, and “quite expensive”

2.7.3 Trigram

Trigram refers to triples of words or letters. It can be 3-gram sequence if characters areusedto represent as “ABC”, “CDE”, or “FGH”, whereas in computational linguistic manner it supposed to be look like “to be or”, “not to be”, “soon to be”, and so on. All these are neutral oriented words that do not provide any opinions. Likewise, trigram also has positive and negative oriented words just like what was shown in bigram. Other than that, for a larger size sequence of words can be found in four-gram, five-gram, and N-gram unknown forecast values.

2.7.4 N-gram

A word N-gram is a contiguous sequence of N from a given sentence which is a value that is used to forecast the evaluation. N-gram is possible to capture one or more than one sentences compared to other sample sequence. “When using N-grams, the feature vector could take on enormous proportions (in turn increasing sparsity of the feature vectors). Limiting the feature vector size can be done by setting a threshold for the frequency of the N-grams, or by defining rule sets” (Erik Boiy, 2007)

2.7.5 Lemmas/stems

Lemmatisation is the process of determining the lemma of a specific word especially in computational linguistic. This means that every single word has its base in the form of lexicon. For instance, the word of “swimming”, “swam”, “swims’ has “swim” as its lemma, the root form of “good” is lemma to the word of “best,” and “better”. Even though lemmatization is able to classify all features to their basic form that is adjacent to dictionary form, sometimes it might be inflect a particular meaning of words, making it a hard task for sentiment classification because in English it comprises different aspects of tenses like part tense, past participant, present tense, and present past participant. In this case, it is very complex to classify a sentence either as positive or negative oriented because the differences between the tenses could possibly generate disparate opinions based on certain circumstances.

2.7.6 Negation

Handling negation can be very important as it relates with opinion orientation classification. Especially to words such as “Not”, “No”, and “But”. For example, “This is not great” and “I don’t like the car” is considered as negations. We have not only noticed about the word “Not” or “no” offers an opposite meaning, but also concern in the reverse proposition of sentences may contain sarcasm or ironic phrases that can be quite difficult to detect.

2.8 Standard evaluation measures

Overall positive and negative content generally is measured using the standard evaluation measures of precision, recall, and score. In a classification task, precision for a class is the number of true positive divided by the sum of true positive and false positive. Whereas, recall for a class is the number of true positive divided by the sum of true positive and false negative.

	Actual Positive	Actual Negative
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

Table 1.6 illustrated the actual and predicted class as below.

The formula of precision and recall are defined as below:

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

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

Fscore is a compound relationship between precision and recall. Both of these are giving almost the same weight measurement. FScore and Accuracy are defined as follow:

$$\text{FScore} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$$

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

2.9 Summary chapter

The incessant growth of social media such as Facebook allows whole world “netizens”to contribute information without borders, share opinions about current issues, even their attitudes towards life. Thus, opinion mining is one of most well-known field of study nowadays.

Social network websites are increasing around the world and trying to drive or attract more users for developing social relation and human interaction in society. People like to share information regarding their opinions, state of mind, lifestyle, emotions, and sentiment expression via different social media.

Sentiment classification classifies opinionated documents into positive or negative categories. There are 3 essential important machine learning algorithms - support vector machine, Naïve Bayes, and maximum entropy.

Chapter 3: Research Methodology

Introduction

Datasets

Sentiment sentence labelling

Baseline tools

Supervised learning

Unit Testing

Summary chapter

3.0Introduction

This chapter describes the baseline method used to conduct the research study in order to achieve the expected outcome. In this section (3.1), I will talk about recommended proposed solution for sentiment analysis. In chapter 2, I have prepared several studies of natural language processing method and numbers of dataset to be used for sentiment analysis. In section (3.2 & 3.3), indicating the ideal machine learning approach and programming tools are chosen for classification. It is critically important to know that sentiment classification can be very complicated. Many approaches of classification are dependent entities on data provided.

3.1 Datasets

I used aset of data to carry out an experiment to evaluate my proposed technique. These data can be collected frompublisher websites that providewell written sentiment analysis papers business industry use.

3.1.2Mejaj dataset HS+

Publicly available corpus polarity dataset HAS+ has been selected to be trained in my system, which consists of 300,000 positive and 300,000 negative processed reviews. This dataset can be easily downloaded fromhttp://nibir.me/projects/mejaj/datasets.html. The dataset is used to classify

emotional feelingsinto positive and negative categories by conducting machine learning techniques.

3.2 Sentiment sentences labeling

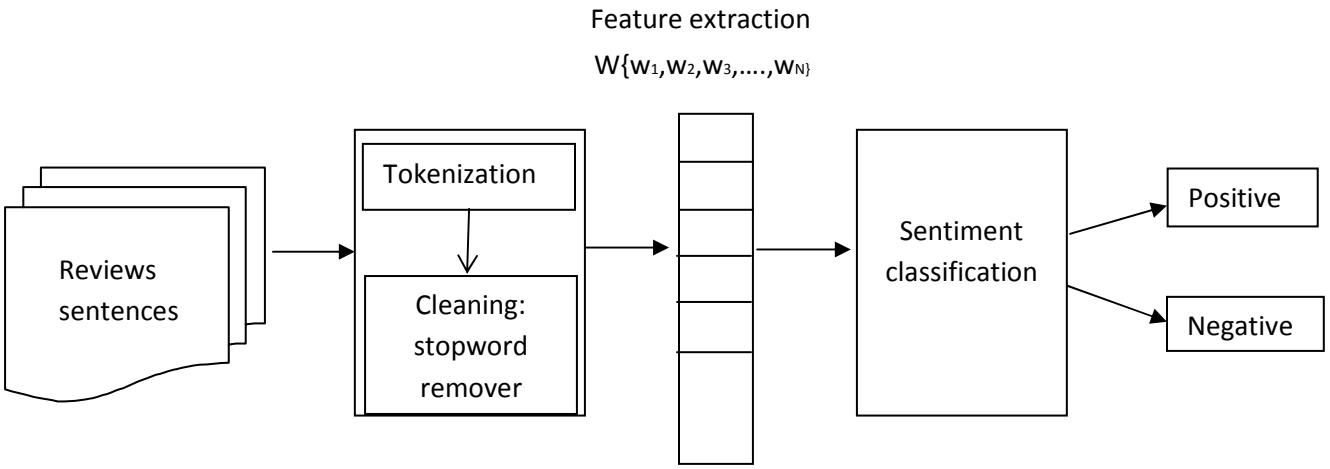


Figure 1.7 Overall sentiment classification processes

In figure 1.7 shows all the phases of sentiment classification. By performing information retrieval from sentences, input may contain a lot of opinion words, personal thoughts, movie reviews, and product features. An opinion sentence can have one or more product features, and one or more opinion word. Opinion words are those words that are used to express opinion about human thoughts. Different persons certainlydo not have similar perceptions about one thing. Hence we can assume that it comes with two perspectives, one is opinion words without product features, and the other is opinion words in general form. The presence of opinion words in a sentence can be very useful

topredictsemantic orientation. Hence, the aim of feature extraction is vital phrase to extract multiple sentences that contain tons of positive and negative opinion words.By using Natural language processing tool to perform feature extraction to transform sentences level into opinion words like good, bad, awesome, and great. Thereafter, features identification has to work in correspondence with features vector to identify the opinion orientation. Feature vector is the characteristic of specific product. For instance, objects referring to movie and feature vector of movie are such as “quality of image”, “storyline”, “Sound effect”, “Movie effect, and etc. The aim of machine learning approach is to classify sets of opinion words associated with feature vectors. Sentiment classification would be used to classify and discriminate sources of positive opinions and negative opinions.

3.3 Baseline tools

Machine learning approach for sentiment classification has been studied by many of scholars with diversity semantic analysis programs. I selected python programming tool as my sentiment classifierbecause python provides higher code-based customization compared to other sentiment analysis tools. The accuracy of performance result was tested by using python built-in tools such as natural language toolkit which comprises a lot of supervised learning and unsupervised learning approaches which are used in practice.

The machine learning approach for opinion mining was created andaccepted in its use by many researchers and scholars. I choose Naïve Bayesclassifier as my artificial intelligent tool for the development of this system.

It was developed by Vapnik since 1963, as known as linear classifier. The program that I used to generate visualized view of data was Rstudion system which transforms the numeric results into meaningful color of plot graph. The system development can be separated into several steps. First of all, preparing a dataset with samples of 500, 1,000, 3000positive and negativelabelled. Second is chosen a desirable program that used to consume all the provided data. Hence, I used python with built-in NLTK tool and Rstudio to develop the system.

3.3.1 Python

Python is widely used nowadays in a wide scope of range from education to industrial fields. It is well-known as high-level programming language. It has gained popularity and entices more programmers to use it because of its clear syntax and readability pieces of code that makes it possible in facilitating programmers to know and understand more easily.

The overall research would be demonstrated by using python, which is a programming language associated with machine learning features that have the potential to perform various capabilities of pre-processing step on the basic properties such as extraction, filtering, and stopwords.

Besides, it also offers validation, evaluation, classification, and regression. All these are important to sentiment orientation. Python also providesthe capabilities of generating standard evaluation measures, it would show the accuracy, precision, recall and descriptive terms of evaluated opinions.I replacedthe obsolete console interface with graphic user interface (GUI) by using

Tkinter function, which offers more visualization and its benefit of being user friendly.

3.3.1.1 Python for tokenization process

All processed data file from excel format is stored into a text file and placed on a location for testing purpose. This is done by manually selecting 500, 1,000,3,000positive and negative from huge mejaj corpus polarity dataset. In python, you have to retrieve the entire positive and negative labelled set from file directory stored at a location. Python would then retrieve and open and read the data automatically. The n.readlines() means that to find all sentences within the document that represent each of positive and negative text file.

The following are the codes indicatehow the documentis to be pointed to positive and negative text files. It is important to create a list of positive and negative categorieswith exact length of data list.

```
for i in range(0,len(negSen)):
    negWords.append('negative')

for i in range(0,len(posSen)):
    posWords.append('positive')
```

Figure 1.8

In particular, Text files need to be pre-processed and split into a sequence of tokens before applying to supervised learning approach. So the

declaration of tokens should be done at the early stage. The declared tokens would therefore create a list of words by joining both positive list and negative list into a large list. The i.lower() functioned likechanging all the possible capital letter of word into lowercase in order to reduce fuzzy duplication of tokens. In addition,The re.findall(r"[\w:;<>/\()=-_^]+") is used to filter unneeded punctuation like fullstop, coma, question mark, and exclamation mark out from sentiment analysis that might influence the accuracy of result. But remains symbols can be used to represent certain meaning.For instance, happy is ":)", grieve is ":(",and so forth.

```
tokens = []

for (word, sentiment) in labelledtokens:
    filtertokens  = re.findall(r"[\w:;<>/\()=-_^]+", word.lower())
    tokens.append((  filtertokens, sentiment))
```

Figure 1.9

From the above 1.9, we can see that there are filter tokens being applied to the system. The purpose of using filter function is trying to split out the words based on their length. Which means that the number of characters they contain. Eg: before transformation "the phone battery life is short", but after being transformed {the, phone, battery, life, is, short"}.

3.3.1.2 Python for stopwords process

Stopwords are words that commonly found useless. It is a bag of words assembling together to create stopword list which are later being used to filter out

tokens which equips to the built-in stopword list. Natural language toolkit provides stopwords corpus that includes 128 englishstopwords. In my proposed cleansing task, I chose English stopword operator to filter unnecessary tokens out from my sentiment analysis. In figure 2.0, it is shown that the wordlist resultscollected from the tokens are filtered out by eliminating the tokens equivalent to the built-in one.

```
def all_of_result(words):
  return dict([(word, True) for word in words if word not in set(stopwords.words('english'))])
```

In figure 2.0

3.3.1.3Python for feature extraction

For the feature extraction process, python will read andprocess all the files attached which are with positive and negative tags. Python thenproduceall of the tokens in a list of words.With its built-in natural languages toolkit frequency distribution to capture those occurrence frequency tokens that appeared in the words.

```
def gettokens(tokens):
    alltokens = []
    for (words, sentiment) in tokens:
        alltokens.extend(words)
    return alltokens


def getfeatures(freqoccured):
    tokensfreq = nltk.FreqDist(freqoccured)
    words = tokensfreq.keys()
    return words
```

Figure 2.1

3.3.1.4 Python with bigram collocation

Bigram refers to pair of words or letters. It could be 2-gram sequence if it is used characters to be presented as “AB”, “CD”, or “EF”; whereas in computational linguistic manner it looks like “to be”, “not yet”, “not be”, and so on. All these are neutrally oriented words that do not provide any opinion. Somehow there is a positive oriented bigram such as “well done”, “pretty good”, and “good job”. However, the terms of bigram like “bad luck”, “no good”, and “quite expensive” contain negative expressions. Unigram will present the result as “no good” into positively oriented opinion due to its single word depicted as “good”. As a result, I decided to use bigram collocation to measure the accuracy of overall performance in my system.

3.3.2 R project for statistical computing

R program is a very useful tool for statistical computing and graphical display analysis. It provides a wide variety of statistical model like classification, regression, clustering, and most beneficial features are resides in its graphical techniques. One of R biggest advantage is its well-designed graphical plots with many choices that can be produced effectively and efficiency.

The accuracy, precision, and recall results would be stored into text file manually. The return results would show 5 samples dataset to be tested in the system. All the gathered results were then multiplied by 100 in order to get the percentage of overall performance.The higher percentage of sample means that the system could provide higher accuracy to identify and classify the sentiment

orientation of words. The result would be put into Rstudio in order to display a graphical plot view. This is because there is no graphical plot available in python. In figure 2.3, it shown that a piece of codes to trigger the plot view from R project.

```
library(ggplot2)
dataset <- c(1,2,3)
samples <- log(dataset)
accuracy <- c(61.08,61.48,60.55)
data <- data.frame(samples, accuracy)
ggplot(data, aes(x=samples, y=accuracy)) +
geom_point(aes(colour = accuracy))
```

In figure 2.3

3.4Supervised learning

In this research, Naïve Bayes classifier is playing two important roles.First, it is used as a tool to classify the positively and negatively oriented opinion sentences. Secondly, it is used to determine the accuracy of result that it has generated. It classifies a class either in positive or negative. For example, a public vehicle may be considered to be a bus in terms of size, seats, and wheels. Naïve Bayes classifier considers all of these properties to generate a series of probabilities to identifying that this vehicle is bus.

One of the advantages of the Naïve Bayes classifies is that it requires less training data to estimate the best option of parameter for classification in opinion mining. Naïve Bayes classifier contains both high speed and optimal feature selection through data that have been put into the system.

I managed to use python for developing a machine learning approach to opinion mining system. The accuracy of Naïve Bayes classifier is tested by using python programming tools with built in natural language toolkit. Python possess the built-in nltk that providing the capabilities to process the data via supervised learning techniques. It cannot be denied that python has many sentiment analysis properties for letting users to have a better experience of semantic analysis.

Training data and testing data are used to test the metric on how well the feature selection works via sentiment classifier. These are processed data used to verify the accuracy of sentiment analysis. The result in decimal would then be converted into percentage before applying to Rstudio statistical model. Those processed data would be fully processed through python. The data is separated into 2 parts. One is to extract ¾ features work as a training data, and the second one is to extract ¼ features work as testing data. From classifier, I used Naïve Bayes provided by python built-in classifier for sentiment classification.

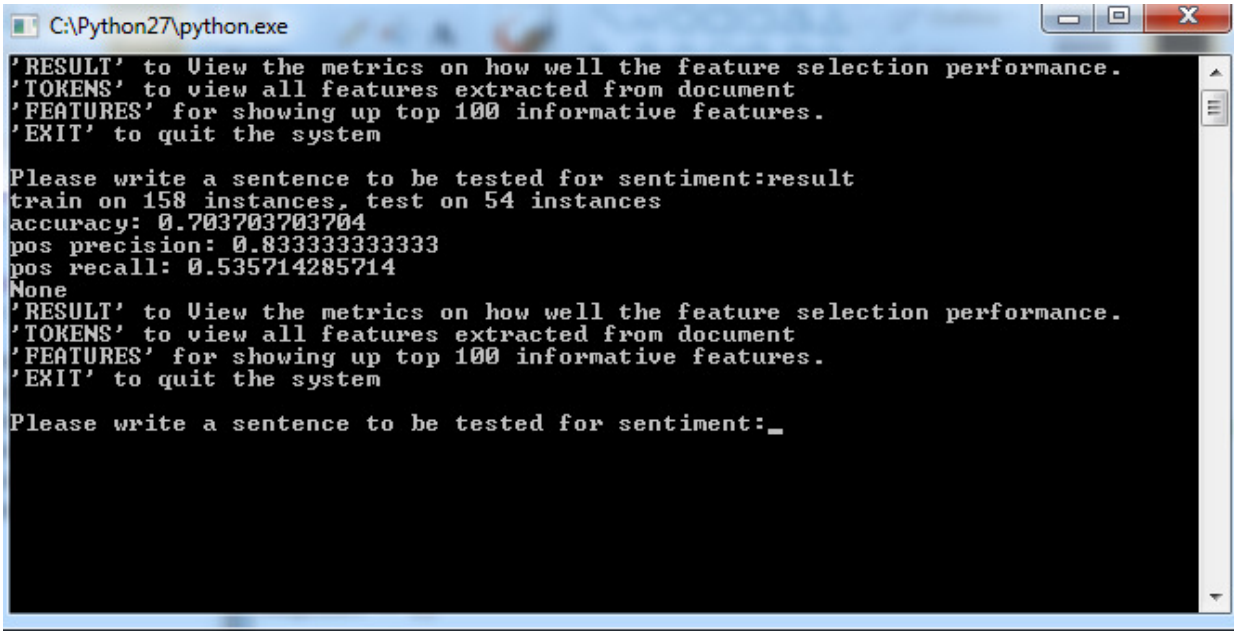
3.5Unit Testing

The purpose of unit testing is to ensure that each unit or module functions properly and works in accordance with program specification. Most of the time programmers will test the individual program. Execute the program using certain data values to ensure that the program handles input and output correctly. Areas of test identify and eliminate execution errors, logic error, and syntax errors. Unit test is commonly focuses on the verification and validation on every single module.

Unit test is also called module testing. The modules will be tested to ensure that there are no existing errors during the run time. Tests on all possible cause of errors such as input data, report, and options selectedare performed in order to minimize the error occurred when the program is running.

There would be a unit testing for the opinion mining by using machine learning approach for sentiment classification. The function of Naïve Bayesis to classify the extracted features from words and then utilized linear coefficient for generating analysis results.

The overall test process is exercised thoroughly with python programming tool where executing a series of codes to read and discretize words in sentences into individual tokens. It isthen tested on machine learning-based approach. The accuracy, precision, and recall of positive and negative result, and the working of coding were shownas follows.



In figure 1.8, Example result of standard evaluation measures

We store the actual set into the index numeric which is known as training label, and also stored the predicted set into the index numeric which is known as test set. If classifier predicts as 'pos' but actual sentiment is 'neg' in training set, the index will be 'neg' in actual set and 'pos' in predicted set.

The formula of precision and recall are defined as below:

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

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

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

This result was generated by using 1,000 positive and 1,000 negative sentiment expressions from mejajcorpus polarity dataset HAS+. However, this was just part of the result generated by the system. The aim of testing is to ensure the accuracy of machine learning approach contributing to the data analysis. Therefore, it is not necessary to use tons of dataset to prove the overall result.

3.6Summary Chapter

In this chapter, the experimental methodology in the research was discussed further. The study used integrity dataset and supervised learning approach to identify and classify sentences into positive and negative opinion in a sentence.The next chapter are focused on the discussion on how to conduct the experiment result, performance analysis, and comparative discussion.

Chapter 4: Experiment and Result

Introduction

Experimental setup

Experiment result

Performance analysis

Comparative analysis

4.0Introduction

This chapter describes how the experiment has been implemented in python. Followed by the application of the proposed evaluation measure used for assessinglarge stream of opinions, they are manually selected and analyzed of500,1,000, 3,000positive and negative respectively. The evaluation is separated into two sections. First, test on capabilities of class pre-processing step like individual word, stopword, and bigram collocation. Second, test on accuracy of classifier. Naïve Bayes and support vector machine are usedthroughoutthe whole processes. Lastly, comparison between dataset and different classifier are carrying out to test the influence of variables to the outcomes.

4.1 Experimental setup

4.1.1 Experimental data

In my research, Iusedmejaj corpus polarity dataset to be applied in my system. These dataset can be easily obtained from respective websites link http://nibir.me/projects/mejaj/datasets.html. The dataset is used to classify emotional review in aspect of positive and negative with conducting of machine learning techniques.

Overall positive and negative contentsare measured using the standard evaluation measures of precision, recall, and score. Fscore is the compound relationship between precision and recall. Both of these are giving almost the same weight measurement. It is crucial to evaluation measure stage when

aggregating all labeled sentence level with positive and negative. Thus, it could give an assessment on accuracy of system conducting the result. Data that has been downloaded must be separated into positive tag and negative tag before passing to the system for feature extracting process. Data that are running on the system can be in various forms such as500, 1,000, 3,000positive and negative.

4.1.2 Experiment process steps

In this research, I did my research by seeking and analyzing through all relevant research or experiment that have been done previously. Based on criteria they made, I look forward to the combine different phases into an integrated work procedure by understanding the procedure of research and discovering which techniques can generate decent results. Pre-processing is necessary step to be performed before applying document into classification phrase. It involved tokenization, andstopword (the, at, is, which, and on) filtering. However word sense ambiguityonly being usedin Wordnet method, which is likely to be a dictionary-based approach in discovering its opinion orientation. But here, supervised learning method is used to perform feature extraction automatically through classifying process.

The aim of this project is to mine out and determine polarity of positive and negative reviews. Therefore, sentiment classification is more crucial for my research purpose compared to the others (feature-based classification and comparative sentence). Also, between supervised and unsupervised classification, I applied supervised learning techniques to classify reviews into positive and

negative associated with machine learning approach. There are 3 supervised learning methods available such as Naïve Bayes, support vector machine, and maximum entropy. I usedNaïve Bayesin my system. In supervised learning, Training and testing data are needed to discover potential predictive relationship and train knowledge into machine learning. Thus, a standard evaluation measureis accurate and is to be used to appraise the performance of the system.

4.1.3 Pre-processing step

4.1.3.1 Tokenization

Tokenization is a pre-processingstep of splitting the text of a document into a sequence of tokens before applying machine learning model. It is useful for languages like English and other languages as well. The main reason of separating the text from document is to assure that theempty spaces are eliminated, andunusable punctuations are totally removed. Here is the example of how we transform the text of a document into every single token. Eg: “the phonebattery life is short” into {the, phone, battery, life, is, short”}. In my proposed method, I used the tokenization tools provided by python with built-in capabilities to split the text out ofpunctuation and empty spaces. This will help to generate tokens into one single word and facilitates the next stopword removal process.

4.1.3.2Stopwords

Stopwords are those commonly found words assembled together to create stopword list which are later being used to filter out tokens equivalent to the built-in stopword list. There is no single stopword list that can be used to filter all the languages. It relies on which type of stopword language you might want to use in your system. In my proposed cleansing task, I chose English stopword operator to filter unnecessary tokens out from my sentiment analysis. The giving example of English stopword list would as shown below figure 1.8.

Stopwords		
a	it	these
about	its	they
again	itself	this
all	just	those
almost	kg	through
also	km	thus
although	made	to
always	mainly	upon
among	make	use
an	may	used
and	mg	using
another	might	various
any	ml	very
are	mm	was
as	most	we
at	mostly	were

Figure 1.8 English Stopwords(Stopwords, 2001)

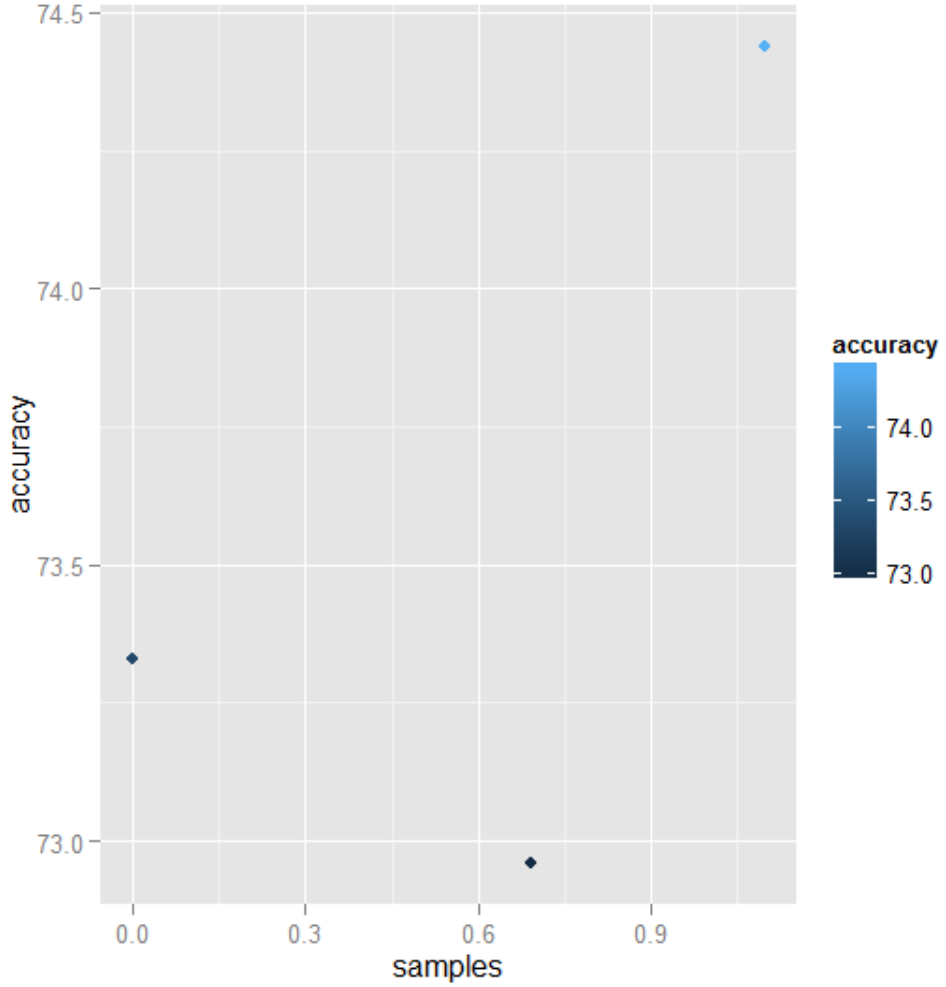
4.1.4 Classification tool –Naïve bayesclassification

Many research studies have shown that sentiment analysis achieved higher precision compared to others when Naïve Bayes approach is used.One of the advantages of the Naïve Bayes classifies is that it only requires less training data to estimate the best option of parameter for classification in opinion mining. Naïve Bayes classifier contains both high speed and optimal feature selection through data that have been put into the system. In my supervised learning, I used Naïve Bayes classifier provided by python built-in classifierfor sentiment classification.

4.2 Experiment result

The measurementevaluation is separated into two sections. First, test is to be done on capabilities of pre-processing approaches about whether it affects the classification accuracy or not. This includes individual word, stopword, and bigram collocation. Second, comparison is provided between Naïve Bayes and support vector linear classifier.

4.2.1Performance analysis on 1000 samples with naïve bayes classifier

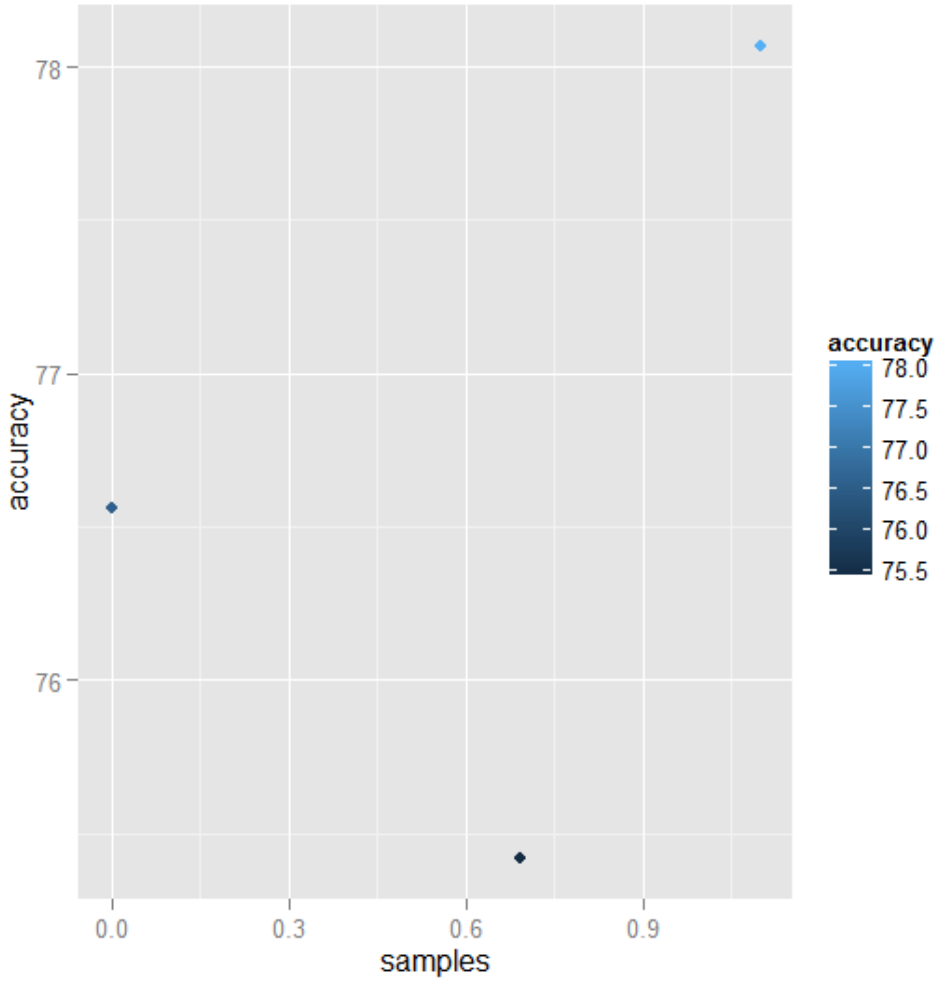


	Individual Word	Stopword	Bigram
Accuracy	73.33	72.96	74.44

The above figures have been multiplied by 100 in order to get the percentage of overall result.

Three color dots represent the probability percentage of accuracy on sentiment analysis with different approaches applied in the system. A result with 1000 samples was analyzed and it has been proven thatbigramprovides higher accuracy followed by individual word and stopword.

4.2.2 Performance analysis on 2000samples with naïve bayes classifier

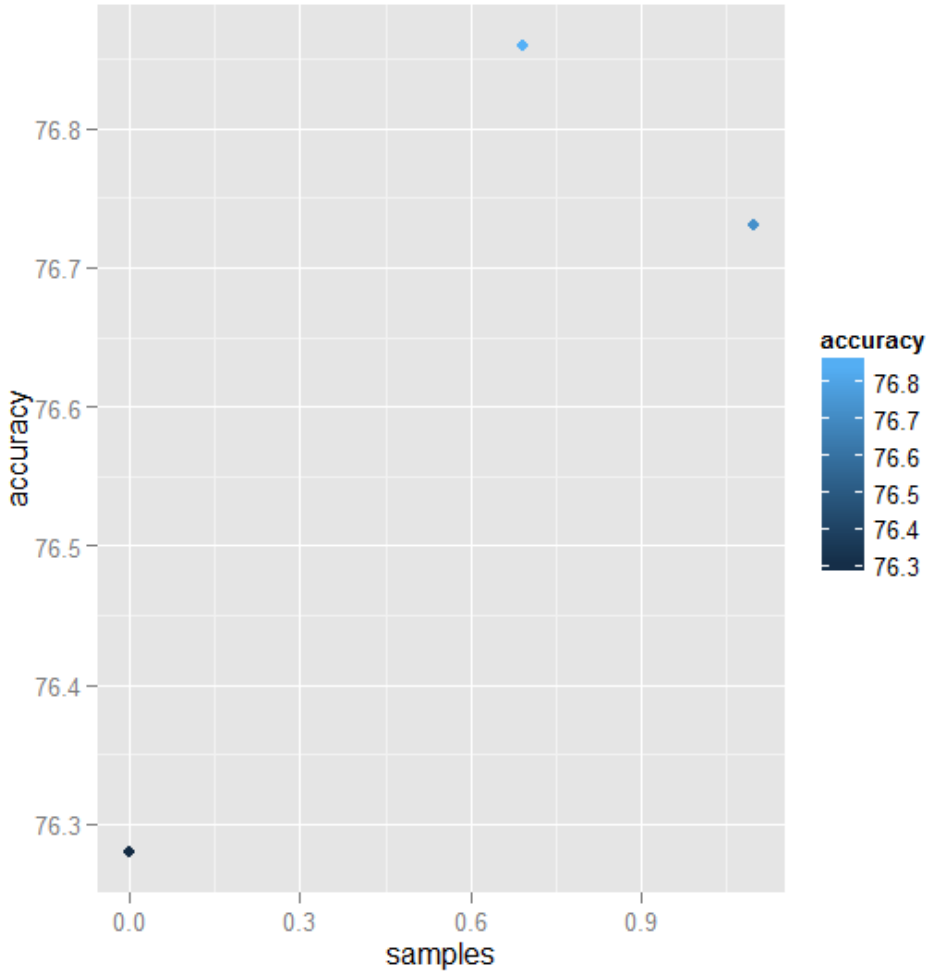


	Individual Word	Stopword	Bigram
Accuracy	76.56	75.42	78.07

The above figures have been multiplied by 100 in order to get the percentage of overall result.

Three color dots represent the probability percentage of accuracy on sentiment analysis with different approaches applied in the system. A result with 2000 samples was analyzed and showed that the higher result is bigram followed by individual word and stopword.

4.2.3 Performance analysis on 6000 samples with naïve bayes classifier

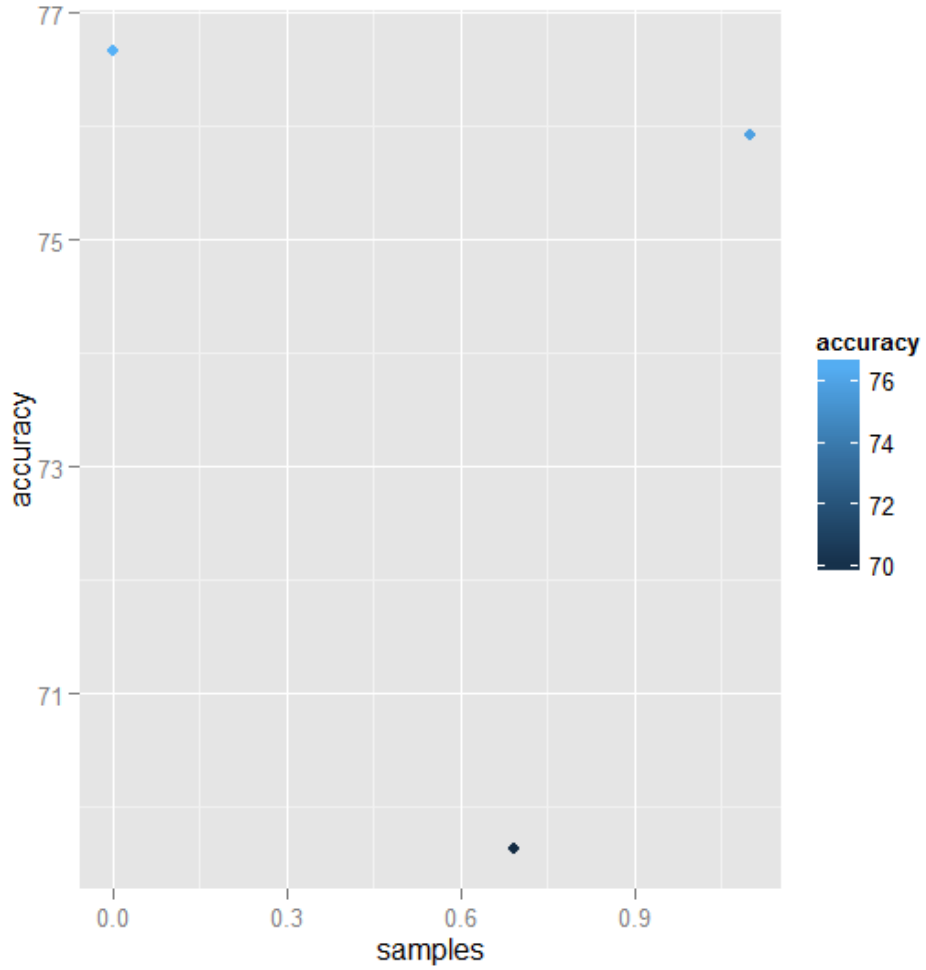


	Individual Word	Stopword	Bigram
Accuracy	76.28	76.86	76.73

The above figures have been multiplied by 100 in order to get the percentage of overall result.

Three color dots represent the probability percentage of accuracy on sentiment analysis with different approaches applied in the system. A result with 6000 samples has been analyzed and it is also proven that the higher accuracy is obtained fromstopword followed by bigram and individual word.

4.2.4 Performance analysis on 1000 samples with support vector linear classifier

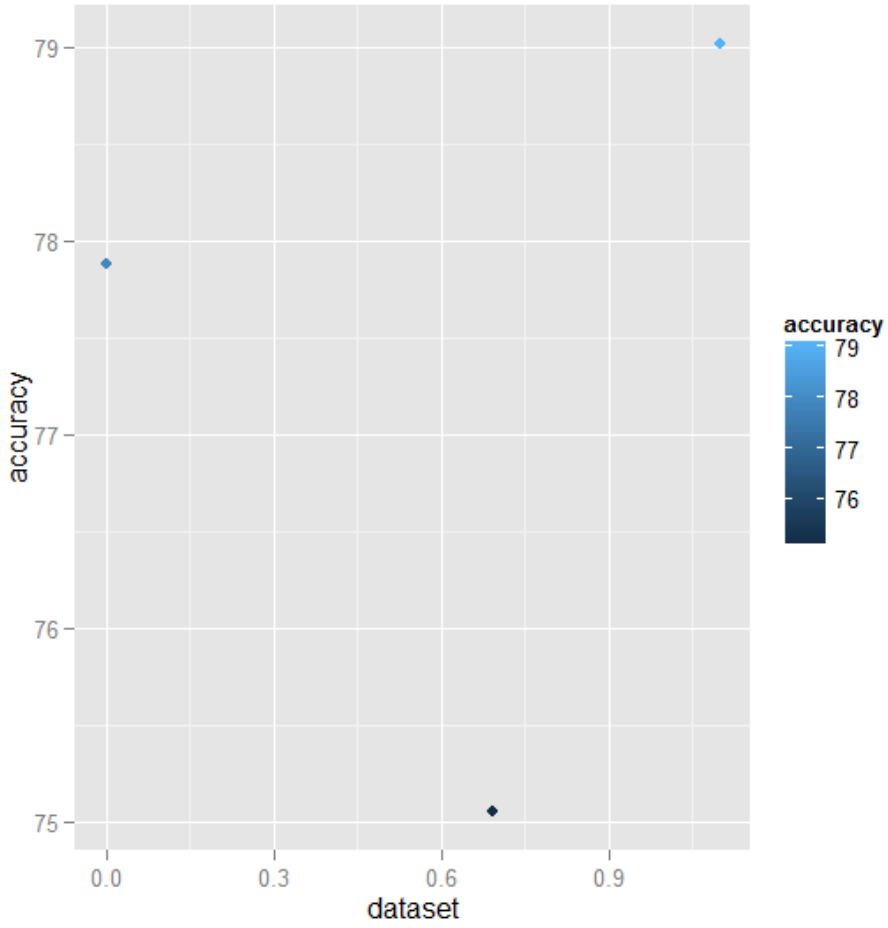


	Individual Word	Stopword	Bigram
Accuracy	76.67	69.63	75.92

The above figures have been multiplied by 100 in order to get the percentage of overall result.

Three color dots represent the probability percentage of accuracy on sentiment analysis with different approaches applied in the system. A result with 1000 samples was analyzed and showed that the higher result is individual word followed by bigram and stopword.

4.2.5 Performance analysis on 2000 samples with support vector linear classifier

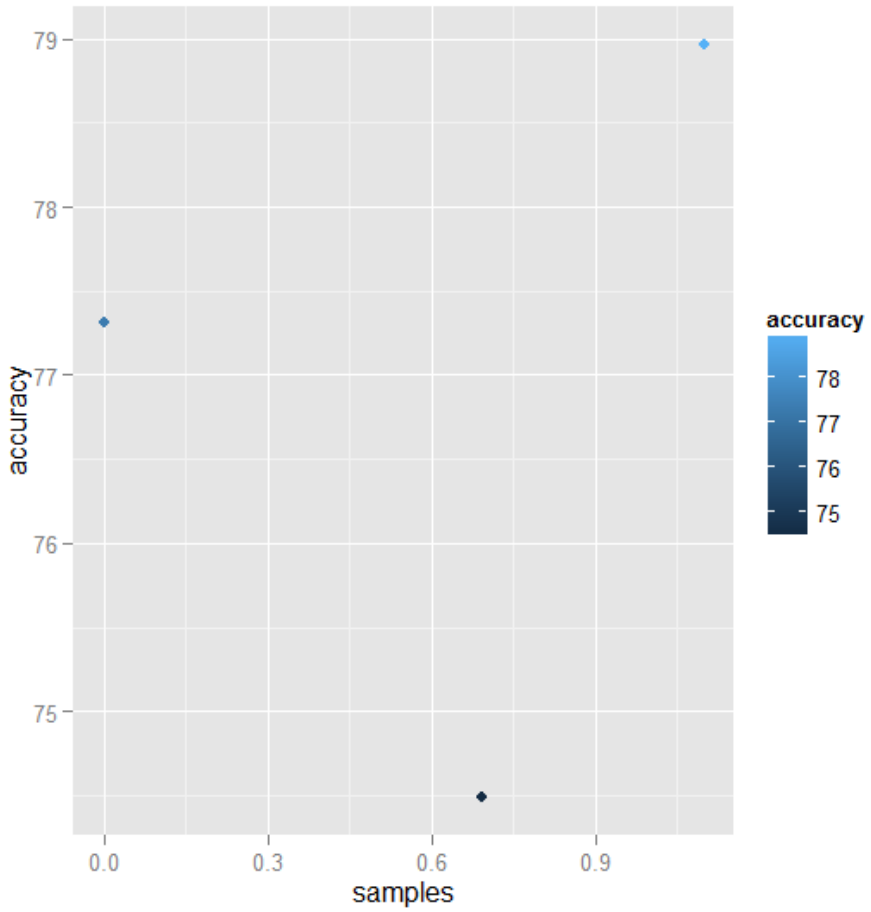


	Individual Word	Stopword	Bigram
Accuracy	77.88	75.05	79.02

The above figures have been multiplied by 100 in order to get the percentage of overall result.

Three color dots represent the probability percentage of accuracy on sentiment analysis with different approaches applied in the system. A result with 2000 samples was analyzed and showed that the higher accuracy is obtained frombigramfollowed by individual word and stopword.

4.2.6 Performance analysis on 6000 samples with support vector linear classifier



	Individual Word	Stopword	Bigram
Accuracy	77.31	74.49	78.97

The above figures have been multiplied by 100 in order to get the percentage of overall result.

Three color dots represent the probability percentage of accuracy on sentiment analysis with different approaches applied in the system. A result with 6000 samples was analyzed and is shown that the higher accuracy is obtained from bigram followed by individual word and stopword.

4.3 Result discussion

Python program would be used together with sentiment analysis. The processed data is analyzed by python built-in nltk classifier. For testing purpose, in order to boost up the processing speed, the processed data should be separated into 500, 1,000, 3,000 positives and negatives sentences. When the program runs, it would only need to process on the said data but not all the 600,000 samples at once.Otherwise, it will make the work become tedious and cumbersome.

All the processed data from excel format would be stored into a text file for testing purpose. Further,Rstudiois used to drawgraphical plot for probabilities result of each performance. All the gathered results would then be multiplied by 100 in order to get the percentage of overall performance.Next, mass amount of results would be pumped into Rstudioto be computed statistically and generate a plot view diagram.

From 6 diagrams mentioned above, they led us to a delve insight and noticeable increase of result accuracy when applying pre-processing step prior to the system at proper manner. Pre- processing also explains that cleanse data is a very important step before feeding data to classifier to analyze and identify their sentiment orientation.

Note that the experiment reported in respect of the 2,000 samples shows that individual word achieves exceptionally higher result than the others. This is probably caused by the content of data, which has verysmall amount false positive to be featured. Also, it might bebecause of the fact that there is less noise in sentences found in those 2,000 samples.

The overall accuracy for sentiment analysis withsupport vector linearclassifier has achieved better result compared to naïve bayesapproach. However, Naïve Bayeswith stopwordcan increase the accuracy of sentiment analysis when there was an increment of samples took place. This means that the greater amount of data is allocated, the higher the accuracy of probabilities will result.

In short, most accurate sentiment analysis would be executedand implemented in the system for sentiment classification purpose. In practice,this system also provides greater accuracy when more accurate result for testing sentimental words.

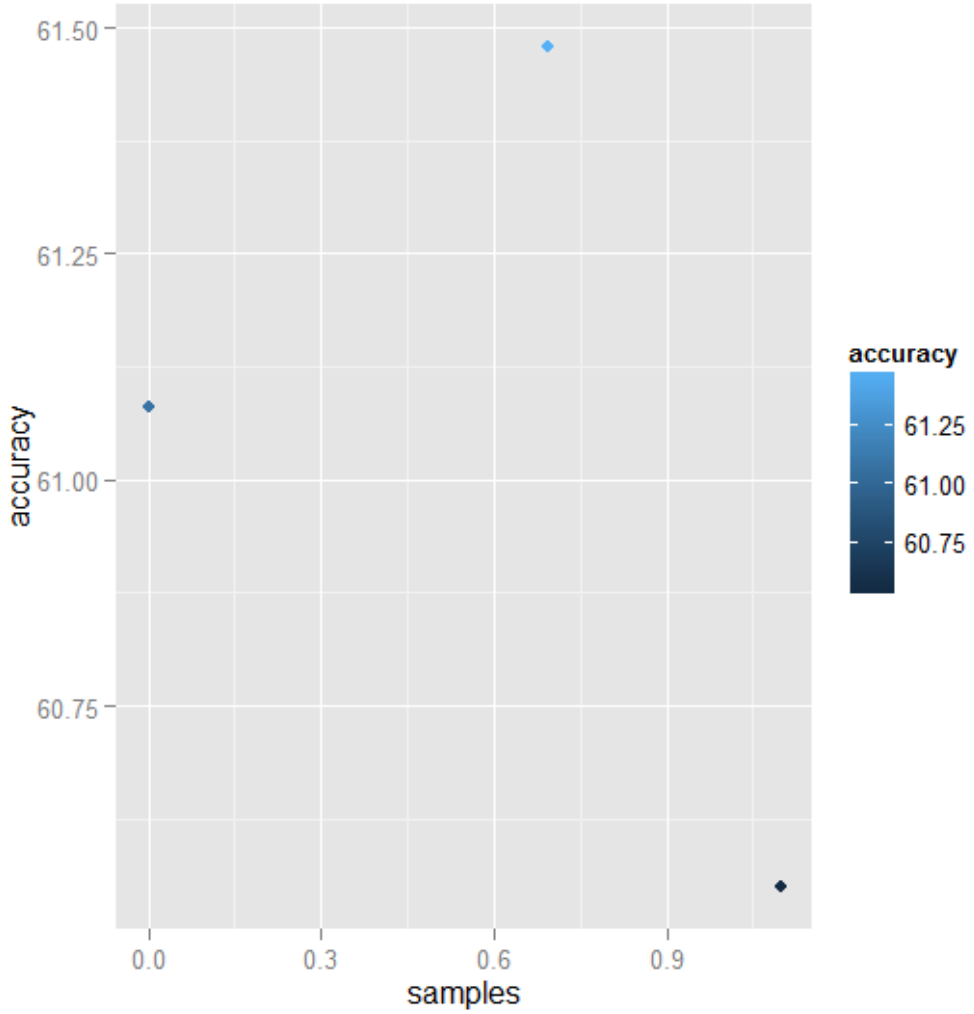
4.4Comparison with sentiment analysis with AFINN wordlist

Many scholars have applied machine learning approach and implemented text mining tool for education purpose and also in industrial field. Opinion mining/sentiment analysis has a long history of being widely applied. There are many researches done, such as natural language processing (NLP), web mining, data mining, text mining, and so on. All carries the same purpose of trying to gather and analyze disparate words of thought and opinions about different topics.

Throughout my research, I found a very meaningful sentiment analysis in AFINN wordlist by Andy (bromberg, 2013). He used R programming language to develop a sentiment analysis using R. He combined all his works with 4 basic areas of expertise- suitable data for analysis, analyze data by using positive and negative polarity corpus, train on data, and also to use train data to classify the same content.

He used AFINN wordlist which contains 2477 words and phrases rated from very negative -5 to very positive +5. He managed to reclassify the wordlist into 4 categories and added in few more words into the AFFIN wordlist. This is how it differs from prevalent manner of which wordlists are normally created by using negative and positive labelled sets.

4.4.1 Performance analysis by using AFINN wordlist system



Sample	Precision	Recall	Accuracy
1000	58.05	83.08	61.08
2000	58.69	80.74	61.48
6000	57.75	80.87	60.55

Sentiment analysis with AFINN wordlist

4.4.2 Performance analysis by using both naïve bayes and support vector linear

Sample	Precision	Recall	Accuracy
1000	87.64	57.35	74.44*Bigram*
2000	88.39	65.30	78.07*Bigram*
6000	83.18	67.85	76.86*Stopwrod*

Best selected result among three withnaïve bayes

Sample	Precision	Recall	Accuracy
1000	78.29	74.26	76.67*Individual word*
2000	78.54	80.60	79.02*Bigram*
6000	79.09	79.29	78.97*Biagram*

Best selected result among three with support vector linear

The above figures have been multiplied by 100 in order to get the percentage of overall result.

According to the confusion matrix above, sentiment analysis in AFINN wordlistsystem does not provide degree of accuracy higher than the one I proposed. The difference becomes more obvious when it is compared tomy proposed method of Naïve Bayes which has been proven to have achieved higher precision level. This means thateverysentence that is identified carries with it approximately a80% to 90% likelihood to be correct. It is because there are less false positive to be found in this class of data. Whereas the AFINN wordlist system achieve higher recall rate compared to my proposed system. It indicates that few false negative to be found in class. Naïve Bayes achieved higher precision among all the other programs, but the recall rate is very low. I also believe that Naïve Bayes is sensitive to sentence noise.

However, in other scenarios, AFFIN wordlist systemmight have higher performance satisfaction level and produce higher test accuracy based on different classification algorithm that is used, or in choosing different wordlist, they can even be measured by using various training data. In other words, different developers may use different approach fortheir system enhancement and maintenance. The improvement of this robust system of analysis is just a matter of time.

4.5Summary Chapter

In this chapter 4, it presents the selected evaluation measurements for calculating precision, recall, and score of the samples. Also, detailed discussionon experimental result gathered from users' opinions by them expressing their thoughts and feelings regarding different topics of conversation such as social issues, politics, economics, movies, merchandises, and etc. Through the experiment which providesimportant insight to the performance analysis. In addition,I made a comparison between sentiment analysis with AFINN wordlistby Andy(bromberg, 2013)and my system .The next discussion is on conclusion of this discussion and some possible future work.

Chapter 5: Conclusion & future work

Introduction

Research Summary

Main contributions

Limitation of research

Future works

5.0Introduction

This chapter describes the contribution dedicated to the proposed research. It presents as anoverall research summary of my findings on sentiment analysis. It will also highlight the research limitations as well assuggestion for future improvement and enhancementfor a more accurate and precise classifying method.

5.1 Research Summary

The incessant growth of social media such as Facebook allows "netizens"around the globe to distribute information without borders, share opinions about current issues, even their attitudes toward life. It is hereby submitted that opinion mining is one of most well-known field of study that has invaluable application value in analysis this world social data.

Social network websites increase in a very fast pace around the world. All try to attract more users to develop their virtual environment social relations and virtual human interaction in the society. Human beings have the tendency to share information about their opinions, state of mind, lifestyle, emotions, and sentiment expression via different social media. Therefore, sentiment analysis is a useful tool to analyze the words used since by expressing one's opinion, views, or feelings, one would inevitably use sentimental words. And this piece of valuable information can be sieved out from various websites further classify them according to positive or negative statement or for other use.

With the help of sentiment analysis,a better and more accurate product analysis can be done in a faster and simpler way. Needless to say, it could be a tool of proper strategies planning. This is because sentiment analysis tools help to make right decision based on user generated content and public opinions. In other words, it is also possible to create a positive impact that can influence the reputation of many politicians. This is because it can be used to predict the outcomes of elections through the words spoken or written by the electorates.

However, in this study, I have been restricted from using other languages. Thus, I only focus on public opinions in English language. Sentiment analysis is a critical task requiring minute preparation. The implicit and explicit opinions expressed by using adjective or adverb sometimes can be very complicated and sophisticated too.

5.2Main Contributions

In my research, I utilised different modeling and text processors, and determined which of the many can generate the best result and has the potential in classifying sentiment orientations. Differentmethods of learning are used to evaluate the performance of system, precision and recall results collect from the balance training data.

I proposed to use Naïve Bayesclassifier to classify positive and negative opined sentences. For the overall process of the study, I used python with built-in natural language toolkit for sentiment classification.

The system is able to correctly classify most of the opinionated sentencesaccording to 1,000 samples given 74.44% of bigram collocation, 72.96% of stopwordswere filtered, and 73.33% were of individual word. For 2,000 samples, result shown was 78.07 % for bigram collocation, 75.42%for stopwords filtered, and 76.56% for individual word. Whereas 6,000 samples given 76.73%for bigram collocation, 76.85%forstopwords filtered, and 76.28% for individual words, which are slightly higher than the others.

5.3Limitation of research

Last step of processing would be the summarization of the result generated by supervised learning method. It is a set of boundaries in the research that specifies in details on the method of the working of machine learning algorithm and potential linear decision boundary. Also, other limitations whichI have encountered are such as some opinion sentence taking the form ofunstructured sentence, spelling mistakes, slang abbreviation, or sarcasms.

5.4Future works

The improvement of machine learning algorithm can provide a more accurate and advanced method in formula calculation. There are many supervised learning algorithm out there ready for use. It is dependent on user preference. Some users might be choosing support vector machine, some goes for Naïve Bayes, or others for maximum entropy, and etc. It seems like there is nosingle

agreed solutions for processing those data. Novel functions should be able to accommodate sentiment analysis in order to fit the user needs.

It cannot be denied that feature selection of indicator words like adjectivesand verbs has its own significance. Many adjectives or verbs phrases contain implicit knowledge of product feature in opinion sentence. For example, the sentence "The smartphone's screen is so small." The sentence expresses a negative opinion on the implicit feature "small". In contrast, for another example like "Thesmartphone screen is small enough to fit into my pocket." The sentence expresses a positive opinion on the implicit feature "small". Every single word can have more than one meaning at different situations.

Another area that could be studied is of entity extraction, entity with different words or phrasesmay be expressed. For example, both "mcd" and "mcds" refer to fast-food chain mcdonald's. The words or phrases should be grouped under the same entity list. Add up a corpora function astranslation of abbreviation can help to improve the accuracy of sentiment analysis. It would be an interesting topic to be discussed in the future.

References

Banitaan, S. a. (2010). A formal study of classification techniques on entity discovery and their application to opinion mining. *Proceedings of the 2nd international workshop on Search and mining user-generated contents* (pp. 29-36). New York, NY, USA: ACM.

Carmen Banea, R. M. (2008). A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (p. 2765). Marrakech, Morocco: European Language Resources Association (ELRA).

Dongjoo Lee, O.-R. J.-g. (2008). Opinion mining of customer feedback data on the web. *Proceedings of the 2nd international conference on Ubiquitous information management and communication.* USA: ACM.

Erik Boiy, P. H.-F. (2007). Automatic Sentiment Analysis in On-line Text. *Proceedings of the 11th International Conference on Electronic Publishing held in Vienna, Austria 13-15 June 2007*, (pp. 349-360).

Hsinchun Chen, D. Z. (2010). Intelligent Systems, IEEE . *AI and Opinion Mining* , 77.

Jindal, N. a. (2008). Opinion spam and analysis. *Proceedings of the international conference on Web search and web data mining* (p. 219). New York, NY, USA: ACM.

Juling Ding, Z. L. (2009). An Opinion-Tree based Flexible Opinion Mining Model. *Web Information Systems and Mining, 2009. WISM 2009. International Conference on* (p. 149). USA: IEEE.

Kim, W. Y. (2009). A method for opinion mining of product reviews using association rules. *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human.* Seoul, Korea: ACM.

Lee, B. P. (2008). *Opinion mining and sentiment analysis.* USA: Now Publishers Inc .

Levene, M. (2010). *An Introduction to Search Engines and Web Navigation.* London: John Wiley & Sons.

Lipika Dey, S. M. (2008). Opinion Mining From Noisy Text Data. *Proceedings of the second workshop on Analytics for noisy unstructured text data* (p. 83). Singapore: ACM.

Liu, B. (2010). Sentiment Analysis and Subjectivity. *To appear in Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau)*, (p. 5).

liu, B. (2006). *Web Data Mining.* New York: Springer Berlin Heidelberg New York.

Ryu, J. S. (2009). Mining opinions from messenger. *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human.* New York, NY, USA: ACM.

Stamper, R. (2008). *Sentiment Mining Of Political Forums.* UMI dissertation publishing.

Xiaowen Ding, B. L. (2008). A Holistic Lexicon-Based Approach to Opinion Mining. *Proceedings of the international conference on Web search and web data mining* (p. 231). Palo Alto, California, USA: ACM.

Bo Pang, L. L. (2002). *Thumbs up? Sentiment Classification using Machine Learning*, 79-86.

Ekman, P. (1992). condition and emotion. *An argument for basic emotions*, 169-200.

Kim, S. M. (2011). *Recognising Emotions and sentiments in text.* sydney: the university if sydney.

Kushal Dave, S. L. (2003). *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews*.

Mike Thelwall, D. W. (2010). *Data Mining Emotion in Social Network Communication: Gender differences in MySpace*, 190-199.

Stopwords. (2001, 3 1). Retrieved 4 4, 2015, from pubmed: http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_170.html

bromberg, a. (2013, 1 5). *Sentiment Analysis in R*. Retrieved 3 25, 2015, from Sentiment Analysis: http://andybromberg.com/sentiment-analysis/

Appendix A

Test with naïve bayes

Sentences	Predicted	Actual
You never fail to amaze me, #	Negative	Positive
birthday admin T birthday admin T :D	Positive	Positive
G Day to my MF nicca since day 1 @WhatABarber...enjoy it bruh!	Positive	Positive
Even if you don't make me , I know one guy that can change my mood instantly. That's why I love him a	Negative	Positive
Lily is young lab who is full of engery and all she wants to do is have FUN!! We are more then to help her	Negative	Positive
#Someday i will be really twetting"Yes @justinbieber follow me it was true we have to #NSN..!	Negative	Positive
@justinbieber please justin, make me ! Follow me! ;D love u!!*--* #35	Positive	Positive
SUMBDY TWEET THE BiRTHDAY SONG TO ME! lOL	Negative	Positive
@DanixJames oh i see. its working-James ya. i havent been this since James became my brother	Positive	Positive
@IRealizedJustinthnx;) I'm glad you're nd of course I can't wait to read the chapter!!	Positive	Positive
@menonannamuffin Yay! Francesca started work with us today! We're very !	Positive	Positive
It makes me to see black ppl coming together standing up for something they believe in. #tsu	Negative	Positive
@miz_missyHee, I love it when you're all . :)	Positive	Positive
@JulieKayJKLD Ensuring that ALL of my customers are customers. #bestday	Positive	Positive
@jeanfranny #realtalkhahaha I tweet my reality. Bahah so dramaa. Ps im so you be tweetin again	Negative	Positive
found another fanboy!! (': @YG_GDragon and belated bday	Positive	Positive
@Mz_ForEver See babe...God works in Mysterious ways :) I'm sooo for you!	Positive	Positive
I love to see people at their best & with their lives with positive minds	Positive	Positive
@justinbieber please justin, make me ! Follow me! ;D love u!!*--* #34	Positive	Positive
Don't let being single be the reason you're struggling to be . A relationship doesn't promise happiness, but God does!	Positive	Positive
We live for tomorrow so wish u all in life & LOVE. ;D	Positive	Positive
@jayysweet that is so true ;) thanks Hun I'm too	Positive	Positive
@Gigi17PW Definitely! I'd be for you! :D	Positive	Positive
@Sbermudez_ awww!! Tell them I said Congrats!!! I'm soooo excited and for them. :D	Postive	Positive
my teacher said i got the highest grade on the test mondayim so ***love dnt change**	Positive	Positive

Sentences	Predicted	Actual
the only thing on my calendar is a party on the 24th. # #iwanttobebusy	Negative	Negative
..i just wish you would stop seeing me as your friend and maybe as more </3..	Negative	Negative
@PowerfulRenata just of all is mad i feel i don somthing wrong	Negative	Negative
My next door neighbor cheats on all her boyfriends, itsreally ..	Negative	Negative
@BiebsMeetMe I get everytime I see this fact :(aww	Negative	Negative
I'm really that the NBA season is over..I miss it already =/	Negative	Negative
They can cure grey hair, but not cancer? How . -.-	Negative	Negative
I was suppose to go out to eat I dnt even have no energy to put clothes on this is a mess	Negative	Negative
face for sure ! :(its times like these you learn to.....	Negative	Negative
Kinda all of my sports are over for now only baseball left... What to do what to do??	Negative	Negative
@Amy_Terry dropped off of youtube too. so .	Positive	Negative
Dontwanna hear your songs i dontwanna feel your pain http://myloc.me/kSw0N	Negative	Negative
to admit ... that beef was the highlight of my day .. I have no life I swear lol	Negative	Negative
No work 2moro, home by 1, n no motive for 2moro, it's to say, but the shift I'm always dreading. I wish I was going to #neversatisfied	Positive	Negative
* The movie was . Two lazy bumms are sleeping. Sigh.	Negative	Negative
So Betty Fox has passed away. #fb	Positive	Negative
I c @All_Ey3z_On_M3 yesterday n he ainttel me hi...smh	Positive	Negative
At PDX, about to say bye to David nd Rachel. face =(	Negative	Negative
i think id get and depressed and feel like a loner and than just sleep.	Negative	Negative
"Cougardating" emails will never not make me .	Negative	Negative
Lots of looking old KISS fans at #ribsontheriver, wait... am I one of those??	Positive	Negative
Last night in Ibiza :(the thought of packing makes me .	Negative	Negative
okay, im :(but i'll be fine.	Negative	Negative
So, a bitche that beat up a bitch with red hair wanna talk shit about OF #	Negative	Negative
its that i get 10 miles to the gallonnn #expensive	Negative	Negative

Test with support vector linear

Sentences	Predicted	Actual
You never fail to amaze me, #	Negative	Positive
birthday admin T birthday admin T :D	Positive	Positive
G Day to my MF nicca since day 1 @WhatABarber...enjoy it bruh!	Positive	Positive
Even if you don't make me , I know one guy that can change my mood instantly. That's why I love him a	Positive	Positive
Lily is young lab who is full of engery and all she wants to do is have FUN!! We are more then to help her	Negative	Positive
#Someday i will be really twetting"Yes @justinbieber follow me it was true we have to #NSN..!	Negative	Positive
@justinbieber please justin, make me ! Follow me! ;D love u!!*--* #35	Positive	Positive
SUMBDY TWEET THE BiRTHDAY SONG TO ME! lOL	Negative	Positive
@DanixJames oh i see. its working-James ya. i havent been this since James became my brother	Positive	Positive
@IRealizedJustinthnx;) I'm glad you're nd of course I can't wait to read the chapter!!	Positive	Positive
@menonannamuffin Yay! Francesca started work with us today! We're very !	Positive	Positive
It makes me to see black ppl coming together standing up for something they believe in. #tsu	Positive	Positive
@miz_missyHee, I love it when you're all . :)	Positive	Positive
@JulieKayJKLD Ensuring that ALL of my customers are customers. #bestday	Negative	Positive
@jeanfranny #realtalkhahaha I tweet my reality. Bahah so dramaa. Ps im so you be tweetin again	Negative	Positive
found another fanboy!! (': @YG_GDragon and belated bday	Positive	Positive
@Mz_ForEver See babe...God works in Mysterious ways :) I'm sooo for you!	Positive	Positive
I love to see people at their best & with their lives with positive minds	Positive	Positive
@justinbieber please justin, make me ! Follow me! ;D love u!!*--* #34	Positive	Positive
Don't let being single be the reason you're struggling to be . A relationship doesn't promise happiness, but God does!	Positive	Positive
We live for tomorrow so wish u all in life & LOVE. ;D	Positive	Positive
@jayysweet that is so true ;) thanks Hun I'm too	Positive	Positive
@Gigi17PW Definitely! I'd be for you! :D	Positive	Positive
@Sbermudez_ awww!! Tell them I said Congrats!!! I'm soooo excited and for them. :D	Positive	Positive
my teacher said i got the highest grade on the test mondayim so ***love dnt change**	Positive	Positive

Sentences	Predicted	Actual
the only thing on my calendar is a party on the 24th. # #iwanttobebusy	Negative	Negative
..i just wish you would stop seeing me as your friend and maybe as more </3..	Negative	Negative
@PowerfulRenata just of all is mad i feel i don somthing wrong	Negative	Negative
My next door neighbor cheats on all her boyfriends, itsreally ..	Negative	Negative
@BiebsMeetMe I get everytime I see this fact :(aww	Negative	Negative
I'm really that the NBA season is over..I miss it already =/	Negative	Negative
They can cure grey hair, but not cancer? How . -.-	Positive	Negative
I was suppose to go out to eat I dnt even have no energy to put clothes on this is a mess	Positive	Negative
face for sure ! :(its times like these you learn to.....	Negative	Negative
Kinda all of my sports are over for now only baseball left... What to do what to do??	Negative	Negative
@Amy_Terry dropped off of youtube too. so .	Positive	Negative
Dontwanna hear your songs i dontwanna feel your pain http://myloc.me/kSw0N	Negative	Negative
to admit ... that beef was the highlight of my day .. I have no life I swear lol	Negative	Negative
No work 2moro, home by 1, n no motive for 2moro, it's to say, but the shift I'm always dreading. I wish I was going to #neversatisfied	Positive	Negative
* The movie was . Two lazy bumms are sleeping. Sigh.	Negative	Negative
So Betty Fox has passed away. #fb	Positive	Negative
I c @All_Ey3z_On_M3 yesterday n he ainttel me hi...smh	Positive	Negative
At PDX, about to say bye to David nd Rachel. face =(	Negative	Negative
i think id get and depressed and feel like a loner and than just sleep.	Negative	Negative
"Cougardating" emails will never not make me .	Negative	Negative
Lots of looking old KISS fans at #ribsontheriver, wait... am I one of those??	Positive	Negative
Last night in Ibiza :(the thought of packing makes me .	Negative	Negative
okay, im :(but i'll be fine.	Negative	Negative
So, a bitche that beat up a bitch with red hair wanna talk shit about OF #	Negative	Negative
its that i get 10 miles to the gallonnn #expensive	Negative	Negative