2017

The Evaluation of Ensemble Sentiment Classification Approach on Airline Services Using Twitter

Zechen Wang
Dublin Institute of Technology

Follow this and additional works at: http://arrow.dit.ie/scschcomdis

Part of the Computer Engineering Commons

Recommended Citation

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License
The evaluation of Ensemble sentiment Classification approach on airline services using Twitter.

Zechen Wang

A dissertation submitted in partial fulfillment of the requirements of Dublin Institute of Technology for the degree of M.Sc. in Computing (Advanced Software Computing)

2017
I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Advanced Software Computing), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute’s guidelines for ethics in research.

Signed: Zechen Wang

Date: 03, January, 2017
ABSTRACT

In the field of sentiment classification, much research has been done on reviews of topics such as movies, software and books. Little research has been done in the airline service domain. In the airline industry, the use of social media as a customer service tool has become a growing phenomenon. The research conducted by Wan and Gao (2015) has proposed an ensemble classification approach for airline service sentiment classification using Twitter data. In accordance, the objective of improving the performance of ensemble classification approach is the primary consideration.

This research proposed new hybrid classification approach that uses the state-of-art approach proposed by Wan and Gao (2015) combining with lexicon based approach on classification of airline service topic using Twitter data. The research evaluated the proposed approach in depth, along with explorations of implementing expansion of tweet content in order to further improve the classification performance.

In this project, the ensemble approach that consists of both machine learning approaches and lexicon based approach was analysed which suggested the improvement of the proposed classification approach performance compare with machine learning only approach on airline service domain conducted by Wan and Gao (2015).

Key words: Natural Language Processing, Twitter, Airline service, Sentiment classification, supervised machine learning,
ACKNOWLEDGEMENTS

This research project cannot be completed without the help of my supervisor Dr. John Gilligan. I would like to express my sincere gratitude for his patience, advice, guidance and consistent support for enabling me to complete my research project.

I also wish to thank our research dissertation coordinator Dr. Luca Longo for providing assistance on developing research skills and advice for selecting the research topic.
TABLE OF CONTENT

ABSTRACT ............................................................................................................. I

ACKNOWLEDGEMENTS .................................................................................. II

TABLE OF FIGURES ........................................................................................... VII

TABLE OF TABLES ............................................................................................. VIII

1  INTRODUCTION ............................................................................................... 1

1.1 ABOUT THIS RESEARCH ........................................................................... 1

1.2 BACKGROUND ............................................................................................. 1

1.2.1 Social Media ........................................................................................... 1

1.2.2 Opinion Mining ....................................................................................... 2

1.3 RESEARCH PROBLEM ............................................................................... 3

1.4 RESEARCH OBJECTIVES ........................................................................... 5

1.5 RESEARCH METHODOLOGY ..................................................................... 6

1.6 SCOPE AND LIMITATIONS ....................................................................... 7

1.7 DOCUMENT OUTLINE ............................................................................... 7

2  LITERATURE REVIEW ..................................................................................... 9

2.1 INTRODUCTION .......................................................................................... 9

2.2 SENTIMENT CLASSIFICATION .................................................................... 9

2.2.1 Overview .................................................................................................. 9

2.2.2 Applications of Sentiment Analysis ......................................................... 10

2.2.3 Sentiment Polarity Classification .............................................................. 10

2.2.4 Subjectivity and Opinion Bearing Detection ............................................ 10

2.2.5 Natural language ..................................................................................... 11

2.2.6 NLP and sentiment analysis ................................................................... 12

2.2.7 Text representation in sentiment analysis .............................................. 14

2.2.8 Sentiment Analysis process ................................................................... 15

2.3 SENTIMENT CLASSIFICATION ALGORITHMS AND APPROACHES .......... 17

2.3.1 Lexicon-Based Approach ................................................................. 18

2.3.1.1 SentiWordNet ..................................................................................... 18

2.3.2 Supervised Machine learning approaches .......................................... 20

2.3.2.1 Support Vector machine (SVM) ....................................................... 22
2.3.2.2 Naïve Bayes .................................................................23
2.3.2.3 Bayesian network..........................................................24
2.3.2.4 Decision tree...............................................................24
2.3.3 Unsupervised Machine Learning........................................25
2.3.4 Sentiment Classification Evaluation..................................26
2.4 Sentiment analysis and airline service ....................................28
  2.4.1 Sentiment analysis on airline service using Twitter Data .......28
  2.4.2 Existing Ensemble approach in airline service ....................29
2.5 Conclusion............................................................................30

3 DESIGN AND METHODOLOGY .............................................32
  3.1 Introduction ........................................................................32
  3.2 Data understanding and opinion mining ................................32
    3.2.1 Data source ..................................................................32
    3.2.2 Dataset analysis and considerations ...............................33
  3.3 Comparison factors ............................................................35
    3.3.1 Classification accuracy ................................................35
    3.3.2 Training dataset ..........................................................36
  3.4 Experiment design ..............................................................36
    3.4.1 Data preparation ..........................................................36
    3.4.2 Baseline classifier .........................................................37
    3.4.3 Ensemble Classifier with Lexicon ..................................37
    3.4.4 Implementation and design .........................................38
    3.4.5 Evaluation ....................................................................39
    3.4.6 Conclusion ....................................................................40

4 EXPERIMENT IMPLEMENTATION .........................................41
  4.1 Introduction ..........................................................................41
  4.2 Experiment process .............................................................41
    4.2.1 Data pre-processing ......................................................42
    4.2.2 Parameter settings ........................................................43
    4.2.3 Data sampling .............................................................43
    4.2.4 Feature selections ........................................................45
    4.2.5 Training set selection ....................................................45
8.2.1 Word Matrix settings ......................................................................................... 76
8.3 Twitter slang word list ..................................................................................... 76
# TABLE OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Information Gain and n-gram level</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>Sentiment classification techniques (Medhat, Hassan, &amp; Korashy, 2014)</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>Graphical representation of a term sense adopted by SentiWordNet (Esuli &amp; Sebastiani, 2006)</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>Graphical representation of a term ‘Estimable’ (Esuli &amp; Sebastiani, 2006)</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>Example of classification in two dimensional space. (Cortes &amp; Vapnik, 1995)</td>
<td>23</td>
</tr>
<tr>
<td>6</td>
<td>Formal expression of Naive Bayes model (Kelleher, Namee, &amp; D’Arcy, 2015)</td>
<td>24</td>
</tr>
<tr>
<td>7</td>
<td>Visualised decision tree sample</td>
<td>25</td>
</tr>
<tr>
<td>8</td>
<td>Experiment process flow</td>
<td>41</td>
</tr>
<tr>
<td>9</td>
<td>Classification performance result with different dataset used</td>
<td>55</td>
</tr>
</tbody>
</table>
TABLE OF TABLES

Table 1 Word matrix example ........................................................................................................... 14
Table 2 (Vohra & Teraiya, 2013) .................................................................................................. 21
Table 3 Multi-class Confusion matrix example ............................................................................... 27
Table 4 Number of tweets distribution for each class ..................................................................... 33
Table 5 Baseline classification result ............................................................................................... 44
Table 6 Accuracy criteria comparison ............................................................................................... 44
Table 7 Performance comparison for training size ............................................................................ 45
Table 8 Performance comparison of combination algorithms .......................................................... 46
Table 9 Environment configuration .................................................................................................. 47
Table 10 Performance result of baseline classifier ........................................................................... 51
Table 11 Performance result of proposed ensemble classifier with dataset A ......................... 51
Table 12 Result for using Dataset B ................................................................................................ 52
Table 13 Accuracy result for B5 ........................................................................................................ 53
Table 14 Result for using Dataset C ................................................................................................ 53
1 INTRODUCTION

1.1 About this research
This project investigates sentiment analysis of airline social media content. Airlines use opinions expressed in this content to improve services, react to customer issues and plan future marketing campaigns. This research also investigates the state-of-art approaches in the field of sentiment classification in regards to airline service domain opinion mining by employing opinion rich platform, Twitter.
This dissertation paper aims to improve the accuracy of sentiment classification approach in the airline service domain. Until recent years only little research has been done for sentiment analysis in airline service with the popularity of social media resources. Based on the research paper conducted in Wan et al (2015), the paper not only contribute a new hybrid classification approach in the field of opinion mining, but also in the airline service domain which intensively discussed the challenges and demand of opinion mining in this domain. Furthermore, the potential for creating and discovering new opportunities in field of sentiment analysis when dealing with various domain specific sentiment classification applications are the motivation of this project.

1.2 Background

1.2.1 Social Media
The development of information communication technologies (ICT) has significantly impacted the business world, creating new business opportunities across various industries. Many companies from different industries have already adopted the new business model that combining with the information communication technologies since the mid-90s when the Internet started to become progressively popular (McIvor, O’Reilly, & Ponsonby, 2003).
Nowadays, in the Internet age, mobile devices or personal computers are often used for paying electricity bills, rent and transferring money without stepping out of the door. Online users can buy quality goods that were used to purchase from the stores, and even large furniture. From the business owner point of view, the Internet has changed the way of the traditional business operations and social life remarkably. The external use of the
Internet also forged a competitive weapon and swiftly accelerated the competitions among global commerce, transportation and financial industries in particular (Ives & Learmonth, 1984).

The rapid increase usage of the internet and web technology has boosted the economy in airline product sales and profitability. The internet has become the central tool for their business strategy development of the airline industry sector. There is a growing impact on airline companies’ revenue with their business strategy advantages involved in the applied internet technologies (McIvor, O’Reilly, & Ponsonby, 2003). The application of internet technologies are not only beneficial for creating business values, but also enables organisations to maintain their competitive position by fully utilising the potential of internet resources. Implementation of social media in marketing has been defined as a ‘connection between brand and consumers.’ (Paquette, 2013). The adoption of social media technologies has already become the driving force for establishing and reinforcing their competitiveness among large number of companies across numerous industries and domains. The potential of using social media as a marketing tool has been recognised by the vast majority of the organisations and apply the tool into their business strategies, to maintain and improve competitiveness in the market. For instance, valuable information can be gathered using a marketing survey conducted by online users, customer feedbacks and reviews written by experienced consumers. The survey then can be analysed in order to provide knowledge on selling products and support for business decision makings. From companies’ perspective, it can be extremely valuable for their future business strategies for determining the customer’s post purchase responses and reviews. Interacting face-to-face with individual customer could be very time consuming and challenging when attempts to determine customers’ opinions on the products sold. Thus, the use of social media tools can be leveraged for essential business strategies, such as interacting with customers and monitoring targeted online customer groups (Bonchi, Castillo, Gionisi, & Jaimes, 2011).

1.2.2 Opinion Mining

To analyse online user activities and understand the post purchase review or comment of the purchased product, many approaches were proposed to achieve this goal. One of the popular approach is the sentiment analysis on social media content.
Not only in business applications, social media sentiment analysis has already attracted many researchers to analysis and study many years ago in sociology (Scott, 2000), the deployment of social networking sites and collections of rich-opinion social media data have significant business impact (Bonchi, Castillo, Gionisi, & Jaimes, 2011). With the growing popularity of the social networks and the opinion resources generated by them, sentiment analysis has provided the opportunities to explore online users’ behaviours (Pang & Lee, 2008).

Sentiment classification also known as opinion mining, which referring to the work carried out by applying computational treatment to analysis the underlying opinion, subjectivity and the sentiment for a given piece of text. The text can be long text in documents or reviews and can be short sentences in micro blogs, news highlight or posts (Pang & Lee, 2008). In other words, the purpose of implementing computational treatment to analyse sentiment of the text is to attempt automating the text classification and opinion recognition procedure for processing large documents and text data (Morinaga, Yamanishi, Tateishi, & Fukushima, 2002). Much extensive research has been done throughout the decades that focus on opinion mining of text. The opinion mining or sentiment analysis approaches can all fall into the category of Natural Language Processing (NLP) techniques. The term opinion extraction has also been used to in this context to compute and predict the sentiment of the text (Yi, Nasukawa, Bunescu, & Niblack, 2003). Essentially, the basic concept of the much research (Yi, Nasukawa, Bunescu, & Niblack, 2003, Pang, Lee, & Vaithyanathan, 2002, Melville, Gryc, & Lawrence, 2009, Wan & Gao, 2015) was proposed which focus on detecting the polarity of text, positive and negative.

1.3 Research problem

The use of social media platform in the airline industry has been rapidly increased to analyse the quality and performance of the services provided by airline companies, due to the effectiveness and inexpensive features of using social media (Paquette, 2013). There are many related research efforts have been conducted that employed Twitter data and performed a series of sentiment classification techniques to evaluate the general public opinion on airline services (Breen, 2012). However, there is need to evaluate and improve the accuracy of the result in sentiment analysis, since it is not yet able to compare with the human judgement on determining the correct sentiment the text should
fall into. Nevertheless, the improvement of the accuracy of sentiment classification approaches are still not yet discovered.

Based on the evidence from many empirical studies, the most accurate text polarity classification method is the ensemble approach, where many individual classifiers are working together with a combination algorithm to predict the polarity of a text. For lexicon based approach, the experiment conducted by Augustyniak et al., (2014), has proven that the accuracy of implementation of ensemble lexicon based classifier on the same dataset outperforms the other lexicon based approaches, achieved a higher F-measure on book and electronic reviews in comparison to supervised machine learning approach. The proposed method is to construct a vector of total count of positive words and negative words appeared in the document, then determine the final result of the text sentiment polarity. This paper uses the sum rule to establish the final result based on the prediction made from each corpus. The state-of-art ensemble machine learning approach in sentiment classification on airline service domain proposed by Wan et al (2015) has achieved 91.7% of overall accuracy that outperformed the other individual machine learning classifiers. The machine learning classifiers used in this paper include Naïve Bayesian method, Bayesian Network method, Support Vector Machine (SVM) method, C4.5 Decision Tree method and Random Forest method. This proposed combination of the supervised machine learning approach employed the Majority vote rules as combination algorithm to produce the final prediction of the text.

As what has been done in other research, ensuring an appropriate degree of accuracy of the sentiment analysis result should be a prior task to complete, in order to meet a high degree of a quality sentiment analysis result and produce valuable informative analysis. However, the incorporation of both supervised machine learning and lexicon based approaches has never been evaluated.

In real world application, a superior sentiment classifier can be implemented to evaluate tweet posted across different types of domains, such as online product reviews (Vidya, Fanany, & Budi, 2015), attraction reviews and airline service reviews. However, it is still required that each time the target domain and objective is changed, the application needs to reconfigured, so that the selected approach can provide an optimal sentiment analysis results for a specific domain. Identifying the most appropriate approach which can be beneficial for improving the performance of sentiment classification process.
WILL THE ENSEMBLE SENTIMENT ANALYSIS PRODUCES MORE ACCURATE SENTIMENT ANALYSIS, INCORPORATING WITH LEXICON-BASED APPROACH BY EMPLOYING LABELED TWITTER DATA ON AIRLINE SERVICE?

1.4 Research Objectives

The core objective of this research is to evaluate the sentiment classification approach in relation to implementing the Lexicon based classification approach incorporate with five state-of-art supervised machine learning classification approaches on polarity mining specifically for airline service domain. According to the discussion made in the previous chapter, the research objectives are:

- Analyse and discuss the related topic in the field of sentiment analysis, Natural Language Processing techniques and lexicon resource creating and implementation approaches.

- Review state-of-art sentiment classification approaches. Investigate the advantages and disadvantages of these approaches.

- Investigate the existing text mining techniques in the field of sentiment analysis.

- Design and implement approaches reviewed from previous objectives.

- Evaluating the classification result obtained from baseline classifier using various measurement techniques investigated.

- Design and construct proposed classification method using equivalent configurations implemented in baseline classification approach.

- Evaluation and analysis results obtained from baseline classifier and proposed classification strategy.

- Critical investigation on obtained results of classification and error of misclassification for the proposed ensemble classification approaches.
Identify and analysis the improvement or demotion of the new analysis strategy comparing with existing approaches on Twitter regrading to airline service.

The objective of this research is to explore the new sentiment analysis approach based on the existing state-of-art sentiment analysis approach on airline service domain. It can be achieved by evolution of the accuracy of the classification results obtained by the proposed method. The accuracy evaluation techniques used for the two sentiment analysis approaches, ensemble sentiment analysis approach and newly proposed ensemble approach with lexicon based analysis, includes recall, precision and f-measure. The research hypothesizes for this paper is shown as follows.

H0: The existing machine learning hybrid analysis approach outperforms the proposed hybrid sentiment classification approach that includes lexicon based approach.

H1: The existing ensemble sentiment analysis can still produce relative high accurate sentiment classification results in the airline service domain.

H2: The accuracy of the proposed ensemble sentiment analysis outperforms the existing ensemble sentiment analysis approach.

1.5 Research Methodology

As part of this thesis, the primary research and secondary research were carried out. The primary research was conducted based on the experiment on tweets polarity classification specifically on airline service domain using proposed ensemble classification approach. The methodologies used are based on the secondary research conducted in the field of opinion mining and machine learning. The primary research was conducted based on three core processes. The experiment was implemented in an iterative fashion which pre-processing were tuned based on each previous result. The processes are:

1. The data pre-processing techniques are applied to the collected dataset.
2. Sentiment classification prediction using classifiers with pre-processed dataset stated above.
3. Sentiment classification results evaluation.

Secondary research is a knowledge gathering process that consisted of review of literature in the field of sentiment classification, text mining, and machine learning. Includes:

1. Conference papers, published literatures, and research journals published on The Arrow DIT, IEEE, Springer, etc.
2. Published books in the field of machine learning
3. Technical documentation website (Java API Documentation, researchGate, Weka API)

1.6 Scope and Limitations

As part of this research, the experiment was concentrated on classifying the polarity of each airline service domain related tweet. The proposed approach employed supervised machine learning algorithms and lexicon based classification approach.

There are several barriers in this research that limit the results and can be further improved with evaluation made based on the result of this paper. First of all, the classification process using a single dataset collected may not be able to represent the overall improvement of the proposed sentiment classification approach across multiple domains. Because of the fact that this paper only focuses on providing insight of sentiment analysis for airline service domains. Furthermore, there are concerns regarding the technical resources can be employed in this research, as sometimes, sentiment classification tasks requires powerful computational resources from a hardware perspective.

1.7 Document Outline

This dissertation consists of seven chapters that are organised as follows.

Chapter 2 discusses the secondary research is conducted based on the objectives of this research. The chapter reviewed the real world applications using sentiment classification
and provided an overview of cross domains sentiment classification. The state-of-art approaches in the field of data mining and sentiment classification are also reviewed in this chapter, along with the commonly used measurement of performance of the approach.

Chapter 3 introduces the design of the experiment for this project, in order to achieve the objectives stated early in this chapter. The considerations of the experiment are also discussed in depth regarding to how and why these procedures should be implemented.

Chapter 4 provides the implementation details based on the experiment processes designed in Chapter 3. The actual implementation of the experiment is then presented with the challenges and limitation identified during the process.

Chapter 5 presents result obtained from the implementation of experiment discussed in Chapter 4. The detailed evaluation of the results are also presented in this chapter to reflect the objectives on this project.

Chapter 6 concludes the experiment implemented and result obtained. The review of the key objective of this research is presented. Furthermore, the contribution and future research directions are discussed in this chapter, based on the overall review of this project.
2 LITERATURE REVIEW

2.1 Introduction

As the objective of this research stated in the previous chapter, examining the sentiment classification approaches on airline service domain using social media content. In this chapter, research literature in the field of sentiment analysis, sentiment classification processes, state of art sentiment classification approaches and application of sentiment classification in airline service domain are presented and discussed. The discussion focused on the current state-of-art domain specific sentiment analysis approaches, using content extracted from social media tools.

2.2 Sentiment Classification

2.2.1 Overview

Sentiment classification can be described as a process that extracts the underlying sentiment of the provided document or text. Sentiment classification is also described in the field of opinion mining and natural language processing. According to Pang and Lee, (2008) the phrase sentiment analysis or option mining is first introduced in the field of marketing research, where authors, Das and Chen, who are interested in finding the market sentiment through web technologies (Pang & Lee, 2008). Then similar discussion and papers were published in the field of Natural Language Processing and Association for Computational Linguistics by Turney (2011) and Pang, Lee, & Vaithyanathan, (2002). The definition of opinion mining is also described , as a recent sub-discipline at the crossroad of information retrieval and computational linguistics which concerned with what the topic of a document is about (Esuli & Sebastiani, 2006). In the early stages of sentiment classification research, the phrase text categorisation is commonly used in the field of study. Text categorisation is the classification of documents into predefined categories (Joachims, 1998). For instance, book genres can be categorised into action, novel etc.
2.2.2 Applications of Sentiment Analysis

The applications of text classification were used across multiple domains, nowadays, such as business, communications and political domains. For example, customer product review websites, message or email spam filters and detection system, search engines, article categorisation, language identification, political debate prediction and recommendation systems. The implementation of email spam filter system employs not only DNS blacklist, but also analyses the content of the email to ascertain whether the email is spam or not. Many of the decision techniques are used in the process, such as analysing the writing style of the email, the regular spam words appeared in the email using suitable corpus built and machine learning techniques such as SVM, Decision tree and Bayesian classifiers (Trivedi, 2016). Search engines are using similar approaches that classify the search queries into different categories. For instance, search keyword ‘apple’, can be divided into categories such as computers or fruit. The implementation of SVM classifier is used in a related field which classifies the result of the search query (Liu, Li, & Lin, 2015).

2.2.3 Sentiment Polarity Classification

The polarity classification is one of the category in sentiment classification. The primary purpose of the polarity classification is to identify if the target document is good or bad, for example in email spam filter, whether it is spam or not and in product reviews, whether the review leads to positive sentiment or negative sentiment. This type of sentiment classification is called sentiment polarity classification or polarity classification (Pang & Lee, 2008). This polarity classification is also known as binary classification, because there are only two types of classification results. As for classifications like categorising search engine query results into multiple predefined categories, it is often referred to multi-class categorisations.

2.2.4 Subjectivity and Opinion Bearing Detection

However, the process of sentiment classification assumed that the opinions already exist in the text, when classifying the underlying sentiment for the specific piece of text. It is necessary to identify the subjectivity of the document and objective evidence (Yu & Hatzivassiloglou, 2003). There are cases where provided document does not contain any opinions or polarities. The given documents are objective and only explain the fact of
some matters. For instance, most articles from newspapers are stating the facts that already happened in the past, meaning that there are no indicators of bias behind the facts described. If the documents are not bias, then the opinion of the document cannot be correctly extracted. Identifying the subjectivity within the given text is essential in sentiment analysis process. Many published research papers have suggested in relation to separate the subjective information and opinions of the documents. According to Wilson et al (2004), determine the strength of subjectivities within deep clause-level of each text can improve the accuracy significantly over baseline (Wilson, Wiebe, & Hwa, 2004). To further extend this concept, the ranking of the sentiment polarity is also interested in research (Wilson, Wiebe, & Hwa, 2004) (Riloff & Wiebe, 2003) (Mehto & Indras, 2016). It is defined as rating inference in Pang and Lee (2008), where the degree of positivity or negativity is classified, such as a positive sentiment is consist of strong positive, medium positive and weak positive. The same rules can also be applied to a neutral document that one scenario can be described as ‘this text is classified as strongly lack of opinion (neutral)’.

Another desired outcome of extracting the opinion of the example review is to summarise why this review is expressed in certain bias. In other words, classifying the cause of the bad reviews and good reviews. Kim and Hovy, (2006) conducted research that proposed a novel approach to automatically identify the cause of the overall sentiment polarity using a collection of marker words appeared in a sentence. However, it is identified by the author that the designed approach is complex and the experiment implemented was only tested with small data size.

2.2.5 Natural language

Sentiment classification techniques and applications are developed due to many reasons. One of the main reasons is because of the rapid increase of opinion text data generated in social media, and potential of classifying these text data can be significantly valuable from many aspects. In this research, there is need to discuss the causes of the formulation of these classification techniques, that is, opinion rich text data itself.

In the field of data mining, the analysis of the dataset is interested in extracting information from structured data. However, in real world situations, due to the complexity and ambiguities of human natural language, textual data that used for opinion mining task is unstructured, which requires deeper analysis and pre-processing
before classifiers can classify the underlying sentiment (Rajman & Besançon, 1998). The textual data do not follow statistical rules, but grammar and syntax rules defined in different languages accordingly. When dealing with unstructured textual data, there are many challenges and issues (Stavrianou, Andritsos, & Nicoloyannis, 2007). In text mining, the main issues can be identified as follows:

1. Stop words considerations
2. Stemming words considerations
3. Existence of noisy data in text.
4. Word sense disambiguation
5. Part of speech tagging
6. Compound or technical terms
7. Tokenisation considerations
8. Word order, context and background knowledge

The above list stated considerations that needed to be addressed when processing textual data for sentiment classifications. Where addressing these natural language issues potentially increases the performance of sentiment classification results. It was proven to be an effective approach in many research papers (Joachims, 1998), (Gaikwad, Chaugule & Patil, 2014), (Kummer & Savoy, 2012). In the field of sentiment classification, much research adopted Natural Language Processing techniques to address the issues in the textual data. The application of NLP in sentiment classification is reviewed in the next section.

2.2.6 NLP and sentiment analysis

One of the concerns discussed in the previous section is the consideration of *stop words* in the textual dataset. For sentiment classifications, stop words such as ‘a’, ‘the’, ‘be’ etc. in content of sentiment classification have less impact on the final result of the sentiment. However, sometimes stop words can be relevant to the content of the document and affect the sentiment classification result depending on the objective. For example, when sentiment strength is considered as one of the classification results, the word ‘very’ should not be removed in the text, as it can be treated as a solid indicator of the sentiment polarity strength.

The stemming technique in NLP has provided a way of reducing the variation of the terms used in the text. For instance, the word ‘cancelled’, ‘canceled’, are all replaced...
with ‘cancel’. The purpose of implementing the procedure in the NLP is tantamount to decrease the number of attributes used during the process. However, it still depends on the objective of the text mining task. Because this process also reduces the information provided for the classifier when training and classifying text.

The noisy data in the text are referring to the incorrectly spelled words, shortened terms such as abbreviations and acronyms, and mark-up language tags. These noisy normally exists in unprocessed data which requires correction of misspelled words and expansion of the abbreviations and acronyms (Stavrianou, Andritsos, & Nicoloyannis, 2007).

The word sense disambiguation (WSD) refers to the same term used in a different context, that expressing completely different meanings. One typical example, can be the term ‘touch down’, it refers to scoring a goal in sports and it could also mean the airplane landing on destinations. The use of this term in a different context could be resulted in distinct meanings. The word sense disambiguation has been an active research topic in recent decades and challenging task to complete with ideal results. Many approaches have developed over the years and they all follow the rules of using external lexical resources to determine the actual meaning of the term, but applying different methods from supervised machine learning and semi-supervised machine learning approach to unsupervised machine learning approach (Pal & Saha, 2015). The WSD in text mining tasks requires special considerations, because of the domain dependencies of different term used.

The part of speech tagging is to determine the grammatical class of the term in the text. In the field of natural language processing, part of speech tagging is typically used in text mining, because the same term can have completely different senses in different grammatical class. For instance, word ‘bass’ is referring to the lowest singing voice in musical domains in adjective, but referring to a kind of freshwater perch in noun. In sentiment classification, tagging the terms can influence the overall sentiment, since it has different sentiment scores in the corpus, such as SentiWordNet (Esuli & Sebastiani, 2006). The technical terms used in sentiment classifications are considered accordingly, the reason being is that the impact of using technical terms is small when only classifying the overall polarity of the text, because they are not always indicators of the emotions in the text.

The considerations of tokenisation in text are referring to the tasks of converting the textual data into different collections of words or terms as tokens, then being used for text mining, which increase the overall performance of the sentiment classification
process and provide an easier manipulation process for text mining (Kummer & Savoy, 2012). Sometimes in long documentation classification process, the tokenisation process can separate the whole document into different segments that can be considered and analysed individually depending on the objective of the data mining task. The tokenisation of the text can be represented in different formats.

2.2.7 Text representation in sentiment analysis

The most common text representation used for sentiment classification is the word vector model proposed in (Salton, Wong & Yang, 1975). The implementation of converting textual data into word vector has created the opportunity for applying statistical analysis in textual data. The example of word vector text representation is illustrated as follows:

<table>
<thead>
<tr>
<th>ID</th>
<th>thanks</th>
<th>cancel</th>
<th>flight</th>
<th>delay</th>
<th>book</th>
<th>.....</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.....</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>.....</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.....</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>.....</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>.....</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>.....</td>
</tr>
</tbody>
</table>

Table 1 Word matrix example

The above example provided an overview of word vector, for each text, the appearance of each word is represented in binary format where 1 indicates the word appeared in the text and 0 is not. The more detailed word matrix is provided in Appendix section 8.2. N-grams is a commonly used as a text representation of the tokenisation in feature selection technique during sentiment classification process. It refers to breaking down a piece of text into different segments where n indicates the number of words kept in one segment. For instance, unigram keeps only one word from each sentence that sometimes may or may not be able to catch the significant emotion indicators in text. The word vector shown in table 1 can represent the unigrams after the tokenization process. The bigrams can sometimes intercept the negations within the given text, as it takes two words as one unit for considerations. For example, ‘I am not happy with the flight’, unigram usually will take ‘not’ and ‘happy’ into consideration separately where bigrams can capture the term ‘not happy’ that satisfies the original sentiment orientation of
negative expression. The trigrams are taking three words into account for one attribute, four-grams are taking four words for one unit, and so on. The pattern of n-grams is to adopt n number of words from the text as one unit for classification considerations. Consideration of n-gram tokenisation should be taken care of with caution, although, the higher n may identify the missing information in a wider context within a text, it also decreases the level of details within a text. A systematic experiment performed out in (Cheng, Yan, Han, & Hsu, 2007) has concluded the information gain ratio when the number n increased.

![Information Gain and n-gram level](image)

**Figure 1 Information Gain and n-gram level**

Figure 1 shows that the decrease of information gain when the increasing number of multi-grams are considered. However, it also provided insight and benefit of using bigrams and trigrams.

Considering the textual dataset used for the purpose of sentiment classification, the collected tweet dataset is transformed into the word vector matrix as a text representation of each tweet.

2.2.8 Sentiment Analysis process

The sentiment classification can be described as a sentiment analysis process that consists of three main stages, sentiment identification, feature selection and sentiment
There are three types of sentiment classification depending on the provided document, documentation, sentence and aspect level. The sentence level classifications are referring to data source collected from micro-blogs, short social media posts such as tweets, customer reviews of any products, books or movies, in some cases, have their character limits which form one sentence. Document level sentiment classification refers to longer text that constructed with multiple sentences such as articles, essays, newspaper reports, book review, or public statements. The aspect level sentiment classification is complex, where the target text is stating the opinions on multiple aspects of one entity. For example, ‘This book is worth to read, also not too expensive”. Identifying the target sentiment type is critical for sentiment classification, as it can influence the final result of extracting sentiment. Because of the information in short text may not be provided with sufficient indicators to extract the opinion behind it. On the contrary, documents with longer text could provide more information that leads to more accurate subjective opinion of the document.

The feature selection task in sentiment classification can be completed by selecting manually or automatically select using statistical methods. The manual feature selection method is frequently used in lexicon based classification approaches (Medhat, Hassan, & Korashy, 2014). The core objective of feature selection is to decrease the dimensionality of the word vector space, and reduce overfitting issues for training data in machine learning classifiers, then to decrease overall computational cost (Kummer & Savoy, 2012). In the context of sentiment analysis, many commonly identified semantic features are presented and used. The term presence and frequency are features that selected based on individual words or n-grams and the word frequency count. Part of Speech, is referring to the features that facilitate classifiers to distinguish emotion indicators, usually adjectives. The Opinion words and phrases are features that obvious strong opinion indication words such as like or hate, good or bad. Negations sometime are frequently appearing in text, which can be illustrated using such example, ‘this movie is not bad’. The overall sentiment of this example expresses a positive sentiment by using a negative word in the sentence with negation word. Identifying these features and select as priori considerations over other features is beneficial for the classification performance (Medhat, Hassan, & Korashy, 2014). Some of the commonly known feature selection methods are introduced and extensively studied in Forman, (2003). Chi-Squared and Information Gain feature selection metric are the two most
commonly used feature selection approaches that maximising the precision score. The Chi-Squared was also implemented in (Ohana & Tierney, 2009), which provided optimal accuracy of two class sentiment analysis with lexical based classification approach. The Information Gain feature selection algorithm is applied in Wan et al (2015) which achieve 91.7 percent of f-measure score.

For sentiment classification, the main approaches are Lexicon based approach and machine learning approaches. The Lexicon-based approach is efficient, but a high error rate and domain dependent method to classify text. The machine learning classifiers are highly accurate, but sometimes can be inefficient when only small size of training data is available. In recent years, the use of machine learning techniques has become the main character in the field of sentiment analysis. As they are recognised as more accurate approaches for sentiment classification and enabled sentiment classification prediction on providing text or documents. (Pang & Lee, 2008) Machine Learning classification techniques can be categorised into two, supervised machine learning and unsupervised machine learning. (Kelleher, Namee, & D’Arcy, 2015)

2.3 Sentiment classification algorithms and approaches

![Sentiment classification techniques](Medhat, Hassan, & Korashy, 2014)
Above figure has provided an overview of state-of-art sentiment analysis approaches in high level perspective. The following section presents the sentiment classification approaches in more details, discussed the implementation of these approaches during sentiment classification process and evaluation of the classifier models.

2.3.1 Lexicon-Based Approach

Initially, the lexicon based sentiment analysis approach is a basic approach for analysing the sentiment of a text. This approach essentially uses a corpus similar to a dictionary, but each word within the corpus is associated with specific opinion strength and polarity. Then, using the words in the corpus to calculate the weighted average of all the sentiment scores of the provided text. For Lexicon based sentiment analysis approach, the use of different corpus will significantly affect the accuracy of the overall sentiment result for lexicon based approach. (Musto, Semeraro, & Polignano, 2014) Selecting a good lexical resource will improve the accuracy of this approach. There are so many lexical resources have been created and used in various domains and research.

The comparative research has been done by Musto, Semeraro and Polignano, which compared some widespread lexicons lexicon using the same experiment procedures and processes. The author compared SentiiWordNet, WordNet-Affect and MPQA through Twitter posts (dataset from SemEval-2013) using the same lexicon methodology. Their experiments discovered that the MPQA and SentiiWordNet are the best performing lexical resources on these datasets. (Musto, Semeraro, & Polignano, 2014) However, there are plenty of other influencing elements are not considered in this experiment, such as the topic specific elements and context of the twitter posts, since there is no way of take them into consideration as this approach breaks a complete sentence into individual word. In most cases, lexicon based approach usually produces a high precision score, but a low recall result (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011).

2.3.1.1 SentiWordNet

As part of this research, the commonly used lexical resource, the structure and logic of SentiWordNet is discussed in more details. Essentially, SentiWordNet consists of six attributes, the Part-Of-Speech (POS) tags which indicates the grammatical class of the term; the ID, the identifier of the term; the positive and negative score of the term in digit with maximum three decimal places, zero if the term has no polarity strength; the
actual content of the term; and finally the gloss of the term (Esuli & Sebastiani, 2006).

The following graph represents the visualized sense of a term. Figure 3 provided a visualized example of the term ‘estimable’ in SentiWordNet, including the positive score and negative score. There are total 115,000 number of synset terms in the SentiWordNet version 3.0

![Graphical representation of a term sense adopted by SentiWordNet (Esuli & Sebastiani, 2006)](image)

**Figure 3** Graphical representation of a term sense adopted by SentiWordNet (Esuli & Sebastiani, 2006)

<table>
<thead>
<tr>
<th>Sense</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>estimable(1)</td>
<td>deserving of respect or high regard</td>
</tr>
</tbody>
</table>
| honorable(5) good(4) respectable(2) estimable(2) | deserving of esteem and respect; “all respectable companies give guarantees”;
ruined the family’s good name |
| computable(1) estimable(3) | may be computed or estimated; “a calculable risk”; “computable odds”; “estimable assets” |

![Graphical representation of a term ‘Estimable’ (Esuli & Sebastiani, 2006)](image)

**Figure 4** Graphical representation of a term ‘Estimable’ (Esuli & Sebastiani, 2006)
2.3.2 Supervised Machine learning approaches

Analysing sentiment of a text or document using supervised machine learning approaches has been discussed and experimented by many researchers in the last decade.

This supervised machine learning approach, in short, is attempting to predict and classify the sentiment of text or documents based on the information collected or “learned” from the past examples. Essentially, supervised machine learning approaches are processed, firstly, employ a set of selected training data with annotated sentiment to the chosen supervised machine learning classifier. Then apply the unlabelled test datasets which are different from the training datasets to the trained classifier model. Finally, predict the sentiment polarity or opinions behind the test dataset. This is also domain sensitive approach, meaning that the trained classifier could produce a poor sentiment classification on a completely different dataset from another domain. This also means, the trained classifier cannot be re-used for topics unrelated with the training dataset. It is required to train the classifier using the training dataset on desired topic related dataset. This, sometimes, generates various degrees of difficulties when dealing with cross domain datasets. Early works conducted by Mullen and Collier (2004) suggested that the domain related documents should be treated with additional attention. Their paper has examined the impact of various features based on different domains or topics. As the experiment carried out in their paper pointed out the importance of the topic specific considerations.

However, the supervised machine learning approaches are able to outperform lexicon based sentiment analysis approaches. Because of the features and effecting elements are taken into account, where lexicon based sentiment approaches not.

Following supervised machine learning classifiers are widespread and often used in sentiment classification and polarity classifications:

1. Support Vector machine(SVM)
2. Naive Bayes
3. Maximum entropy Classifiers
4. Decision tree
5. Bayesian network classifier
6. Random Forest
The Bayesian network classifier and the Naive Bayes approaches are probability based approaches and both derived from Bayes’ Theorem. The distinct feature of these two approaches is that the Naive Bayes only considered the probabilities of each occurred feature in a particular text, and Bayesian network takes co-occurrences between each features into account and then calculate the probability of the sentiment polarity.

Support vector machine has been increasingly implemented throughout majority text classification applications in recent years, by reason of the high performance and sentiment classification accuracy. (Xu & Schuurmans, 2005) In (Xu & Schuurmans, 2005) has proposed a novel approach that attempts to unify and generalise the Support Vector Machine classifier for unsupervised and semi-supervised learning.

According to the comparative research conducted by Vohra et al, the SVM technique has produced a greater overall precision rate among others (Vohra & Teraiya, 2013).

<table>
<thead>
<tr>
<th>Paper</th>
<th>Dataset</th>
<th>Technique (precision, %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pang et al. [11]</td>
<td>IMDB</td>
<td>NB (81.5), ME (81.0), SVM (82.9)</td>
</tr>
<tr>
<td>Terney [2]</td>
<td>Epinions</td>
<td>PMI(66)</td>
</tr>
<tr>
<td>Dave et al. [6]</td>
<td>Amazon, CNET</td>
<td>SVM (85.8-87.2), NB (81.9-87.0)</td>
</tr>
<tr>
<td>Hu and Liu [4]</td>
<td>Amazon, CNET</td>
<td>Lexicon (84.0)</td>
</tr>
<tr>
<td>Abbasi et al. [12]</td>
<td>U.S. &amp; Middle Eastern web forum postings</td>
<td>SVM(95.55)</td>
</tr>
<tr>
<td>Aj Khan et al. [5]</td>
<td>IMDB,Skytrax, Tripadvisor</td>
<td>Lexicon(86.6)</td>
</tr>
<tr>
<td>Zhang et al. [7]</td>
<td>Luce, Yoka</td>
<td>Lexicon(82.62)</td>
</tr>
<tr>
<td>Pang et al. [16]</td>
<td>Multi-Domain Sentiment Dataset</td>
<td>ML + Lexicon (66.8)</td>
</tr>
<tr>
<td>Zhang et al. [14]</td>
<td>Twitter</td>
<td>ML + Lexicon (85.4)</td>
</tr>
<tr>
<td>Mudinas et al. [15]</td>
<td>CNET, IMDB</td>
<td>ML + Lexicon (82.30)</td>
</tr>
</tbody>
</table>

*Table 2 (Vohra & Teraiya, 2013)*

There is a preliminary research conducted which indicated that human produced opinion classification is relatively poor performance results in movie reviews. (Pang, Lee, & Vaithyanathan, 2002). However, unlike the sophistication of movie reviews sentiment analysis, the sentence level analysis in twitter may produce completely different outcomes for airline service review (Pang, Lee, & Vaithyanathan, 2002).
The research conducted by Wan et al. (2015) has proposed an ensemble sentiment classification system that combined five supervised machine learning classifiers working together to achieve an improvement on sentiment analysis of twitter about airline service. The current approach to improve the accuracy of sentiment analysis can be identified as a combination of multiple sentiment analysis. The result of the research has verified an effective approach to improve the sentiment analysis on Twitter for airline services (Wan & Gao, 2015).

Erik Cambria also has proposed, in his recent article that using a combination of semantic knowledge and machine learning approach could complement each other’s flaws (Cambria, 2016). Furthermore, the combination of Lexical knowledge and text classification has been verified as an improvement of each of the individual approach, regarding the sentiment analysis on blogs (Melville, Gryc, & Lawrence, 2009).

According to the research conducted by Wan et al (2015), the improvement of sentiment analysis approaches on airline service is to use the ensemble analysis system, a combination of the five sentiment analysis approaches, including: Naïve Bayesian method, Bayesian Network method, Support Vector Machine (SVM) method, C4.5 Decision Tree method and Random Forest method (Wan & Gao, 2015).

2.3.2.1 Support Vector machine (SVM)

Support Vector Machine (SVM) is also known as Support Vector Network (Cortes & Vapnik, 1995). The SVM is originally designed for binary, two-class classifications (Duan & Keerthi, 2005). With the development of SVM classification technique, the performance and advance classification strategy are recognised in the field of sentiment classification. Many researchers have proposed more complex SVM system to classify not only two classes, but multi class sentiments. However, much research has proposed various approaches to handle multiclass classification using SVM. Most of the proposed approaches are following the rules of combining multiple SVM classifiers to classify multi-class problems. Popular approaches are “One-against-One”, “Ong-against-All” (Hsu & Lin, 2002), DAGs (Platt, Cristianini, & Shawe-Taylor, 2004) and Error-Correcting output codes (Dietterich & Bakiri, 1995). The empirical study of multiclass SVM classification system conducted by Duan and Keerthi (2005), suggested that the
One-against-One combination of the SVM approach outperformed the other classifiers. Furthermore, the comparative study conducted by Hsu and Lin (2012), also provided strong evidence that One-against-One is stronger than other implementation approaches, when dealing with inadequate training dataset. The following diagram shows the single Support Vector Machine classifier.

![Figure 5 Example of classification in two dimensional space. (Cortes & Vapnik, 1995)](image)

The above figure shows the core strategy of SVM classifier. The supervised SVM classifier requires the training dataset in order to calculate and produce the optimal hyperplane that separates and classify the input data into two categories as shown (Figure 1). The extra feature is added by the kernel function when optimal hyperplane is not present in the current dataset. When handling multi class issues, the two dimensional space is replaced with higher level of multi-dimensional space, by implementing a combination of several single SVM classifiers.

### 2.3.2.2 Naïve Bayes

One of the child product of the Bayes’ Theorem approach is the Naïve Bayes approach. The difference is that the text features are independently considered in Naïve Bayes approach. In other words, the Naïve Bayes assumes each feature is independent and no relationships created between each class. Kelleher, J et al (2015) in their research suggested the Naïve Bayes can be described as:
'A naive Bayes model returns a MAP prediction where the posterior probabilities for the levels of the target feature are computed under the assumption of conditional independence between the descriptive features in an instance given a target feature level.'

\[
\mathbb{M}(q) = \arg \max_{l \in \text{levels}(i)} \left( \prod_{i=1}^{m} P(q[i] | t = l) \right) \times P(t = l)
\]

Figure 6 Formal expression of Naive Bayes model (Kelleher, Namee, & D'Arcy, 2015)

The equation calculates the posterior probabilities of each occurrence of the class, positive, negative and neutral, independently based on the training data provided. The prediction of final classification is determined by the highest probability score over the others. The implementation of Naive Bayes approach for text classification is widely used, due to the advantages such as ease of training, able to handle streaming data, and efficiency.

2.3.2.3 Bayesian network

The Bayesian Network Classification technique is also a probability based classifier, which extend the logic of Bayes’ Theorem. It specifies the joint conditional probability distributions and uses a direct acyclic graph to represent the relationship between the subset of features. As the dependencies are considered in the Bayesian network classifier, the relationships between features are highly dependent on each other. For instance, in the text representation word vector, the bigram attributes are highly dependent on unigram attributes.

2.3.2.4 Decision tree

The Decision tree classification approach is a technique that making classification predictions by carrying out a series of true or false decisions. The decision tree approach is the foundation for the implementation of Random forest modal and C4.8 decision tree model.
A visualised Decision Tree example is produced using the Weka GUI tool, which generated using 500 training data, see figure 3. For purpose of text classification, the tree consists the word ‘great’ as the root node of the tree, other words such as ‘passengers’, ‘today’ are represented as internal nodes. Finally, the class words, positive, negative are leaf nodes. The decision tree approaches provides an efficient way of classifying text without any domain knowledge needed and parameter configurations. However, with that being said, the decision tree also introduces the overfitting problem (Tan, Steinbach, & Kumar, 2005).

2.3.3 Unsupervised Machine Learning

Early work, Turney (2002) conducted, has suggested an unsupervised machine learning approach to classify the documents only using two polarity seed words “excellent” and “poor” that achieved the accuracy of 74% overall based on the four topics experimented on. The topics analysed involves the reviews of movies, automobiles, banks and travel destinations. The (Atserias et al., 2006; Padró & Stanilovsky, 2012) has proposed the approach, which is an unsupervised dependency parsing-based approach using a lexicon resource, created by means of an automatic polarity expansion algorithm and natural language processing techniques. Their approach was based on determining dependencies between lemmatized tagged words using a sentiment propagation algorithm that took into account
and distinguished between key linguistic phenomena, namely, intensification, modification, negation and adversative and concessive relations.

The literature suggested an approach that leverages a variety of natural language processing techniques and sentiment features primarily derived from sentiment lexicons. These lexicons were created by the means of a semiautomatic polarity expansion algorithm in order to improve accuracy in a specific application domain. The proposed approach consists of three main procedures. First, Lexical and syntactic analysis, which the input document is tokenised and each word is pos tagged, then transformed into the dependency tree using FreeLing Parser (Atserias et al., 2006; Padró & Stanilovsky, 2012). Secondly, the creation of lexicon resources employs some polarity lexicons, such as SO-CAL, WordNet, etc. Then, they improved the lexicon coverage by acquiring polarities for subjective words not present in generic dictionaries and adapt their scores using the available data. Thirdly, sentiment analysis through propagation, which they defined the rules for dealing with specific cases such as intensification, modification, negation and adversative/concessive relations.

As discussed in the previous section, employing a unified and generalised SVM technique to the unsupervised machine learning approach can also be effective for text classifications. (Xu & Schuurmans, 2005) Similar approach are also conducted in (Shafiabady et al., 2016), which combined the automatic test clustering techniques with the Support vector machine technique together to achieve an unsupervised learning approach for classifying documents. The proposed method employs Self Organising Map as text clustering component. Then apply the clustered data with unlabelled texts as a test dataset to the SVM model, which will predict the sentiment of texts. In addition, Shafiabady et al. (2016), has suggested that the work proposed is feasible where experts’ knowledge are not available.

2.3.4 Sentiment Classification Evaluation

The previous section introduced in detail regarding to the state-of-art sentiment analysis approaches. It is necessary to discuss the evaluation of the classification model, in order to analysis the performance and result of the classifiers. First of all, the confusion matrix needs to be discussed. It is introduced in (Apté, Damerau, & Weiss, 1994), which provided a perspicuous view for analysing and evaluating the performance of
classification results. The confusion matrix for multi-class classification can be illustrated by following table:

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>Actual class</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>True positive (1,1)</td>
<td>False Negative (1,2)</td>
</tr>
<tr>
<td>Class 2</td>
<td>False positive (1,2)</td>
<td>True positive (2,2)</td>
</tr>
<tr>
<td>Class 3</td>
<td>False positive (1,3)</td>
<td>False Negative (3,2)</td>
</tr>
</tbody>
</table>

*Table 3 Multi-class Confusion matrix example*

As the predictions made by sentiment classifier, the number of correct and incorrect prediction result is shown within the table shown above. The green cells indicate the correctly classified instances, the true positive results. The white cells indicate the number of incorrectly predicted class.

There are many evaluation measurements for sentiment classification models available in the literatures, Recall, precision and f-measures are the most commonly used evaluation strategies for sentiment classifications (Jain & Nemade, 2010). Recall and precision extensively use the advantages of confusion matrix to evaluate the performance of the classifier (Apté, Damerau, & Weiss, 1994).

The F-measure is developed to eliminate the imbalance of precision and recall measure created. It is the harmonic mean of precision and recall. Also, the weight of precision and recall is defined in the formula. “Recall is the fraction of the correctly classified instances of one class of the overall instances in this class” (Melville, Gryc, & Lawrence, 2009). The recall can be described as, the result of the actual correct result of each model (number of correctly classified positive tweets), divided by the total number of relevant data (total number of positive tweets). The number of correct classified tweets is determined by human labelled sentiment polarity (Apté, Damerau, & Weiss, 1994).

“Precision is the fraction of the correctly classified instances for one class of the overall instances which are classified to this class.” The precision can be described as, the number of correct classified results (number of correctly classified positive tweets) divided by the total number of classified results (Total number of positive tweets classified).

The following equation expressed the calculation of F-Measure :
The parameter $\beta$ represent the weight of the precision and recall, in this case, it will be 1, which the weight of precision and recall is the same (Apté, Damerau, & Weiss, 1994).

2.4 Sentiment analysis and Airline service

2.4.1 Sentiment analysis on airline service using Twitter Data

Sentiment analysis in the field of airline service is similar to the product review sentiment analysis in such way that both are classifying the feedback of a purchasable product. However, there are still many differences in terms of domain focus and concentration of the result. For instance, classifying book reviews from Amazon, ‘This book is worth reading,’ is expressing a positive opinion regarding to the book. The word ‘book’ is a noun in this sentence, but in airline service the word ‘book’ has a higher chance to act as a verb, such as ‘I am not be able to book this flight’

In the field of sentiment analysis, much research has been done using different resources collected, and most of them are domain dependent. It is discovered that the number of research published regarding to airline sentiment analysis using Twitter data was little. Although there were voluminous unpublished experimentations conducted which can be found online. These experiments employed many different technologies such as Java, R, and Python to perform the sentiment classification tasks. For instance, an approach using R as a classification tool to perform the classification process was proposed in the post (Breen, 2011). The visualised sentiment results are also produced in his blog, so that it provides an intuitive understanding for the general public. Based on the evidence analysed from these experiments, the similar classification approaches were used, including machine learning techniques and lexicon based techniques.

“One in two Twitter users says Twitter content is influential in their consideration of travel brand” (Elrhoul, 2014). The Twitter blogger Meghann Elrhoul has conducted a survey with Millward Brown that shows how travel brands are using Twitter to boost their image. The blog highlighted that the implementing customer service strategies base on Twitter platform can directly reflect the travel brand image (Elrhoul, 2014). Discard whether this blog is trustworthy or not, the evidence of increasing number of official travel companies’ active Twitter accounts has proved the importance of adopting the
Twitter platform for travel companies. In airline service domains, for instance, many airline companies have provided 24/7 customer service using Twitter, which increased the responsiveness especially when dealing with complaints and inquiries. Twitter also enabled companies to notify latest promotion products to customers and receive direct customer opinions regarding the deal when they retweet or comment on it, which is another advantage of using the Twitter platform. In this context, implementing automatic sentiment analysis techniques can be very effective comparing with using human resources when marketing analysis tasks are required for a specific promotion. Additionally, many consumers tend to consider the feedbacks and opinions posted by other Twitter users who have experience with the products and services before purchasing. This can significantly influence the consumers’ decisions on which airline company they are about to choose.

2.4.2 Existing Ensemble approach in airline service

Further sentiment analysis research in relation to the airline services domain has been done by Wan and Gao. The research proposed an ensemble sentiment classification system that combined five classifiers working together to achieve an improvement on sentiment analysis of twitter about airline service. Erik Cambria also has proposed, in his recent article that using a combination of semantic knowledge and machine learning approach could complement each other’s flaws (Cambria, 2016). Beside the combined machine learning classification techniques, the combination of Lexical knowledge and text classification has also been verified as an improvement of each of the individual lexicon based approach, in relation to sentiment analysis on blogs (Melville, Gryc, & Lawrence, 2009). However, it is very difficult and time consuming to compare these two combination approaches as there are many differences are implemented. The result of the research proposed by Wan and Gao on airline service has verified it is an effective approach to improve the accuracy of sentiment analysis using Twitter for airline services (Wan & Gao, 2015). The research focused on applying sentiment classification processes on collecting Twitter data relevant to airline service domain using commonly used classification algorithms, lexicon based approach, probability approach and decision tree approach individually in comparison with the ensemble classifier to validate the proposed approach. The research discussed in detail regarding the sentiment
classification processes and individual classifier, also provided empirical contribution on sentiment analysis in the airline service domain.

The proposed ensemble supervised machine learning approach employed the Majority vote rules as combination rule to produce the final prediction of the text. However, the need for identifying the best suitable combination algorithm is required, which has not been addressed in Wan and Gao’s research. On the other hand, the combination of lexicon based sentiment classification approach, the combination algorithms were analysed in Ohana et al (2011) research. The empirical study conducted suggested that Sum Rule has outperformed other rules such as majority vote and Max Rule combination algorithm, across many domains, file, hotels, electronics, books, apparel and music reviews. The best sentiment classification accuracy achieved in this research is 80.23% in hotel reviews. The validation of combination algorithms is still required in the airline service domain, especially the classification approach is not lexicon based, in Wan and Gao’s research.

To systematically summarise the research conducted by Wan and Gao (2015), there is a need to discuss the experiment undertook in this research. First of all, the Twitter Search API was used to collect airline service related tweets using keywords ‘flights’ with one airline company’s name at a time for each search query, including Delta Airlines, United Airlines, SouthWest Airlines, Air Canada, JetBlue Airways, etc. The dataset then, is manually labelled with three classes, ‘positive’, ‘negative’ and ‘neutral’. Then the data pre-processing tasks were applied which include removal of all symbols, hashtag signs, links, emoticons and punctuations. Information Gain was used for feature selection algorithm, and 10 fold cross validation was used for evaluation of the classifiers. The five machine learning classifiers were selected based on the evaluation result of each individual classifier which then formed the hybrid classification system. The accuracy of the system then was later analysed using precision and recall techniques combining with f-measure, as they are the common measure of the sentiment classifier performance. The Wan and Gao’s hybrid classification system managed to achieve a satisfactory performance result of 84.2% in precision, 84.2% in recall and f-measure.

2.5 Conclusion

This chapter discussed in details regarding the state-of-art research papers in the field of sentiment analysis. The chapter started with a higher level review of sentiment analysis
from real world applications of sentiment classification such as email spam filters and recommendation systems. Then the categories of sentiment analysis were discussed such as the strength of the sentiment, the polarity of the sentiment, and the aspect level sentiment classification. The processes of sentiment classification are also concluded. The related field involved during sentiment classification process such as Natural Language Processing and text representation of the document is extensively discussed. Then, section 2.3 intensively discussed the state-of-art machine learning approaches in sentiment classification, including Naive Bayes, Bayesian Network, SVM, decision trees and random forest. The lexical-based classification approach is further reviewed with a commonly used corpus SentiWordNet 3.0 (Esuli & Sebastiani, 2006). This chapter ended with a discussion of the relationship between sentiment classification and airline service domain and existing studies for sentiment classification in airline service domains, especially the research conducted by Wan et al (2015). Furthermore, the dataset used for airline service sentiment analysis is discussed, in terms of its role and its relevance for airline service opinion mining. The requirements of the experiment for this project were also presented.

To this end, the next chapter will echo the discussion made in this chapter and illustrate the design of sentiment classification experiment, in order to achieve the objective of this project.
3 DESIGN AND METHODOLOGY

3.1 Introduction

In this chapter, the design of this research experiment is discussed in detail. As discussed in the Introduction Chapter, the experiment for this research consists of three distinct procedures. First, the collected datasets are pre-processed and transformed into required format in order to feed the classifier. Then, the training dataset and test dataset are constructed from the original dataset, which the training dataset will be used to train the Lexicon Ensemble classification system. Finally, the test dataset is used to allow the trained Lexicon Ensemble classification algorithm to perform prediction of the sentiment polarity and evaluation of the model and results.

In addition, the structure of the twitter dataset is presented and analysed in detail. The detail considerations of the experiment procedure are also explained and analysed in this chapter. The configuration and construction of the Lexicon Ensemble Classifier are discussed and presented at the end of this chapter.

3.2 Data understanding and opinion mining

3.2.1 Data source

As part of the research, the dataset used during sentiment classification process plays a very important role, as it can significantly impact the classification performance. According to the review of state-of-art approaches in the field of sentiment classification, the selection of the sentiment classification dataset depends on many factors, the objective of the classification, the domain focus, the data structure and so on. Considering the objective and the domain focus discussed in Chapter 1, the dataset is required to be related to opinions on airline services closely and consists of polarity sentiment regarding the service. With the increasing popularity of employing Twitter data for sentiment classification purpose (Wan & Gao, 2015), (Wakade, Shekar, Liszka & Chan, 2012), employing Twitter data is also considered in this research.

The dataset considered is obtained from CrowdFlower, as it fits the requirement of this research and efficient to use. Because this dataset does not require further labelling of
the sentiment, it has already been labelled manually to state the actual sentiment of the tweet, along with the reason of the sentiment. Additionally, the dataset selected contains full content of the tweet where the Twitter Search API can only obtain only limited content of the tweet. It is also the most up-to-date dataset that contain only airline service relevant tweets.

The dataset obtained contains 13572 number of labelled tweets with the tweet created date, username, content, sentiment confidence and labelled class, etc. However, the information gathered other than the sentiment confidence, the content of the tweet and labelled sentiment are less important in this research as they only provide additional data of a tweet such as create time of the tweet and the owner of this tweet. The dataset note provided by Crowdflower stated that the sentiment confidence is the probability of the class being labelled. This attribute contains numeric values in range zero to one with three decimal numbers, which one indicates the labelled sentiment is certain and zero is not. In this research, the content of each tweet is considered as the determine factor for the result of tweet sentiment. The tweets collected in each dataset are labelled in three classes, positive, negative and neutral. The number of tweets distribution for each class is shown as follows:

<table>
<thead>
<tr>
<th>Number of Tweets</th>
<th>Positive</th>
<th>Negative</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2151</td>
<td>8561</td>
<td>2862</td>
</tr>
</tbody>
</table>

Table 4 Number of tweets distribution for each class

In the collected dataset from Crowdflower, the number of airline service providers obtained are Virgin America airline, United airline, Southwest airline, Delta airline, US Airways and American airline.

3.2.2 Dataset analysis and considerations

The data collected and used in this research are varied in different forms, also due to the nature and the characteristics of social media post. The content of each tweet is informally written in many different ways that sometimes human beings are not able to understand its underlying opinion without looking up the Internet. Human languages are subtleties that one word can be represented with many characters and symbols or a combination of these. For instance, the word ‘wait’ can be represented as ‘w8’ in a tweet, ‘4’ can be read as word ‘for’, ‘nvr’ can represent the word ‘never’ and etc. There are so many other informal ways it can be represented, but also keep the underlying opinion of
the tweet remain intact. This also an issue that this research will be considered in order to produce more accurate results on tweet sentiment classification model.

The hashtag is one of the widespread feature in Twitter, and has been formally used in online blogs and other popular social media platforms. Hashtags are the user specified topic keywords in tweet prefixed by ‘#’ (Wang & Zheng, 2014). Hashtags increase the exposure of the tweet itself, because adding hashtags can be treated as a categorization base on the keywords used. In other words, the hashtags in tweets are similar to keywords under abstract section in literatures, however, they are not bound to any language or grammar rules. It may contain incorrectly spelled words, abbreviations, acronym and etc. Moreover, hashtags must not contain any spaces between words. It has to be written in consecutive letters. A representative example from the collected dataset could be, ‘#worstflightever’, which is a sequence of three words, worst, flight and ever joint together with no space in between. For tweet sentiment analysis, hashtags are similar to the example stated above are critical information that must be included and feed into classification process. The reason is that the result produced from lexicon based classifier are entirely depending on the weight of each word contained in the tweet, especially for adjectives such as ‘worst’ in the example. Expand the hashtag is also an indication of an increase of statistic information gathered for machine learning classifiers. The research conducted by (Prusa, Khoshgoftaar, & Seliya, 2015) suggested that increase of the dataset used for training could produce classification results in a higher accuracy rate. In the collected dataset, hashtags in tweet are also frequent phenomenon. Thus, expanding the hashtag in the experiment can be taken into consideration.

In many twitter sentiment classification related research, the existence of emoticons in microblogs have been discussed and many have proposed different approaches to deal with classification of emoticons in the tweet. Also experiments result discovered by Wang and Castanon (2015) in their research, suggested two things when classifying tweet sentiment are, First, emoticons are strong and reliable signals for sentiment polarity classifications and second, the emoticons exist in tweet should be treated with caution as most of the emoticons sometimes are expressed very complicated in different contexts. One of the approaches when dealing with classifying the polarity of tweet with emoticons, is to replace the emoticons with suitable emotion words. The replacement of emoticons with emotion words works best with the lexicon classifier, since the emotion words can directly reflect and enhance the weight of the emoticons (Sahu, Rout, &
Mohanty, 2015). The accuracy of the tweet classification with emoticons replaced is achieved with 80% in Sanket, Suraj and Debasmit’s research.

In addition, considering the expansion of hashtags and replacement of emoticons. The collected datasets also reveal the frequent appearances with abbreviations and use of slangs including Internet slangs and cultural slangs. Especially for micro-blogs like Twitter, with the limitation of 140 characters for each post, many long spelling words are abbreviated into a few characters. Phases or terms such as ‘laugh out loud’ become ‘lol’, ‘direct message’ becomes ‘dm’ and ‘private message’ becomes ‘pm’ in tweets, etc.

In twitter polarity classification process, abbreviations will be ignored as they are not considered as useful words can be used for classification, particularly for lexicon classifiers which it depends on the weight of each word appeared in a tweet.

Further discovery on the datasets collected are made includes the appearance of html entities such as ‘&amp;’ and appearance of airport code in both capital characters and non-capital characters. The portion of the tweets collected in dataset consists of destination or departure airport code. For polarity classifications, the airport names are less significant comparing with emotion words in lexicon classifiers, due to the weight they are assigned. However, they could provide an increase of statistical information for machine learning classifiers included in the Lexicon Ensemble Classification system.

Considerations of HTML entities are also taken into account, which they can be un-escaped into its actual character.

### 3.3 Comparison factors

#### 3.3.1 Classification accuracy

For most of the sentiment classification research conducted, the accuracy of the classification result produced are evaluated by calculating the ratio between correctly classified tweet and incorrectly classified tweets. However, as it is discussed in the previous chapter, the calculation of precision, recall and f measure are known to be common accuracy evaluation method for sentiment classification result. The experiment adopts this measure across the whole experiment process.
3.3.2 Training dataset

As a part of this research, the machine learning algorithms require training resources before implementing it for the test data or unknown class dataset. In order to maximise the best classification results with machine learning techniques. The training dataset size was determined by comparing the accuracy of different size of training dataset applied. It has been discussed in Chapter 2 that 10 fold cross validation technique is used in Wan et al (2015). However, it is still required to evaluate other training dataset selection techniques such as 7 fold cross validation or 3 fold cross validation, which used in (Ohana & Tierney, 2009). In this project, the 7 fold cross validation and 10 fold cross validation were used for compression purpose and then selected accordingly.

3.4 Experiment Design

The process of this experiment will reflect the considerations discussed in the previous section. The detail design of the experiment is discussed in the following sections.

3.4.1 Data preparation

During the data preparation phase, the collected datasets are processed based on the consideration discussed. However, the general text pre-processing techniques are adopted as required for text polarity classification. The processes are explained as follows:
1. All tweets are decapitalised into lower case letters.
2. Any appearances of blank characters such as tabs, enters, multiple spaces are all replaced with a single blank space character.
3. The special characters in each tweet discovered except emoticons are removed.
4. Extract each word within a tweet and compare it against the US airport code collected. The airport code is expanded to full address. For instance, ‘ACY’ will be replaced with ‘Atlantic City International Airport, Egg Harbor Township, NJ,’
5. Extract each word within a tweet and compare it against the Twitter slang words dictionary constructed, then replace the matching word with its corresponding full text. For instance, ‘THX’ can be changed to word ‘Thanks’.
6. Unrelated text, username and URL are deleted.
7. Extracted hashtags in each tweet is broken into correctly spelled words, such as ‘#nothappy’ is broken into ‘not’ and ‘happy’.
8. Finally the tweet that contains emoticons will be expanded. The emotion is replaced with suitable alias. For instance, the sad face is replaced with text ‘sad’.

3.4.2 Baseline classifier

In sentiment classification research, baseline classification results are produced during the early stage of the experiment. As the ensemble classification system has proposed in the research conducted by Wan et al (2015). This approach consists of five machine learning classifiers that collaborate together to classify one tweet, then the final result of the polarity of a tweet is decided by the most classified results by each classifier, ie Majority Vote. In accordance of this research, the ensemble classification system proposed by Wan et al (2015) will be used as baseline classifier, in order to compare with the proposed model in this research. In this research, the baseline classifier is built with five machine learning classifiers using Weka text sentiment analysis Java API, as it was described in Wan et al (2015). This baseline classifier will then evaluate and predict the polarity of each tweet individually. After five classifiers have completed with classification on one tweet. The final result is produced using the Majority Vote algorithm. Where each result classified by each classifier will be considered a vote to produce the final result. The weight of each vote is equally divided. When votes are tied, the result will be assigned to an arbitrary result between the two results. Because there are five votes, the only tie scenario is when either of two classes have two votes each and the other class has only one vote.

The datasets processed in section 3.3.1, then is used as input for the baseline classifier.

3.4.3 Ensemble Classifier with Lexicon

The Lexicon ensemble classifier consists of six classifiers which include, Naïve Bayesian method, Bayesian Network method, Support Vector Machine (SVM) method, C4.5 Decision Tree method, Random Forest method and finally the Lexicon-based classifier.

Based on the baseline classifier built, the additional lexicon classifier is integrated and implemented at the same level with other classifiers. In other words, the prediction made by lexicon based classifier for each tweet will be considered along with predictions made by other machine learning classifiers before final prediction is made. According to the
baseline classifier, the equivalent configurations are applied to the five machine learning classifiers in the proposed system, and the lexicon based classifier will be built based on the commonly used lexical resource, SentiWordNet word list (Esuli & Sebastiani, 2006), as discussed in Chapter 2.

3.4.4 Implementation and design

With the considerations made in Section 3.3.1, the implementation of this experiment will be carried out in an iterative fashion. In order to maximise the accuracy and identify the impact of the data pre-processing on final result, each feature within the tweet is considered individually. Changing the content of the tweets during pre-process phases may or may not influence the result of sentiment prediction. In this case, each change of the feature is implemented incrementally as per experiment, such as expanding slang words, hashtags etc.

First of all, the general data pre-processing process will be carried out which does not directly modify the content of each tweet. This pre-processing task includes removing duplicate tweets and retweets. Then, each character within the tweet is decapitalised. The processed tweets, then will be used as the base dataset for the first classification experiment and further changes of pre-processing tasks will be conducted on the base dataset. Feature selection process will be performed against the base dataset as previously discussed in section 3.3.4.

The classification of the base dataset was first classified using both proposed Ensemble Classifier with Lexicon classifier and Ensemble Classifier using five machine learning techniques to produce the baseline result of this experiment. The baseline of the experiment will be repeated with the five combination rules using fixed configuration for each individual classifier. Then the combination rule which produced the best accuracy result will be selected as the base classifier.

After selecting the base classifier, the experiment will be implemented cumulatively with tweet content modified by each consideration made in the previous section.

- The first stage of the experiment evolved several steps that could contribute for comparison purpose against the final experiment results. First, remove all
usernames and URL. Then remove all non-alphabetic characters such as Unicode (emoji), digits and punctuations.

Both ensemble classifiers will be fed with previously processed dataset for training and testing.

- The consideration of expanding the content of each tweet will be applied to the base dataset, using the base classifier selected in the previous step. The abbreviations and emoticons will be replaced with appropriate full words and aliases, by using the defined Twitter abbreviation list and Full Emoji Library Java API. As discussed in the previous chapter, some of the aliases of emoji are replaced with suitable weighted words to reflect the appropriate positive or negative score for lexicon classifier. The username and URL in the tweet are removed as they make little impact on the overall sentiment orientation. Expansions of hashtags will be applied to the base dataset, according to the discussion in Chapter 2.

The processed tweet dataset, then feed into both, the proposed ensemble classifier with Lexicon and ensemble classifier with five machine learning classifiers, to train and predict the tweet sentiment in three classes, positive, negative and neutral. The result produced will then be used for evaluation and comparison with previous result and base line classification result.

3.4.5 Evaluation

The designed experiment analyses two sentiment analysis approaches, the ensemble sentiment classification system, by Wan et al, which includes the Naïve Bayesian, Random Forest, Bayesian Network and newly proposed ensemble sentiment analysis approach with Lexicon based sentiment analysis combined. The F-measure technique will be used as the performance evaluation for the proposed classification approach. The F-measure is defined as so far the most appropriate methodology for determining the accuracy of the text classification for many types of sentiment analysis approaches (Pang & Lee, 2008). It is proven that the ensemble approach performs much better than any individual approach based on recent research conducted by Wan et al (2015). The
The implementation of F-Measure provided a balanced comparison between precision and recall, thus, it can be used as a suitable classification performance measure.

In order to achieve the objective of the research, the hypotheses proposed in this research will be testified using the designed experiment. As the principle of f-measure described previously, if the newly proposed ensemble sentiment analysis approach has achieved the highest f-measure score, then h2 will be accepted. The same principle will apply for all hypotheses proposed.

The Confusion matrix will also be presented as an evaluation of the classifier performance, as discussed in Chapter 2. In the Weka API, the confusion matrix is also provided that can be used for evaluating the performance of the classifier.

### 3.4.6 Conclusion

The main objective of this chapter is discussed in detail regarding the design of the experiment procedures will be carried out in order to achieve the objective discussed in Chapter 1. Early in this chapter includes the discussion on the dataset employed in the experiment, its structure and content. This chapter also analyses in detail that the consideration of the factors involved when modifying and expanding the content of the tweet, such as hashtags, URLs, abbreviations and the use of slangs. Next, the comparison factors of the experiments are discussed to echo the studies conducted in Chapter 2, which the evaluation of the performance of classifiers and the considerations of the training dataset is discussed.

At the end of this chapter, the detailed design of the experiment processes is discussed, in order to achieve the objective of this research. There are three main datasets which will be obtained using different pre-processing procedures discussed based on the consideration of tweet content, the general processed tweets with duplicate tweets removed and decapitalised content, the dataset with tweet content modified and expanded, then the dataset with tweet features removed. The proposed classifier and baseline classifier will then be used to classify each dataset and the results will be evaluated.

The implementation of the designed process of the experiment is discussed in the next chapter.
4 EXPERIMENT IMPLEMENTATION

4.1 Introduction

In Chapter 2, the state-of-art sentiment classification approaches were presented and discussed. The sentiment classification processes were also studied in detail in order to perform the sentiment classification experiment. Chapter 3 presents the related considerations on tweet dataset, the factors that the experiment required for comparison purpose and the evaluation strategies for result of this classification experiment. The design of the experiment will be discussed in detail: data pre-process considerations, experiment phases and the proposed ensemble classification approach.

This chapter echoed the designed experiment processes introduced in the Chapter 3 and discuss the implementations of the experiment. This chapter presents the implementations of designed experiment using proposed sentiment classification approach. The objective of this chapter is to determine the performance of the proposed method in comparison with the baseline classification results produced using the evaluation strategies stated in the previous chapter.

4.2 Experiment process

This section is mainly focused on discussing the designed sentiment classification processes carried out in the implementation phase. The outline of the design of the experiment processes can be illustrated in figure 11:

![Figure 8 Experiment process flow](image)
4.2.1 Data pre-processing

The dataset is pre-processed and stored into three different entities in the MySQL database. They are labelled with A, B and C, so it can be used during the classification stage:

A. General pre processed
   A.1 All characters are decapitalised. Any appearance of multiple blank spaces are replaced with a single space.

B. Tweet Content Modified (using data A from step 1)
   B.1 The US airport codes within the tweet are expanded with its full name. The airport code is expanded to full address. Unrelated text, username and URL are deleted.
   B.2 Extract each word within a tweet and compare it against the Twitter slang words dictionary constructed, then replace the matching word with its corresponding full text. For instance, ‘THX’ can be changed to word ‘Thanks’. The full slang list can be found in Appendix section 8.3
   B.3 Extracted hashtags in each tweet is broken into correctly spelled words, such as ‘#nothappy’ is broken into ‘not’ and ‘happy’.
   B.4 Finally the tweet that contains emoticons will be expanded. The emotion is replaced with suitable alias. For instance, the sad face is replaced with text ‘sad’.

C. Partial tweet content deleted (Using data A from step 1)
   C.1 Usernames, URLs and Unicode (Emojis) are removed from Tweets. Remove all non-alphabetical characters from tweet.

The dataset A is the original tweet dataset from CrowdFlower that has not been pre-processed and modified. The purpose of this dataset A is to allow baseline results to be produced and used for further pre-processing for dataset B and C.

Dataset B is pre-processed based on the original dataset A, the dataset B focused on modifying the content the tweet. The purpose of this dataset is to produce the classification result with tweet content modified which can be used to compare against baseline result.
The dataset C is pre-processed upon original data, dataset A. The dataset C aims to produce the dataset for sentiment classification when the features of Twitter data are not considered, such as Username, URLs, Emoji and non-alphabetical characters.

4.2.2 Parameter settings

As part of the sentiment classification experiment, the proposed model employed five machine learning classifiers and one lexical-based classifier working together to produce the final classification results for each tweet. The parameters for each individual classifier are discussed in this section, as each classifier is required to be configured to perform classification tasks.

According to the discussions made in Chapter 2 and the Weka API used for this experiment, the parameters were set for each machine learning classifier as the code provided in Appendix A section 8.1. The parameter configurations are set based on the discussion made in Wan et al (2015) as the experiment is closely related. As discussed in Section 3.4.3, the parameter settings for lexicon-based classifier is not applicable.

As discussed in chapter 2, the text representation of trigram, was used in this experiment. In the Weka API, the n-gram tokenizer was adopted to this experiment. The n-gram tokenizer transforms the text data into word matrix before it was used for classification. The N-gram tokenizer was set with maximum size of 3 and minimum size of 1. The Java code snippet was given in Appendix A 8.2.1 for text to word matrix setting.

4.2.3 Data sampling

It is noticed that the three classes have unequal number of tweets with large differences in the dataset. There are 8561 negative tweets, but only 2151 neutral tweets. In this case, the existence of imbalance class will affect the performance of the training dataset and the prediction of the result from each classifier. Thus, in order to select a balanced dataset for both training and testing purposes, the lowest 2151 number of tweets were selected from each class, 6453 number of tweets were used in total. As the majority of the tweets are negative, the sentiment confidence attribute is considered as a selection factor for selecting 2151 number of neutral and 2151 number of negative tweets. According to the data source, the dataset contains the labelled sentiment with the sentiment confidence that reflects the probability of the labelled sentiment. In the dataset sampling phase, this sentiment confidence attribute is considered as an important factor.
for selecting 2151 numbers of tweets from each negative class and neutral class. The tweets were sorted in descending order by sentiment confidence scores, and the top 2151 tweets were then selected for this experiment.

The initial experiment for baseline classification was carried out using the 6453 numbers of tweet from the dataset, 2151 for each class. However, the performance of the initial experiment was really below expectations, since it requires nearly 3 hours (214 minutes to be exact) to complete one dataset classification using 7 fold cross validation, without feature selection process involved. The classification result can be shown in the confusion matrix as following tables:

<table>
<thead>
<tr>
<th>Classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>Neutral</td>
</tr>
<tr>
<td>1577</td>
</tr>
<tr>
<td>164</td>
</tr>
<tr>
<td>410</td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td>336</td>
</tr>
<tr>
<td>882</td>
</tr>
<tr>
<td>933</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>329</td>
</tr>
<tr>
<td>220</td>
</tr>
<tr>
<td>1537</td>
</tr>
</tbody>
</table>

Table 5 Baseline classification result

<table>
<thead>
<tr>
<th>Evaluation criteria (Overall)</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Accuracy</td>
<td>61.92 %</td>
</tr>
<tr>
<td>F-Measure</td>
<td>61.20 %</td>
</tr>
<tr>
<td>Recall</td>
<td>61.90 %</td>
</tr>
<tr>
<td>Precision</td>
<td>63.8 %</td>
</tr>
</tbody>
</table>

Table 6 Accuracy criteria comparison

Note: (Simple accuracy is the percentage of correct classified tweet)

The baseline classification task using 10 fold cross validation, was also carried out, however, the Java heap size exceeds exception was thrown during programme run time. With evidence of the issue investigated, and attempts to solve this issue. The limitation of technical resource was discovered, and reducing the size of dataset was found the most effective solution for this issue, as there are 4,5621 number of attributes generated after work tokenisation task. In the final test, there are 1800 number of tweets were selected, 600 tweets for each class. Based on the performance of the classifier the classification task for one dataset was improved to 8 mins for completion.
4.2.4 Feature selections

As discussed in section 2.2.5, the feature selection processes are implemented against the target dataset individually for comparison and select the best result produced. The Chi-squared metrics are adopted in Ohana and Tierney (2009) which enable irrelevant features been removed during the process in order to maximise the outcome. However, the Information Gain feature selection method is adopted in Wan et al (2015) for airline service sentiment classification process. Thus, it is also employed in this experiment. The core of implementing the feature selection technique is to maximise the performance of the classification process, as discussed in Chapter 2. Thus, the feature selection task was performed against the 1800 number of tweets. In Weka, the wrapper class AttributeSelection is used, which implements the Information Gain evaluation and a Ranker search algorithm to sort the attributes based on the information gain of the attribute. Based on the result produced by Attribute selection algorithm, there are 1801 attributes were selected in the dataset.

4.2.5 Training set selection

As discussed in Chapter 3, the size of the training data set can directly affect the accuracy of supervised machine learning algorithms. Selection of training data size also depends on the complexity and the quality of the whole dataset. The 10 fold cross validation is used in Wan et al (2015), which produced promising overall accuracy result. However, other research has produced optimal result using 7 fold cross validation rules. In this experiment both 7 and 10 fold cross validation were performed to determine the training data size.

As discussed, the comparison of using 7 and 10 fold cross validation was made using the baseline classifier with five machine learning models and Majority Vote algorithm.

<table>
<thead>
<tr>
<th>Fold</th>
<th>F-measure</th>
<th>Simple Accuracy</th>
<th>Time in minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>71.104 %</td>
<td>71.000 %</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>71.987 %</td>
<td>72.000 %</td>
<td>7</td>
</tr>
</tbody>
</table>

*Table 7 Performance comparison for training size*

Based on the table above, the results indicated that the 10 fold cross-validation rule has provided better accuracy result than 7 fold cross validation rule, although, the 7 fold cross validation is 2 minutes faster than 10 fold cross validation. Thus, the 10 fold cross-validation rule is selected in this experiment for further use.
4.2.6 Combination rules

In accordance of discussion in Chapter 2 and 3, the comparison of using different combination algorithms was performed, in order to select the appropriate approach for the system. The Weka API provided with five combination algorithms, including the schemes presented in (Kittler, Hatef, Duin, & Matas, 1998).

<table>
<thead>
<tr>
<th>Combination algorithm</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAJORITY VOTING RULE</td>
<td>50.866%</td>
</tr>
<tr>
<td>AVERAGE RULE</td>
<td><strong>51.768%</strong></td>
</tr>
<tr>
<td>MAX_RULE</td>
<td>47.277%</td>
</tr>
<tr>
<td>MIN_RULE</td>
<td>43.862%</td>
</tr>
<tr>
<td>PRODUCT_RULE</td>
<td>43.862%</td>
</tr>
</tbody>
</table>

*Table 8 Performance comparison of combination algorithms*

The evidence suggested that the average rule combination algorithm outperformed the others. Thus, it is selected for further classification process.

4.2.7 Sentiment classification

As discussed in Chapter 3, the implementation of the Twitter classification process is carried out in three stages. First, the baseline classification is executed for both baseline classifier and proposed classifier, using the dataset processed from A. Secondly, the classification process was implemented in the same configurations, but using B dataset. Thirdly, the C dataset is used in the same process as the previous ones.

The baseline classification is carried out against both classifiers using the three pre-processed datasets. The purpose of this process is to establish the baseline result not only using the baseline classifier, but also to establish a baseline result before the content of the tweet is modified and expanded. Whether there are any of the improvements of classification performance or decrease of the performance can be evaluated based on the result produced.

During the classification phase, there are total six different datasets classified for comparison purpose. As discussed earlier in this chapter, three main datasets were used, the dataset A, dataset B and C. In dataset B, the classifications were implemented by modifying tweet content cumulatively, as discussed in chapter 3. During the process of implementing this experiment, additional classification was made to evaluate the
classification result, this is referring to sentiment classification of tweets with only emoji replaced with text representation. The results are discussed in the next chapter. The next section discussed the project environment which the experiment was implemented in.

### 4.3 Environment Setup

#### 4.3.1 Configuration

In this research, the MySQL database is implemented to accomplish the seamless connection with Java application. The following table shows the configured environment applied:

<table>
<thead>
<tr>
<th><strong>Hardware</strong></th>
<th><strong>Software/OS</strong></th>
<th><strong>Processing Language and package</strong></th>
<th><strong>Dataset</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>6GB RAM Intel i5-4310M 2.70GHz dual-core</td>
<td>Windows 10 Intellij IDE MySQL Server/WorkBench Weka 3 GUI</td>
<td>JAVA 8 JDBC connector Weka Java API FullEmoji V4.0 API SentiWordNet 3.0 word list</td>
<td>Twitter airline service collection</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Dataset</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter airline service collection</td>
</tr>
</tbody>
</table>

*Table 9 Environment configuration*

#### 4.3.2 Configuration Discussion

In order to achieve the objectives of this research, Weka open source Java library is adopted in the experiment. Weka API is written in Java and designed specifically for data mining tasks. The Weka Java library provides collections of machine learning
algorithms that can be invoked directly from Java code. The library includes functionalities for pre-processing, classification, regression and visualisations. Weka also provided with User interface application, which can be directly implemented with selected data mining tasks. However, in this project, only the Weka Java library is considered due to the hybrid classification is highly customised for the specific classification process. It provides the opportunity of integrating the lexicon based classification approach with the five machine learning classifiers using the Java API provided by Weka. Although, there are many other data mining tools available to use, such as RapidMiner tool used in (Ohana & Tierney, 2009), Weka Java API is still the primary choice for integrating lexicon based classifier with machine learning classifiers.

The MySQL database is employed for storing the tweet datasets purpose, due to the features it provides and the programming language used for this experiment. The FullEmoji API is applied for the replacement of emoticons task within each tweet. The FullEmoji API provides a list of emoji characters (2389 number of emoticons to be exact) in various format, Unicode, alias, and keywords. More importantly, it is written in Java programming language that can be invoked directly from the sentiment classification project. Based on the objective of this research, the task of implementing FullEmoji API to replace emoji in tweets with textual representation is only considered and the workflow of the API is not concerned. Additionally, as the observations made on the text representation of each emoji in Fullemoji chart. The some emojis are not replaced with textual data, such as the ‘thumb up’ and ‘hundred points’ emojis. In FullEmoji API, the ‘thumb up’ emoji is represented with ‘+1’ and the ‘hundred points’ is represented with ‘100’. As these representations cannot be accepted based on the consideration of lexicon based classification approach, the terms are replaced with high positive score words, ‘great’ and ‘superb’.

As the exceed Java heap size exception discovered during the implementation of this experiment, one possible cause is the limitation of the environment setup. Thus, increasing the computational power and RAM size can also be an improvement when dealing with large dataset for sentiment classification tasks.

4.4 Conclusion

This chapter aims to contribute to achieving the objectives of this project and validate the proposed classification using processed datasets. Early this chapter tends to echo the
procedures designed in Chapter 3. The considerations made in Chapter 3 were further implemented and discussed in this chapter. The implementation of the designed experiment was discussed in the first section, including data pre-processing tasks, dataset sampling, feature selection, training dataset selection and combination algorithm selection. During the implementation of the experiment, the limitations and challenges were discovered, which led to using data sampling techniques to resolve the limitation discovered. The solution provided has effectively resolved the issue and has produced satisfactory result, however, the limitations or side effects of this solution were not discovered during the process. Later in this chapter presents the environment of the implemented experiment. The discussion on the environment setup is also presented.
5 RESULT AND EVALUATION

5.1 Introduction
This chapter presents the result of the implemented experiment which discussed in chapter 4. In the first section of this chapter, additional comparison results for baseline results are presented. As discussed in the last chapter, the baseline results are presented and discussed with comparison of incremental changes made to the datasets. Later in this chapter, the evaluations of experiment results are conducted along with the analysis made on misclassification and limitations.
This chapter aims to provide structured experiment result to validate the statement made in Chapter 1 and attempt to achieve the objectives of the project.

5.2 Experiment results
This section is essential for validating the objective of this research. The baseline results are presented with evaluation criteria discussed in Chapter 2. The classification performance results are presented including baseline result and further experiment results.
Due to both of the classifiers are implemented against all six numbers of datasets, the results presented are shown based on two classifiers
1. Ensemble classifier (only machine learning classifiers)
2. Proposed ensemble classifier (machine learning with lexicon classifier)
The result shown in the next section is the weighted average accuracies for different experiment phase.

5.2.1 Classification result
First of all, the baseline results are presented in table 10. These results were produced using the baseline classifier built using five machine learning classifiers. According to the objective of this research, the dataset A is used for baseline classification. The tweets in dataset A are not being deep processed, except general processes, all tweets are decapitalised into lower case letters, any appearances of blank characters such as tabs, enters, multiple spaces are all replaced with a single blank space character. The result of the baseline accuracy classification results of the tweets are shown as follows:
1. Baseline Classification result:

<table>
<thead>
<tr>
<th>10 fold cross validation with baseline classifiers</th>
<th>Dataset A (without changes)</th>
<th>Dataset B (Expansion of tweet)</th>
<th>Dataset C (tweet feature removed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Accuracy</td>
<td>72.0 %</td>
<td>71.61%</td>
<td>71.61%</td>
</tr>
<tr>
<td>Precision</td>
<td>73.06 %</td>
<td>73.12%</td>
<td>73.12%</td>
</tr>
<tr>
<td>F-Measure</td>
<td>71.98 %</td>
<td>71.63%</td>
<td>71.63%</td>
</tr>
<tr>
<td>Recall</td>
<td>72.00 %</td>
<td>71.61%</td>
<td>71.61%</td>
</tr>
</tbody>
</table>

Table 10 Performance result of baseline classifier

The dataset A in this experiment is represented as the baseline result produced by the baseline classifier with dataset A employed, which achieved 72% in simple accuracy, 73.06 % in precision, 71.98 % in f-measure and 72% in recall. The further classification process was performed against the other datasets where the content of the tweet is modified. The purpose of the performed task is to establish comparison with the result produced using the proposed ensemble classification approach. The highest overall accuracy results produced when Dataset C used, is likely due to the noisy data introduced in Dataset B where Dataset C eliminated the possible noisy data. As discussed in chapter 3 and 4, the proposed ensemble classification model is implemented with pre-processed dataset. The results are shown in the following section.

2. Proposed Ensemble classification approach:

<table>
<thead>
<tr>
<th></th>
<th>Dataset A (without change)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Accuracy</td>
<td>57.89 %</td>
</tr>
<tr>
<td>Precision</td>
<td>96.96 %</td>
</tr>
<tr>
<td>F-Measure</td>
<td>71.11 %</td>
</tr>
<tr>
<td>Recall</td>
<td>57.89 %</td>
</tr>
</tbody>
</table>

Table 11 Performance result of proposed ensemble classifier with dataset A

Based on the result achieved using proposed ensemble classification approach, the basic accuracy of the classifier reached 57.89%, precision achieved a relatively high score of 96.96%, F-measure 71.11 % and recall 57.89%. It is noticeable that the precision score achieved is quite high, one of the possible reasons could be the imbalance of the each
class exists in the training dataset, caused by the feature selection process. More details of the result evaluation will be discussed in section 5.3

<table>
<thead>
<tr>
<th>Dataset B (content modified)</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Simple Accuracy</strong></td>
<td>56.39 %</td>
<td>52.98 %</td>
<td>54.38 %</td>
<td><strong>61.30 %</strong></td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>97.67 %</td>
<td>97.68 %</td>
<td>97.46 %</td>
<td><strong>98.30 %</strong></td>
</tr>
<tr>
<td><strong>F-Measure</strong></td>
<td>70.26 %</td>
<td>67.32 %</td>
<td>68.69 %</td>
<td><strong>74.56 %</strong></td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>56.39 %</td>
<td>52.98 %</td>
<td>54.38 %</td>
<td><strong>61.30 %</strong></td>
</tr>
</tbody>
</table>

*Table 12 Result for using Dataset B*

The result presented in table 12 contains the results produced by dataset B implemented with the proposed ensemble classifier. This experiment was performed cumulatively as discussed in Chapter 4. First, the dataset B1 was used, which the airport codes existed in each tweet was expanded. It achieved 56.39 % in basic accuracy, 97.67 % in precision, 70.26 % in f-measure and 56.39 % in recall. Comparing to the baseline result, the simple accuracy was decreased.

The dataset B2 then is used for classification. The B2 dataset contains the tweets with expanded airport codes and slang words (The list of slang words used is provided in Appendix section 8.3). The simple accuracy result obtained after the slang words were replaced has decreased by 3.41 %. However, there is a slight increase in the precision score.

The dataset B3 was classified and the results are presented in Table 12. By comparing with the results from B2, an improvement of the simple accuracy, f-measure and recall were noticed. The dataset B3 has modified the content of tweets with the expansion of airport codes, slang words and hashtags. However, the overall accuracy result is still relatively lower than result obtained using dataset B1 with only airport codes expanded.

The classification experiment using dataset B4 was performed and the results are presented in table 12. As it was shown in the table, the overall accuracy performance has achieved the best result so far as accumulative modifications were made to the content of each tweet. The improvement of 3.14% on simple accuracy was discovered, comparing with accuracy results obtained from table 11, when no changes applied to the content of each tweet.

As discussed in Chapter 4, the improvement of classification results has motivated the experiment of only replacing the emoji in the tweet with its corresponding text.
representation to be performed. This task was performed in order to further validate whether only change the emoji can improve the result based on classification results obtained using dataset B4. The details of the results are shown in table 13.

<table>
<thead>
<tr>
<th></th>
<th>Dataset B4</th>
<th>Dataset B5 (only replace emoji in tweets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Accuracy</td>
<td>61.30 %</td>
<td>61.29 %</td>
</tr>
<tr>
<td>Precision</td>
<td>98.30 %</td>
<td>97.79 %</td>
</tr>
<tr>
<td>F-Measure</td>
<td>74.56 %</td>
<td>74.34%</td>
</tr>
<tr>
<td>Recall</td>
<td>61.30 %</td>
<td>61.29%</td>
</tr>
</tbody>
</table>

*Table 13 Accuracy result for B5*

As the result shown in table 13, there is no significant improvement found when classification process was implemented with only emoji in tweets were replaced with text.

The table 14 presents the final classification results using dataset C where all the features of tweets were removed, including usernames, URLs and Unicode (Emojis), only alphabetical characters from tweet were kept in each tweet.

<table>
<thead>
<tr>
<th></th>
<th>Dataset C (tweet feature removed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Accuracy</td>
<td>63.08 %</td>
</tr>
<tr>
<td>Precision</td>
<td>97.94 %</td>
</tr>
<tr>
<td>F-Measure</td>
<td>76.03 %</td>
</tr>
<tr>
<td>Recall</td>
<td>63.08%</td>
</tr>
</tbody>
</table>

*Table 14 Result for using Dataset C*

As the result shown in table 14, it has achieved the best accuracy result with 63.08% in simple accuracy and 76.03% in f-measure, comparing with dataset B when the content of the tweets is modified.

Based on the empirical study of this experiment, further experiment was performed using the baseline classifier, and the results were presented earlier in this section, as discussed. In addition, the above results shows relative high percentage of precision. One of the possible reason can be the impact of neutral class tweets, as the multi-class classification contains neutral tweets that can create meaningless data that affects prediction result. The feature selection process is also an influencing factor that requires precise threshold adjustment to maximize the noisy data elimination process.
The classification experiment results were presented in this section, the evaluation of the results and classifier performance will then be discussed in the next section.

5.3 Evaluation and analysis

This section presents a detailed analysis and evaluation on results obtained from the classification experiment with the baseline classifier and proposed classification system. The objective of the research is also reviewed and validated based on the result achieved from the experiment implemented in Chapter 4. In this section, the evaluations are made based on two main factors. First, the evaluation of the performance results conducted by different classifiers, the proposed classifier and baseline classifier. Second, the evaluation of the performance results produced when modified datasets were used.

5.3.1 Accuracy evaluation of classifiers

As the results presented in section 5.2, the proposed sentiment classification approach that combing lexicon classifier has achieved the best simple accuracy result of 63.08 % when the features of the tweet are removed from the text. However, the baseline classifier has achieved the best simple accuracy result of 74.17 % with the same dataset used. In accordance with the result observed in section 5.2, the baseline classification accuracies remain to be the highest percentages achieved throughout the experiment process, discard the considerations of different datasets applied.

Based on observed accuracy results produced, the influencing factors can vary, one could be considered as the integration of the lexicon based classifier with machine learning classifiers has many rooms to be improved. On the other hand, the precision accuracy results produced by the ensemble classifier needed to be highlighted as it achieved the best result of 98.3%. This high percentage of precision score has indicated that average 98.3% of the tweets were predicted correctly for each class, which performed well for the true positive classification result. However, the high precision score often leads to low recall. In this case, the recall result is 61.3 %, meaning that for each class, only 61.3 % of the tweets are correctly predicted. For instance, there are 100 numbers of positive tweets, only 61 number of positive tweets are correctly predicted. In this case the f-measure result should be considered carefully. Because it provides valuable performance result based on a balance of precision and recall. The f-measure calculated based on the classification results produced by both classifiers, suggested that
the proposed classification approach outperforms the baseline classifier by 1.69%. However, there is need to consider the trade-off when using proposed classification approach, that is, the low recall of classification results. Furthermore, based on the evaluation made, the H2 hypothesis stated in Chapter 1, is accepted.

5.3.2 Accuracy evaluation on dataset used

As discussed at the beginning of this section, the comparison is also conducted based on the dataset used for each classification process, in order to achieve the objective of this research. The following figure shows the performance changes based on how the content of the tweet is modified with the implementation of the proposed classification approach.

![Classification performance result with different dataset used](image)

Based on the flow of result presented in figure 9, tweet document classification with data pre-processing techniques applied to dataset C has produced the highest f-measure, accuracy and recall among others. The comparison of these classification performance validated that modification against tweet content can affect the performance of the classifier. The above results indicated that removing the features within each tweet such as the usernames, URLs and emoji can improve the performance of the proposed classifier. The comparison results from B1 to B4 also provided such insight when the content of the tweets is expanded, little affects were made on the performance of the proposed classifier. When expansion of the airport code applied in dataset B1, the performance of the classifier was declined, as well as the expansion of tweet slangs was
applied in dataset B2. On the contrary, when the accumulative expansion of hashtags applied, the classification of dataset B3 results in a slight improvement on all accuracy measures. Finally, when the replacement of emoji task was performed on the dataset. The improvement of classification performance was exposed. Thus, it is verified that expanding the hashtags and replacing emoticons within the content of the tweet can be applied during data pre-processing phase based on the purpose of improving proposed classifier’s accuracy when classifying Twitter documents on airline service domain. The above analysis focused on evaluation on correctly classified tweet documents, therefore, discussion on incorrectly classified tweets will be made in the next section.

5.3.3 Misclassification Analysis

As many sentiment classification experiments conducted in past research, the analysis of misclassification can be critical for improving the performance of the classifiers in further studies. In the experiment carried out for achieving the goal of this project, the three class sentiment classification was implemented using the proposed ensemble classifier which contains lexicon based classifier and the five machine learning classifiers. The dataset used for this experiment contains customer opinions on various airline service providers. For analysing the misclassification of the tweet document, following two incorrectly classified tweets are extracted from dataset C.

1. “united i would like your baggage damage number as well another great thing from your trained staff whats the number please claim time” predicted positive, actually negative.

2. “united the lack of customer service is astounding daysofhell” predicted positive, actually negative.

In the first case of the misclassification, it is noticeable that the appearance of sarcasm in the tweet. For the ensemble classifier used in this case. The probability based machine learning classifiers can hardly capture this feature, as it may not be provided in the training dataset. For lexicon based classifier, it can be really challenging, as the sarcastic word used has a higher positive score which increases the chance of producing positive results. Additionally, the second case of misclassification has the similar phenomenon of having positive words and negative word together in one tweet, if considered from lexicon based classifier. The word ‘lack’ has weak negative score and ‘astounding’ has a higher positive overall score in SentiWordNet. As the probability machine learning
approach concerns, the appearance of the word ‘astounding’ was appearing more often in the positive tweets and more frequent than the probability of the word ‘lack’ appeared in the negative tweets. Thus, it was predicted as positive tweet. In the second tweet, it is also noticeable that the term ‘daysofhell’ is used, which is a strong indicator of negative tweets from human perspective. However, neither lexicon based classifier nor machine learning classifier were able to capture it. One of the reasons can be considered is that three words in the term used are not correctly separated, which cause the failure of identifying the negative indicators. Furthermore, as the investigation made based on original tweet, the term ‘daysofhell’ is the hashtag used within this tweet. Based on the purpose of increase the accuracy of the prediction, expanding hashtags can be considered which was already implemented during the experiment process. However, the overall performance result suggested otherwise. The reason being is that some of the tweets with hashtags expanded may not provide useful indicators for correct predictions. For instance, when the sarcastic words were used in hashtags.

5.4 Conclusion

This chapter presented experiment results combining with the performance of the classifiers based on the processes implemented and described in Chapter 4. The results were shown and organised based on the results obtained from baseline classifier and proposed classifier. The chapter highlighted the comparisons of classification performance between the baseline classifier and the proposed classifier by employing the comparison criteria discussed in Chapter 2. Then the impact of modification and expansion of the tweet content of classification performance of both classifiers was discussed. It includes the discussion on declines of the classification accuracy when airport codes and slang words were expanded, the increase of classification accuracy when expanded hashtags and replaced emoji dataset employed for classification.

Later in this chapter, evaluation of the classification result was conducted for each result obtained from the experiment. Additionally, the analysis of misclassification was also presented in detail.

Additionally, based on the results obtained, the key findings of this experiment and evaluation can be presented as follows:
1. The proposed ensemble classification approach outperforms the machine learning combined ensemble classification approach when classifying twitter data on airline service domain, as additional lexicon based classifier is added.

2. Expansion of the hashtags and emoji within tweet content can improve the accuracy of the sentiment classification results, due to the additional information they can contribute to sentiment classification.
6 CONCLUSION

6.1.1 Introduction

This chapter concludes this dissertation project. Early of this chapter reviewed the objectives, limitations, challenges and the advantages during the research in sentiment classification on airline service. The discussion of the contribution of this research is also presented. The third section reviewed the overall result of the experiment carried out in the project in terms of reflecting the objectives of the project. Finally, this chapter concluded with work performed in this research, and discussed in detail regarding the future opportunities, research direction in the field of sentiment analysis on airline service domains.

6.1.2 Research overview and objective

The research objectives stated in Chapter 1 have been discussed and reviewed throughout each chapter. As the experiment performed in Chapter 4 and the results obtained in Chapter 5, the following objectives are achieved based on the completion of the experiment.

- Analyse and discuss the related topic in the field of sentiment analysis, Natural Language Processing techniques and lexicon resource creating and implementation approaches.
- Review state-of-art sentiment classification approaches. Investigate the advantages and disadvantages of these approaches.
- Investigate the existing text mining techniques in the field of sentiment analysis.
- Design and implement approaches reviewed from previous objectives.
- Evaluating the classification result obtained from baseline classifier using various measurement techniques investigated.
- Design and construct proposed classification method using equivalent configurations implemented in baseline classification approach.
- Evaluation and analysis results obtained from baseline classifier and proposed classification strategy.
- Critical investigation on obtained result of classification and error of misclassification for the proposed ensemble classification approaches.
- Identify and analysis improvement or demotion of the new analysis strategy comparing with existing approaches on Twitter regrading to airline service.

6.1.3 Problem definition

This research reviewed the state-of-art sentiment classification approaches and existing approaches in airline service domain. As the valuable insights that implementation of sentiment analysis techniques can provide for airline service providers based on the goal of improving their services, determine the most appropriate and best performance approach to employ is necessary. Much research in the field of sentiment analysis has been done in order to improve the performance from many aspects. First lexicon based classification approach was introduced, then the implementation of machine learning techniques dominated the field of sentiment classification. However, the machine learning techniques have their own limitations and objectives that cannot be achieved. To complement the limitations of each machine learning classifiers, the hybrid approaches were introduced. One of which the research conducted by Wan and Gao (2015) has suggested the significant improvement of classification performance using social media data on airline service domain. Furthermore, the research on hybrid sentiment classification approach was not only limited combining only machine learning approach or only lexicon based approach, but the hybrid approach of implementing lexicon and machine learning approaches collectively (Mudinas, Zhang, & Levene, 2012) which the increase of performance was shown in software and movie reviews. In this project, the ensemble approach that consists of both machine learning approaches and lexicon based approach was analysed which suggested the gain of its classification performance compares with machine learning only approach on airline service domain (Wan and Gao, 2015).

6.1.4 Experiment, evaluation and results

During the implementation experiment phase, the designed experiment processes were successfully carried out, as discussed in Chapter 4. The original tweet dataset collected was divided into three different datasets with various pre-processing techniques applied, the general pre-processed dataset, the dataset with tweet content expanded and the dataset with all features of the tweet removed. These datasets were then used for comparison and validation purpose discussed in Chapter 3.
The dataset sampling process, feature selection process and selection of training dataset size were implemented accordingly during the experiment phase. The dataset sampling process was implemented in order to resolve the limitations of computational power of environment setup and the time spent to complete the classification process. The implementation of feature selection employed Information Gain evaluation strategy, which also contributed to this issue, but the main purpose is to maximise the accuracy of the classifier by selecting the features that can contribute the most to the classification prediction. In addition, according to the results produced by the 10 fold cross validation is more accurate than the 7 fold cross validation strategy. The 10 fold cross validation was selected in the experiment.

In accordance of the experiment carried out in this project, the results obtained were satisfactory for the purpose of this research, by comparing the accuracy measures outlined in Chapter 3. The best baseline performance results achieved by the ensemble machine learning classifiers were 74.19% in F-Measure, and 74.17% in simple accuracy. In comparison with the proposed classification approach, it improved and achieved f-measure of 76.03%, however, only obtained simple accuracy with 63.08%.

Based on the evaluation experiment results discussed in Chapter 5, the improvement observed on f-measure was originated by the high precision score, which indicated the advantages of implementation of this approach. Another reason of the appearance of this phenomenon can also be considered, due to the machine learning classifiers used and the combination algorithm employed in the proposed system. Further discussion was made in Chapter 5 section 5.3.2 regarding the accuracy performance evaluations when modified tweet content was used for classification. The result obtained implied the improvement of classification result can be originated when expansion of hashtags and replacement of emoji characters tasks was applied to each tweet. However, classification on tweet dataset with removal of the irrelevant components and non-alphabetical characters such as usernames, URLs and emoji has obtained the highest accuracy result, discard the different classifiers used.

In the context of airline service domain, the proposed method is ideal for airline marketing research for determining the quality of the provided services in customer perspective, due to the high precision score achieved using this system.
6.1.5 Contributions and impact

The objective achieved based on the experiment implemented in this research has contributed to the field of opinion mining in many aspects. The outcomes of this research can be featured as the contributions to the field of study.

The core element of the research experiment was to implement the sentiment classification process designed and predict the underlying sentiments on social media content collected which consists of customer opinions on airline service providers. The project presents the proposed classification approach with traditional lexical based classification approach and the supervised machine learning approaches in parallel for classification of Twitter documents. The proposed method was then evaluated based on state-of-art classification measurements, namely, recall, precision and f-measure. The result obtained from the experiment can be used as reference in the related field of future research.

As the experiment implemented, the project demonstrated the ensemble classification approach with a combination of lexicon based classifier and machine learning classifiers can produce improved classification result using three-class Twitter dataset related with airline service topic, comparing with the existing approach (Wan and Gao, 2015). In a wider context, the results accomplished by the proposed classification approach validated the potential improvement can be made in the airline service sentiment analysis. It is recognised that the pattern of high precision score results obtained by the proposed method can provide significant support and enable productive airline service research to be executed by the airline service provider.

The project also identified the impact on modification of Twitter document for sentiment classification tasks. The expansion of Twitter features such as slang words, hashtags and emoji can improve the overall classification performance. However, it is still a relatively little improvement, comparing with these features removed from the text. On the other hand, this also facilitates to eliminate the unnecessary effort required for the data pre-processing task, as it still can produce more accurate classification results when these features are removed in each tweet.

6.1.6 Future work and research

The research was performed based on the knowledge contributed by researchers in the field of opinion mining, NLP and machine learning. As the results of this project were
observed and discussed during experiment evaluation phase, many limitations are still exist to overcome. Many concerns were covered and implemented in the experiment of this project, however, there are still many other factors not considered, such as other techniques of modifying text content. The future work considered based on the result obtained in this project can be listed as follows:

- As the modification of the twitter content was made cumulatively, individual modification tasks can be considered for further classifications experiment. There are possibilities that the accumulative experiment performed in this project when using the subset in Dataset B can affect the overall classification result and it cannot represent the improvement or downside of changes made for each element modification in the tweet. Thus, perform and evaluate each tweet modification process (Expand hashtags, replace emoji with text, expand slang words) individually can be significant to demonstrate their influences on sentiment classification prediction process.

- Considerations of implementing various lexicon resources for lexicon based classifier in the proposed classifier, other than SentiWordNet. There are many lexical resources available for use, the corpus created by Bing Liu (2015), VADAR and etc. However, using the existing external lexical resources sometimes can be insufficient when dealing with domain specific sentiment classifications.

- Constructing custom lexicon resources can be beneficial when dealing with domain specific sentiment classifications. In this case, creating a new corpus is the main focus on future work of this research. Although it can be time consuming to construct the lexicon, it still can maximise the performance of the system for airline service sentiment analysis. Creating new lexicon resource requires the words to be assigned with correct and suitable sentiment score. This can be achieved with the features extracted by using existing reviews from other product review websites. Because of the unique features that product review websites provide, the strength of the sentiment can be extracted along with the review content. In this case, extracting the sentiment indicators combine with the strength of each review, the weight of the terms in the lexicon can be adjusted with the most appropriate sentiment score based on the topic chosen, in order to achieve the improvement of classification performance.
7 BIBLIOGRAPHIES


Paquette, H. (2013). Social Media as a Marketing Tool: A Literature Review, University of Rhode Island, Rhode Island, Kingston, USA


Kummer, O., & Savoy, J.. (2012). Feature Selection in Sentiment Analysis. CORIA.


8 APPENDIX

8.1 Parameter settings

8.1.1 Random Forest:

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag Size Percent</td>
<td>100</td>
</tr>
<tr>
<td>Number of Execution Slots</td>
<td>1</td>
</tr>
<tr>
<td>Break Ties Randomly</td>
<td>False</td>
</tr>
<tr>
<td>Calculate Out Of Bag</td>
<td>False</td>
</tr>
<tr>
<td>Max Depth</td>
<td>0</td>
</tr>
<tr>
<td>Number of Decimal Places</td>
<td>2</td>
</tr>
<tr>
<td>Number of Features</td>
<td>0</td>
</tr>
<tr>
<td>Number of Iterations</td>
<td>100</td>
</tr>
<tr>
<td>Seed</td>
<td>1</td>
</tr>
</tbody>
</table>

8.1.2 J48 Decision Tree

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary Splits</td>
<td>False</td>
</tr>
<tr>
<td>Collapse Tree</td>
<td>True</td>
</tr>
<tr>
<td>Confidence Factor</td>
<td>0.25</td>
</tr>
<tr>
<td>Minimum number of object</td>
<td>2</td>
</tr>
<tr>
<td>Number of fold</td>
<td>3</td>
</tr>
<tr>
<td>Number of Decimal Places</td>
<td>2</td>
</tr>
<tr>
<td>Reduced Error Pruning</td>
<td>False</td>
</tr>
<tr>
<td>Subtree Raising</td>
<td>True</td>
</tr>
<tr>
<td>Seed</td>
<td>1</td>
</tr>
<tr>
<td>Unpruned</td>
<td>False</td>
</tr>
<tr>
<td>Use Laplace</td>
<td>False</td>
</tr>
<tr>
<td>Use MDL correction</td>
<td>True</td>
</tr>
</tbody>
</table>
8.1.3 Support Vector machine (LibSVM)

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coef0</td>
<td>0.0</td>
</tr>
<tr>
<td>Cost</td>
<td>1.0</td>
</tr>
<tr>
<td>Degree</td>
<td>3</td>
</tr>
<tr>
<td>EPS</td>
<td>0.001</td>
</tr>
<tr>
<td>Gamma</td>
<td>0.0</td>
</tr>
<tr>
<td>Kernel Type</td>
<td>Radial basis</td>
</tr>
<tr>
<td>Loss</td>
<td>0.1</td>
</tr>
<tr>
<td>Normalization</td>
<td>False</td>
</tr>
<tr>
<td>Seed</td>
<td>1</td>
</tr>
<tr>
<td>Nu</td>
<td>0.5</td>
</tr>
<tr>
<td>Shrinking</td>
<td>True</td>
</tr>
<tr>
<td>Probability Estimates</td>
<td>False</td>
</tr>
<tr>
<td>Number of Decimal Places</td>
<td>2</td>
</tr>
</tbody>
</table>

8.1.4 Bayesian Network

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimator</td>
<td>simpleEstimator</td>
</tr>
<tr>
<td>Search Algorithm</td>
<td>K2</td>
</tr>
<tr>
<td>Max num of parent</td>
<td>1</td>
</tr>
<tr>
<td>Score Type</td>
<td>Bayes</td>
</tr>
<tr>
<td>UseADTree</td>
<td>False</td>
</tr>
<tr>
<td>Number of Decimal Places</td>
<td>2</td>
</tr>
</tbody>
</table>

8.1.5 Naïve Bayes Multinomial

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Decimal Places</td>
<td>2</td>
</tr>
</tbody>
</table>
8.2 Word Matrix (Trigram)

8.2.1 Word Matrix settings

```java
StringToWordVector stringToWordVector = new StringToWordVector();
try {
    stringToWordVector.setIDFTransform(true);
    stringToWordVector.setTFTransform(false);
    stringToWordVector.setDoLog(false);
    stringToWordVector.setAttributesIndices("first-last");
    stringToWordVector.setDoNotCheckCapabilities(false);
    stringToWordVector.setMinMaxTermFrequencies(false);
    stringToWordVector.setInvertSelection(false);
    stringToWordVector.setLowerCaseTokens(true);
    stringToWordVector.setMinTermFrequencies(1L);
    stringToWordVector.setPeriodicPruning(10L);
    stringToWordVector.setSaveDictionaryInBinaryForm(true);
    stringToWordVector.setStopWordsHandler(new Rainbow());
    nGramTokenizer = new NGramTokenizer();
    nGramTokenizer.setNGramMaxSize(3); nGramTokenizer.setNGramMinSize(1);
    stringToWordVector.setStringTokenizer(nGramTokenizer);
    stringToWordVector.setWordsToKeep(9999);
    stringToWordVector.setInputFormat(instances);
    //data instances that you are going to input to the filter
    Instances instances1 = filter.useFilter(instances, stringToWordVector);
    ArffSaver saver = new ArffSaver();
    saver.setInstances(instances1);
    saver.setFile(new File("StringToU.arff"));
    saver.writeBatch();
} catch (Exception e) {
    e.printStackTrace();
}
```

8.3 Twitter slang word list

Abt=About
AMPG=Above my pay grade
Ack=Acknowledge
AFAIC=As Far As I am Concerned
BRB=Be Right Back
b/c=Because
B4=Before
BTW=Bye the way
CC=Carbon Copy
Civ=Civilization
CO=Commanding Officer
CYA=Cover Your Ass
CTFU=Cracking the Freak Up
Ctfu=Cracking the Freak Up
FISH=Fck It Shit
F-U=Fcked You
FU=Fcked You
FM=Fcking Magic
FB=Facebook
FOCL=Falls Off Chair Laughing
F/U=Follow Up
FFT=Food For Thought
4=For
FCO=For Crying Outloud
FCOL=For Crying Outloud
4COL=For Crying Outloud
FFS=For Fuck Sake
FTL=For The Loss
FTR=For The Record
Forward=For The Win
FWIW=For What Its Worth
4U2C=For You To See
FYA=For Your Action
FYI=For Your Information
FYR=For Your Records
YT=YouTube
AFAIK=As Far as I Know
RT=Retweet
B4=Before
BFN=Bye for now
BGD=Background
BH=Blockhead
BR=Best regards
BTW=By the way
CD9=parents are around
CHK=Check
CUL8R=See you later
DAM=Dont annoy me
DD=Dear daughter
DF=Dear fiance
DP=used to mean profile pic
DS=Dear son
DYK=Do you know
EML=Email
EM=Email
EMA=Email address
FTF=Face to face
F2F=Face to face
FB=Facebook
FF=Follow Friday
FOTD=Find of the day
FTW=For the win
FUBAR=Fed up beyond all repair (slang from the US Military)
FWIW=For what its worth.
GMAFB=Give me a fing break
GTFOOH=Get the f out of here
GTS=Guess the song
HAGN=Have a good night
HAND=Have a nice day
ZOMG=OMG to the max
ATM=At The Moment
ATW=All the way
TQRT=Thanks for the retweet
TQRF=Thanks for the follow
DM=Direct message
PM=Private message
LOL=laughing out loud
NBD=no big deal
NVM=never mind
TBH=to be honest
IDC=I don't care
IMO=in my opinion
NVR=never