ENHANCEMENT OF SENTIMENT ANALYSIS SCORING MECHANISM – A CASE STUDY ON MALAYSIAN AIRLINE INDUSTRY

RAYVENDRAN VISVALINGAM

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

University of Malaya

ENHANCEMENT OF SENTIMENT ANALYSIS SCORING MECHANISM – A CASE STUDY ON MALAYSIAN AIRLINE INDUSTRY

RAYVENDRAN VISVALINGAM

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF COMPUTER SCIENCE

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

University of Malaya

UNIVERSITY OF MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Rayvendran Visvalingam

Matric No: WGA140053

Name of Degree: Master of Computer Science

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Enhancement of Sentiment Analysis Scoring Mechanism-A Case Study on Malaysian Airline Industry

Field of Study:

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

University of Malaya

ABSTRACT

Sentiment polarity calculation is a method to gage the strength of a sentiment extracted from a text. Many tools have been developed with their respective scoring mechanism in order to produce an effective sentiment score. Semantic Orientation Calculator (SO-CAL) is one of the lexicon-based tool that is incorporated with important features (such as intensifiers, negation and etc.) to calculate the sentiment polarity of a text. However, this tool has its limitation in processing misspelled word especially in repeated letters or characters that may lead to sentiment inaccuracy. The accuracy of SO-CAL is when processing social media text that mostly contains misspelled word is low. Thus, an enhanced scoring mechanism (LexiPro-SM) was developed to improve the sentiment scoring considering misspelled word especially on words that contain repeated letters. The LexiPro-SM was tested on the posts that were collected from the Facebook official pages of two major airline industries in Malaysia, which will be referred to Airline A and Airline B respectively. Three important phases were involved the development of LexiPro-SM which are, data collection, data cleaning and data analysis. Data collection was performed with the aid of Facebook Graph API to collect three months' posts from the both airlines. Data cleaning was performed by removing noise leaving only text that contains alphabets and exclamation mark. Improvement was made on the scoring mechanism and incorporated in LexiPro-SM with the features that can process misspelled word and also other improved features such as negation and exclamation mark. Then clean data of the airline was analyzed with LexiPro-SM and SO-CAL. A web-based portal was developed to visualize the LexiPro-SM's result of the two airlines, where each airline has own page with overall score chart, polarity group chart and sub-services chart. Sub-services chart is a new idea implemented in this research to categorize the overall services into sub-services such as customer service, price, preflight and facility. This would be helpful for the airline management to improve their

University of Malaya

service by narrowing down their attention into a particular service. The airline pages are also linked in order to show the comparison results between Airline A and Airline B. Based on these results, a case study was conducted between the two airlines where the observation shows that Airline A achieved a high positive score than Airline B. Moreover, to assess the effectiveness of LexiPro-SM , the both results of LexiPro-SM and SO-CAL was compared by performed evaluation measures using evaluation metrics (such as accuracy, recall, precision and F1-score) with the reference of human expert results. From the evaluation it shows LexiPro-SM achieved higher accuracy (90.7%) than SO-CAL (58.33%). Overall, in LexiPro-SM the improvement made has increased the accuracy of sentiment detection and produced a better result than SO-CAL. This concludes processing misspelled word is an important process in social media sentiment analysis. This is further proved with the reference to the case study, where a conclusion was formed as Airline A providing a better service than Airline B.

University of Malaya

ABSTRAK

Pengiraan markah sentimen ialah satu kaedah untuk mempertingkatkan nilai sentimen yang diekstrak daripada teks. Pelbagai aplikasi telah dibangunkan dengan mekanisma pemarkahan bagi menghasilkan pengiraan sentimen yang lebih berkesan. Semantic Orientation Calculator (SO-CAL) adalah salah satu alat berasaskan “lexicon” yang dibina dengan unsur-unsur penting seperti “intensifiers”, “negation” untuk mengira kekukuhan sentimen sesuatu teks. Walau bagaimanapun, aplikasi ini mempunyai had dalam memproses perkataan yang salah dieja, terutamanya dalam huruf berulang, ia boleh menurunkan prestasi ketepatan dalam penentuan sentiment. Ketepatan SO-CAL akan berkurang apabila memproses teks daripada media sosial yang kebanyakannya mengandungi perkataan yang salah dieja. Oleh itu, peningkatan mekanisma pemarkahan (LexiPro-SM) telah dibangunkan untuk meningkatkan pengesananan sentimen dalam teks, terutamanya dalam memproses perkataan silap eja yang mempunyai huruf berulang. LexiPro-SM telah diuji dengan mengunakan komen-komen yang diekstrak daripada dua syarikat penerbangam iaitu dinamakan sebagai Airline A dan Airline B. Tiga jenis fasa mempengaruhi pembangunan LexiPro-SM iaitu pengumpulan data, pembersihan data dan analisis data. Dalam fasa pengumpulan data, tiga bulan data telah dikumpul daripada kedua-dua laman rasmi penerbangan di Facebook. Dalam fasa pembersihan data, data yang telah dikumpul akan dibersihkan dengan menghilangkan unsur-unsur yang tidak perlukan dalam analysis ini. Maka selepas pembersihan data, data tersebut hanya mempunyai abjad dan tanda seru. Dalam fasa analisa data, data tersebut akan diproses dengan mengunakan LexiPro-SM, dimana LexiPro-SM dibina dengan fungsi baru untuk memproses perkataan salah eja dan juga telah meningkatkan fungsi lain yang sedia ada seperti “intensifier” dan “negation”. Data yang bersih tersebut akan juga diprosess dengan mengunakan SO-CAL untuk tujuan perbandingan antara kedua-dua makanisma sentiment. Selain itu, laman portal telah dibina, untuk

University of Malaya

menganalisis keputusan LexiPro-SM (dalam bentuk graf) antara kedua-dua syarikat penerbangan. Graf sub-kategori adalah idea yang baru dalam penyelidikan ini dimana markah keseluruhan bagi perkhidmatan penerbangan telah dibahagikan kepada sub-kategori ia itu perkhidmatan pelangan, harga, pra-penerbangan dan kemudahan. Perlaksanaan sub-kategori ini akan menolong syarikat penerbangan untuk menility dalam perkhidmatan yang diberi kepada pelangganya. Selain daripada itu, portal ini juga akan dikaitkan dengan rekod analisa antara kedua-dua syarikat penerbangan, untuk tujuan membuat perbandingan. Dengan mengunakan keputusan analisa LexiPro-SM, satu kajian kes telah diadakan dimana , keputusan kajian tersebut menunjukan Airline A telah menerima lebih komen positif daripada Airline B, ia membuktikan bahawa Airline A telah memberi perkhidmatan yang lebih baik daripada Airline B. Selain daripada itu, keberkesanan LexiPro-SM telah diuji dengan membuat pengiraan keberkesanan terhadap LexiPro dan SO-CAL, ia menunjukan LexiPro-SM lebih effektif dan telah mencapai ketepatan yang tertinggi (90.7%) daripada SO-CAL (58.33%). Ini menyimpulkan, fungsi memproses perkataan salah eja adalah penting dalam media sosial, ia dapat membantu untuk mempertingkatkan kualiti analisa sentiment.

University of Malaya

ACKNOWLEDGEMENTS

I would like to express my highest gratitude to my supervisor, Dr. Vimala Balakrishnan for her continuous guidance and support to help me in completing this thesis. Then, I am hugely indebted to my senior, Ms Wandeep Kaur whose wonderful collaboration supported me greatly and provided the right direction to complete my dissertation successfully. I truly appreciate her kindness and efforts. Besides, I would also like to thank my family members and friends who are always there for me and for their wise counsel and sympathetic ear. I am blessed to have them as my constant source of motivation.

University of Malaya

TABLE OF CONTENTS

University of Malaya

University of Malaya

University of Malaya

University of Malaya

University of Malaya

LIST OF FIGURES

University of Malaya

University of Malaya

LIST OF TABLES

University of Malaya

LIST OF ABBREVIATIONS

API	:	Application programming interface
DFD	:	Data flow diagram
ERD	:	Entity-relationship diagram
GPL	:	General Public License
HTML	:	Hypertext Markup Language
LIWC	:	Linguistic Inquiry and Word Count
ML	:	Machine Learning
NLP	:	Natural Language Processing
PMI	:	Pointwise mutual information
POS	:	Part-Of-Speech
RAD	:	Rapid Application Development
SA	:	Sentiment Analysis
SDKs	:	Software development kits
SDLC	:	System Development Life Cycle
SHC-pt	:	SentiHealth-Cancer
SO	:	Sentiment Orientation
SO-CAL	:	Semantic Orientation Calculator
SVM	:	Support Vector Machine
SWN	:	Sentiwordnet
TC	:	Term counting
TCAvg	:	Term counting average
URL	:	Uniform Resource Locator
WWW	:	World Wide Web

University of Malaya

CHAPTER 1: INTRODUCTION

1.1 Introduction

Sentiment analysis, also known as opinion mining is a major study that plays a significant role in information processing technology. The purpose of this study is to analyze opinions of an entity, focusing on mining factual information from opinion elements such as emotions, sentiments, evaluations, attitudes and appraisals (Liu, 2012). The sentiment analysis technique has evolved to related tasks such as emotion detection, opinion spam detection, mood detection, or subjectivity analysis (Serrano-Guerrero et al., 2015; Peng et al., 2014; Roberts et al, 2012).

Dave et al. (2003) described opinion mining as a tool to process a set of results, generating attributes and aggregating opinions. The term "sentiment analysis" was first coined by Nasukawa et al. (2003), who described methods to capture sentiment using natural language processing. However, the research into this field has been initiated since year 2000 (Liu, 2012).

Humans are able to make better decisions when they are better informed. For example, seeking opinions of others with experience in a specific entity, is one of the ways decisions are made. In our current world, businesses and organizations, being the major entities that require public opinion for their product or service improvement conduct activities such as surveys, opinion polls and interviews, often find these methods time and cost consuming (Liu, 2016). However, these negative impacts have been minimized by applying sentiment analysis technique towards the business and organization entities (Ziora, 2016; Tarlekar & Kodmelwar, 2015).

Moreover, now with the tremendous growth of social media and internet usage, large amounts of information can be easily accessed at any given time. As information is readily available, information overload poses a problem. Thus, automated sentiment analysis tools are needed to analyze opinions to promote better decision making.

University of Malaya

1.1.1 Opinion

The definition of opinion can be defined in mathematical form as quadruple (*g, s, h, t*) where *g* represents target opinion, *s* is the sentiment identified in target opinion, *h* is the opinion holder and *t* is the time. This formula shows the relation between sentiments (positive, negative and neutral), entity and time of the opinion expressed (Liu, 2012). Entity is also defined in a similar way, where an entity (such as service, product, issue, topic, organization, person, event or business) has hierarchy with parts and each part has its attributes (Liu, 2012). The mathematical form for entity is *e: (T, W)*, where *e* is entity, *T* is hierarchy and *W* is attributes of entity.

Opinion can be classified into regular and comparative opinions (Liu, 2012). According to Jindal and Bing (2006a; 2006b), regular opinions can be further divided into two sub-groups; direct and indirect opinion. Direct opinion is the idea of an entity (or an aspect of an entity), whereas indirect opinion is the opinion of an entity (or an aspect of an entity) that interrelated to the effects on other entities. Example of direct and indirect opinion sentences are:

Direct opinion sentence - "*Diuretics drug is very good*"

Indirect opinion sentence - "*After taking Diuretics drug, my blood pressure rises*"

From the above example, it shows direct opinion generally states that Diuretics drug as good, however in indirect opinion it states the Diuretics drug as good, by describing the good effects of the drug.

Comparative opinion on the other hand is based on comparison made between entities by looking through on their common aspects or features. For example:

Comparative opinion sentence - "*Pepsi tastes better than Coke*"

Above example shows comparison made between the tastes of the both carbonated drinks. The word "*taste*" is the entity of this sentence.

University of Malaya

Apart from opinion classification, two concepts that provide significant meanings to opinions are subjectivity and emotion. Subjective sentence will be expressed in many forms such as desires, suspicions, opinions and allegation (Riloff et al., 2006; Wiebe & Janyce, 2000). Example of subjective:

Subjective sentence - *"I like Asian food"*

In this example, the word "*like*" expressed desire towards the Asian food. Emotion is mainly related to sentiment and produce different level of intensity that would determine the strength of a sentence (Liu, 2012). Example of subjective and emotion sentence:

Emotion sentence - "*This is the best food I have tasted in Malaysia"*

In this example, the word "*best*" gives strength for the sentence, by describing the taste of food in Malaysia.

1.1.2 Level of sentiment analysis

In general, sentiment analysis can be divided into three levels; document level, sentence level and entity/aspect level.

Document level is to express sentiment as positive or negative for an entire document and it is only applicable on a single entity or aspect (Pang et al., 2002; Turney, 2002). For example, a document describes many aspects such as reviews of two or more products; it is not applicable in this level, because document level is only allowed to have review for one or single product.

Sentence level is to determine sentiment as positive, negative or neutral (no opinion) for each sentence (Liu, 2012). This will be suitable for the document that describes many aspects. Moreover, this level could identify the intensity of sentiment for each sentence.

Entity/aspect level formerly has been described as feature level (feature-based opinion mining and summarization) (Hu & Liu, 2004). This level could perform fine-

University of Malaya

grained analysis and can identify the actual meaning or idea of a sentence rather than identifying the number of sentiments present in a sentence. For example:

"*Your restaurant looks clean but the service is very bad*"

This sentence belongs to two aspects which are environment and customer service. So, the unstructured sentence will turn into structured data (based on aspects present in sentence) and determine the sentiment for each of the data (Liu, 2012; Bongiwar, 2015; Mary & Arockiam, 2016).

1.1.3 Sentiment analysis techniques

There are two techniques that involve in sentiment analysis: machine learning and lexicon based.

Machine learning is closely related to artificial intelligence, which is dealing with algorithms that allow computers to determine sentiment in a sentence or document (Schrauwen, 2010; Vohra & Teraiya, 2013; Shelke et al., 2016). This technique is more effective compared to lexicon based, but it is unable to produce quality result for the sentiment analysis that involves different domains (Aue & Gamon, 2005; Muhammad et al., 2015). Generally, machine learning can be categorized into two groups: supervised and unsupervised learning.

Supervised learning is using two sets of documents: training and test set. Training set is used to learn the differentiation in documents characteristic by automatic classifier and test set is to analyze the performance of the classifier (Vohra & Teraiya, 2013). Supervised learning is a successful method in traditional classification in machine learning. The most known and effective algorithms are Naïve-Bayes and Support Vector Machine (SVM). However, the major problem of this supervised learning is it may become less effective with the presence of difference quantity and non-quality training data (Arora et al., 2015; Vohra & Teraiya, 2013).

University of Malaya

Unsupervised learning is a method that involves textual classification, where the classification based on fixed syntactic pattern (composed of part-of-speech POS tags) to express opinion (Soni & Patel, 2014). For example: in a sentence the same word can become noun, verb or adjective, for instance the word "book" (noun: "the book on chair" or verb: "to book a room"). Here POS tags are crucial to determine for which sentiment the word "book" belongs to. The most used standard POS tags were Standard Penn Treebank POS tags (Liu, 2012). However, the disadvantage of this method is, the fully unsupervised model will produce incoherent results due to absence of training data that caused difference between analysis objective and human judgments (Arora et al., 2015).

There were many studies conducted on machine learning technique. For example: sentiment analysis on movie reviews (Narendra et al., 2016), intensified sentiment analysis of customer product reviews using acoustic and textual features (Govindaraj & Gopalakrishnan, 2016), applying machine learning to text mining with Amazon S3 and RapidMiner (Wunnava , 2015), and Arabic sentiment analysis using supervised classification (Duwairi, 2014).

Lexicon based approach is mainly using lexicon to perform sentiment analysis on text which calculates scores based on semantic orientation of words or phrases (Turney, 2002; Liu, 2016; Serrano-Guerrero et al., 2015). The word or phrase bearing sentiment context will form a sentiment lexicon (Liu, 2012). For better understanding, lexicon is a text that may belong to positive or negative sentiment. This approach can be divided into two methods; dictionary based and corpus based (Serrano-Guerrero et al., 2015).

Dictionary based method identifies opinion seed word and search for the synonym and antonym in dictionary. One of the examples is WordNet dictionary, which is used to develop SentiWordNet (Aurangzeb & Baharum, 2011). The main disadvantage of this method is, it rigidness to adapt to a domain or specific context (Martín-Valdivia et al., 2014 ; Shelke et al, 2012). This is because, dictionary based is contain seed word that

University of Malaya

belongs to general meaning which may cause poor performance in domain specific. For example: the word “cheap” generally it contributes negative sentiment, but when comes to domain specific like movie ticketing the word “cheap” contributes positive sentiment.

In the corpus based method, a dictionary related to a specific domain is created with a list of seed opinions and searches related opinion word using statistical or semantic techniques (Shelke et al., 2016). One of the examples of this method is Latent Semantic Analysis (LSA) (Serrano-Guerrero et al., 2015).

Broadly, the lexicon-based approach would be more stable among different domains and suitable to analyze social media text which is diverse in domain and context (Muhammad et al., 2015).

There were many studies conducted on lexicon based approach such as lexicon-based sentiment analysis of teachers’ evaluation (Rajput et al., 2016), lexicon-based sentiment analysis for reviews of products in Brazilian Portuguese (Avanco & Nunes, 2014), a lexicon-based approach for hate speech detection (Gitari et al., 2015), and lexicon-based sentiment analysis of Arabic tweets (Al-Ayyoub et al., 2015). The main technique used in this research is lexicon-based approach. Thus, in chapter two the overall discussion will be mainly focused on lexicon-based approach. The Figure 1.1 shows sentiment analysis techniques.

University of Malaya

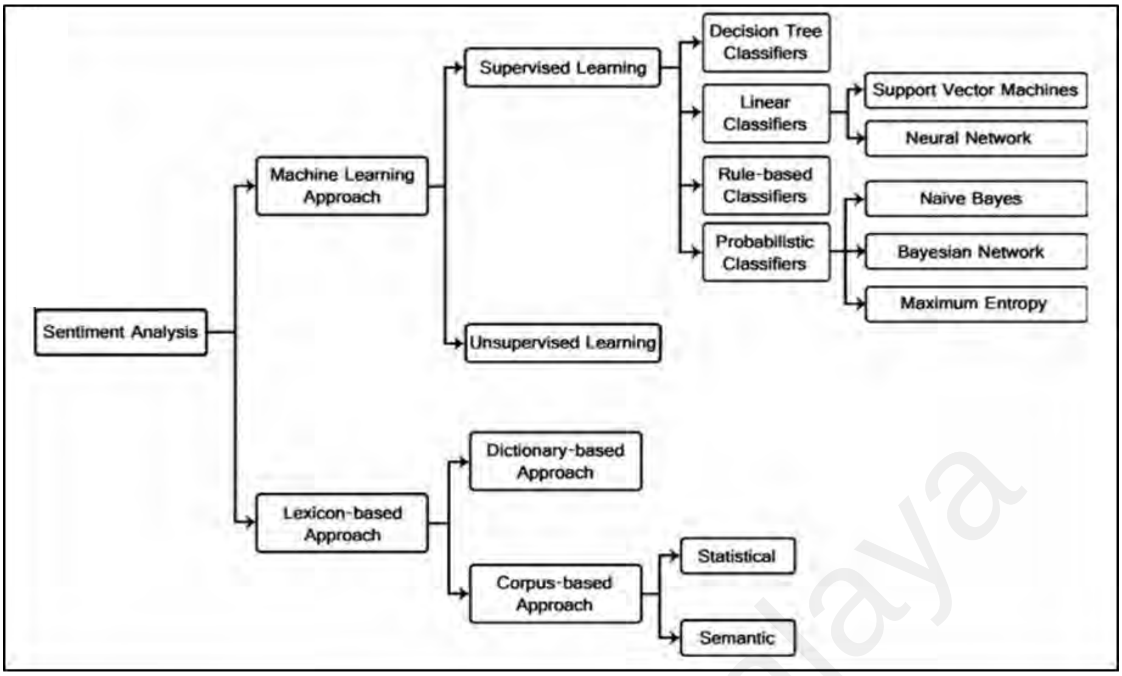


Figure 1.1 Sentiment analysis techniques (Medhat et al., 2014)

1.2 Existing lexicon based sentiment analysis tool and its limitation.

In this section, a brief explanation will be given for the existing lexicon based sentiment analysis tools that are related to this research. Then, the main limitation identified for each tool is described, which would be helpful to determine the problem statement for this research.

1.2.1 SentiHealth-Cancer (SHC-pt)

This SentiHealth-Cancer (SHC-pt) tool was developed to analyze Portuguese text post from online cancer communities, which can help to improve the mood detection of cancer patients in Brazil. This tool is context-based that uses specific information of cancer in order to improve the accuracy of lexical detection, however this tool is unable to support misspelled word, slang, irony and sarcasm (Rodrigues, 2016).

University of Malaya

1.2.2 Sentiment Strength 2

This tool is the enhancement of the first version of SentiStrength 1.0, where more features have been added to improve the scoring mechanism such as idiom list, negation, intensifier, word correction algorithm and repeated letter. It has been tested among different social web data sets and proven as robust in different domains. Nevertheless, it is only able to analyze short informal social web text which has limitation on length of text (Thelwall et al., 2010; Thelwall et al., 2012).

1.2.3 SmartSA

SmartSA was developed to integrate strategies to capture contextual polarity from local and global text. Hybridize lexicons of SentiWordNet and Genre-specific vocabulary and sentiment were incorporated in this tool. SmartSA was analyzed in different social media and comparison evaluation was performed with the results obtain from this tool. The comparison between existing tools shows that the sentiment classification of this tool has been significantly improved. The limitation of SmartSA is its incompatibility between genres (text possess more than one meaning) that causes ambiguity (Muhammad et al., 2015).

1.2.4 Linguistic Inquiry and Word Count (LIWC)

Linguistic Inquiry and Word Count (LIWC) is widely used by psychologists, sociologists, computer scientists, social media domains, and linguists in a number of researches (Crossley et al., 2016). LIWC has evolved to capture psychological phenomena (conscious and unconscious) that are related to emotional affect, cognition and personal concerns. However, this tool does not incorporate sentiment features such as intensifier and negation (Salas-Zárate et al., 2014). Besides, this tool is not freely available but it is able to process text with the absence of an internet connection (once purchased and installed on a personal desktop).

University of Malaya

1.2.5 Semantic Orientation Calculator (SO-CAL)

Semantic Orientation Calculator (SO-CAL) uses dictionary containing words annotated with semantic orientation and incorporates semantic features such as negation and intensification. Despite its stable performance across different domains, SO-CAL is unable to process misspelled words where the misspelled words will be removed (as a part of data cleaning) before analyzing the document (Taboada et al., 2011).

This is because, misspelled word in social media such as acronym and repeated letters can express sentiment especially the word with repeated letters could emphasis word strength in the form of stress and intonation (Ghorbel & Jacot, 2011; Agarwal et al., 2011). By ignoring this feature in SO-CAL, it caused limitation when processing social media data that mostly present with informal text (misspelled word).

University of Malaya

1.3 Problem Statement

In this research, the main focus will be given on lexicon-based approach. Thus, further discussion of this research is only related to lexicon based approach.

The application of lexicon-based method for sentiment analysis has grown and spread to all contexts and domains such as consumer products, healthcare, services, social and political services (Liu, 2012; Muhammad et al., 2015). Although most research focuses on texts written in English, interest in researching on other languages such as Portuguese (Rodrigues, 2016), Spanish (Martín-Valdivia, 2014), Arab (Al-Kabi et al., 2014) and Chinese (Wu et al., 2014) have also peaked. Moreover, many tools have been developed to analyze the presence of sentiment in document and each tool possesses its own scoring mechanism (lexical algorithm) to calculate the intensity of sentiment. Some tools use polarity score calculation that can identify the strength of a document or sentence (Taboada et al., 2011).

Lexicon and corpus based dictionary have a list of positive and negative lexicons and each lexicon has its own value which has been built by human coders (Taboada et al., 2011). The lexicon value range may differ in different research. For example: Semantria score range is -2 (very negative) to 2 (very positve) and SO-CAL score range is -5 (very negative) to 5 (very positive) (Rodrigues, 2016; Taboada et al., 2011). Lexicon algorithm is a method to analyze sentiment of a text based on the orientation and the occurrences of sentiment (Thelwall et al., 2012).

This analysis method may involve other features such as emoticon, negation and semantic rules (Neviarouskaya et al., 2007; Taboada et al., 2011). Besides, the incorporation of non-sentiment or non-lexicon modifiers such as capitalization and sequence of repeated letters in lexicon algorithm could enhance the scores or sentiment strength of document (Muhammad et al., 2015).

Through observation of the research on the existing tools, not every tool discussed in previous section has incorporated non-lexicon modifier especially "repeated letter or characters". According to Brody and Diakopoulos (2011), the lengthening (repetition of letters in word) is strongly related with sentiment and subjectivity. Based on literature, it is discovered that there are two tools that have applied this feature in their own scoring mechanism albeit not exactly defining the strength of repeated letter or characters. This is because both tools are only giving a general score for the detection of repeated character.

Below are the two tools that have incorporated repeated letter feature in their scoring mechanism:

a) SentiStrength 2

In this tool, the scoring mechanism assigning additional score as 1 for the word that has two or more repeated letters (Thelwall et al., 2012). For example, if the score of "happy" word is 2. These words "*happpy*" or "*haaaaapppppy*" will be denoting the same score value as 3.

University of Malaya

Below method shows a clear picture of the score calculation:

Original score for "*happy*" =2.

Score for repetition letter = 1.

Sentence without repetition letter:

"*I am happy with your service*",

Total score= 2.

Sentence with repetition letter:

i) "*I am happppy with your service*"

Total score = 2 (original score of happy word) +1 (score for repetition letter) =3.

ii) "*I am haaaappppy with your service*"

Total score = 2 (original score of happy word) +1 (score for repetition letter) =3.

b) SmartSA

In this tool, a sequence of repeated letters is considered as intensification for sentiment detection. Thus, the same intensification value for the word "*very*" will be assigned for the misspelled word that has repeated letters (Muhammad et al., 2015).

For example, if the score of "happy" word is 2 and intensification percentage for "*very*" word is 25%. These words "*happpy*" or "*haaaaapppppy*" will be denoting the same score value as 2.5. Below method shows a clear picture of score calculation:

Original score for "*happy*" =2.

Percentage for repetition letter = 25%.

Score for repetition letter= 2 (original score of happy word)* 25%= 0.5.

Sentence with repetition letter:

i) "*I am happppy with your service*"

Total Score = 2 (original score of happy word) +0.5 (score for repetition letter) =2.5.

ii) "*I am haaaappppy with your service*"

Total Score = 2 (original score of happy word) +0.5 (score for repetition letter) =2.5.

University of Malaya

The general value given for repeated letters by not considering the length of repeated letter, does not define the strength of the word. This is because, Brody and Diakopoulos (2011) have explained that the length of repetition letters in a misspelled word has the capability to emphasis sentiment's strength, which means the more repeated letters present in a word, the higher the sentiment value for a word. However, there is no tool developed with the incorporation of sentiment score calculation for length of (number of) repeated letters.

1.4 Research Objective

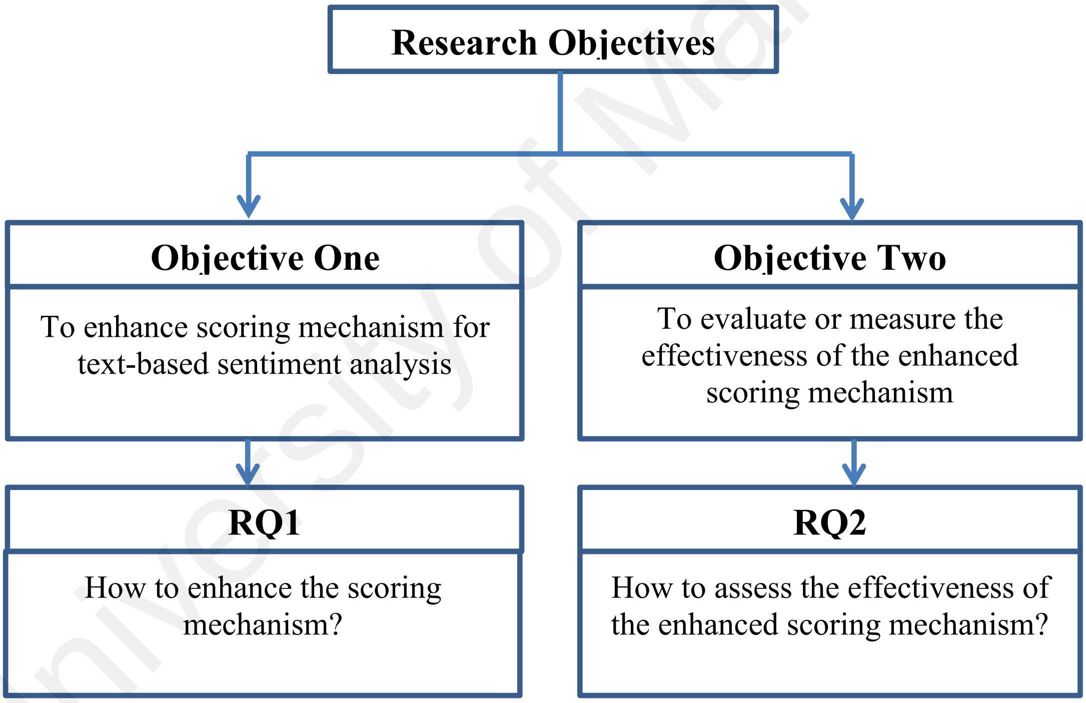

Research Objectives
Objective One
To enhance scoring mechanism for text-based sentiment analysis
Objective Two
To evaluate or measure the effectiveness of the enhanced scoring mechanism
RQ1
How to enhance the scoring mechanism?
RQ2
How to assess the effectiveness of the enhanced scoring mechanism?

Figure 1.2 Research Objective

Figure 1.2 shows the research objectives and research questions that identified from the problem statement discussed in previous section.

First objective: To enhance scoring mechanism for text based sentiment analysis

The main goal of this research is to enhance scoring mechanism for text-based sentiment analysis.

University of Malaya

Although there are many existing sentiment analysis tools that able to produce an effective result, still many have their own limitations that could affect the accuracy of the sentiment detection. Thus, this research could be a part of improvement on the accuracy of sentiment detection in lexicon-based sentiment analysis.

Second objective: To evaluate or measure the effectiveness of the enhanced scoring mechanism

Besides, the second objective of this research is to evaluate or measure the effectiveness of the enhanced scoring mechanism by making comparison with an existing sentiment analysis system. The enhanced scoring mechanism will be addressed as LexiPro Scoring Mechanism (LexiPro-SM), this would be helpful to differentiate among other existing systems throughout this research.

In this research, SO-CAL plays an important role to develop the proposed scoring mechanism. The main feature such as dictionaries, score ranges value and intensification has been taken from SO-CAL to build the LexiPro-SM. Additionally, SO-CAL tool which is freely available on internet will be used as a reference tool to make comparison evaluation with the results obtained from the LexiPro-SM.

1.4.1 Research questions

(a) ***RQ1. How to enhance the scoring mechanism?***

SO-CAL will be used as a reference tool for this research. So there were ideas taken from SO-CAL scoring mechanism to develop the LexiPro-SM. It could be a part of improvement for SO-CAL scoring mechanism. Many features are incorporated in scoring mechanism such as lexicon dictionary, algorithm, negation, intensification and non-lexicon modifiers (capitalization, repeated letters and emoticon). The alteration of these features could improve the scoring mechanism performance. The main feature will be improved on LexiPro-SM is “repeated letter” which will be elaborated in research contribution section. Besides, the LexiPro-SM has also will be

University of Malaya

incorporated lexicon and non-lexicon modifiers such as negation, intensification and capitalization.

(b) ***RQ2. How to access the effectiveness of the enhanced scoring mechanism?***

Evaluation is one of the important parts to show either the proposed objective has been achieved or not. There are many strategies to perform evaluation. For example: analyzing system with human coded data or any existing trained data. For this research the evaluation of LexiPro-SM and SO-CAL results will be analyzed by using human expert results (will be discussed further in chapter three).

1.5 Project scope

Social media has become an important means of communication in the human daily life (Beigi et. al., 2016). The explosive growth of social media, allows everyone to perform many activities such as editing, posting, sharing and manipulating content (Beigi et. al., 2016). Such activities contribute to an increased amount of data that can be accessed easily by the end user. As mentioned previously, the main technique used in this research is lexicon-based approach, which is suitable to analyze social media informal text that is divers in domains and context (Muhammad et al., 2015). Thus, for this research, Facebook will be used as the source of social media to extract data and perform the research analysis.

Furthermore, the context of this research focuses on the Malaysian Airline Industry. This is because airline services in Malaysia at large involve both local and foreign travelers (Tand & Yap, 2015). This leads to an increase in the number of comments on the airline official Facebook page (Socialbakers, 2012). Therefore, it would be favorable to collect sufficient data for this analysis. Besides, by doing this research, a case study of airline industries would be conducted between two major airlines in Malaysia. In this research, the two major airlines will be referred as Airline A and Airline B respectively due to information confidentiality purpose.

University of Malaya

The main competitive factor between these two airlines is the price factor. Despite both airlines providing different service experiences, there is minimal difference when it comes to domestic flights or international flights to destinations close to Malaysia (Kee & Ghazali, 2011).

By conducting this sentiment analysis research, it helps to measure customer satisfaction level for both airlines. Moreover, comparisons can be made between these airlines based on the statistical results obtained from the enhanced scoring mechanism.

Besides, the improvement of scoring mechanism not only applied for the overall service, it has been divided into sub-services Customer Service, Price, Preflight and Facility that could be helpful to identify and improve a particular service in the airline industry.

1.6 Research contribution

The research focus will be on improving score calculation for one of the non-lexical modifiers: which is known as repetition of letters, where a lexicon will be identified in typo word that has repetition of letters and will be producing scores based on the length (number) of repeated letters contains in typo word. Besides, the scores improvement also will be applied on other modifiers/features such as intensifier, negation, exclamation mark and auto-word correction mechanism. This scores improvement would facilitate to increase the degree of sentiment (positive or negative) in a document or sentence. The details of the enhancement scoring mechanism are presented in Chapter 3.

Although the basis of the scoring mechanism is referenced to SO-CAL, the calculation of scores has been built by integrating own perspective, which is believed may be helpful to further improve the scoring mechanism to produce an effective result.

University of Malaya

1.7 Dissertation layout

This dissertation consists of six chapters, and they are organized as per below:

i. **Chapter 1: Introduction**

This chapter is serves as an introductory for sentiment analysis (or opinion mining) technique. The importance and other information of this technique has been clearly explained. Besides, it presents the problem statement, objective and project scope-contribution that has been identified for this research.

ii. **Chapter 2: Literature review**

This chapter discusses the review of literature on the relevant concept of research. The discussion mainly focuses on lexicon based sentiment analysis. A general study has been carried out towards the similar existing tool or related research.

iii. **Chapter3: Research methodology**

This chapter provides the detail overview of the planning to perform the research. It discusses the different stages involved in this research. This includes design of the data collection, data cleaning, and data analysis for the scoring mechanism.

iv. **Chapter 4: Implementation and testing**

This chapter discusses the implementation of all plans that have done during the designing phase, which involve programming works to build the scoring mechanism and perform testing to ensure the functionality of the mechanism works accordingly. During this phase the collected data will be analyzed in enhanced scoring mechanism and SO-CAL tool.

University of Malaya

v. **Chapter 5: Result and discussion**

This chapter will discuss the evaluation of the enhanced scoring mechanism by performing comparison between enhanced scoring mechanism and existing tool SO-CAL results, which may involve accuracy and precision testing.

vi. **Chapter 6: Conclusion, limitation and future work.**

This chapter will conclude the research; highlight the limitations and future study that would be helpful to improve the current enhanced scoring mechanism.

University of Malaya

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

In this chapter, the discussion will focus on lexicon-based sentiment analysis, as this is the main technique used to develop LexiPro-SM scoring mechanism. Thus, the discussion comprises of problems faced in natural language processing (NLP), an explanation about sentiment lexicon, approaches in lexicon based technology, lexicon modifiers and finally an explanation and review on the existing systems (lexicon approaches) that relate to this research.

2.2 Natural language processing (NLP) problems

Natural language processing (NLP) is a field of computational linguistics concerned with interaction between computers and humans. However, such an interaction can be a challenge especially when it comes to accuracy in sentiment detection. One of the major problems in (NLP), is the level of ambiguity and complexity of semantic information (Phillips, 1999).

Level of ambiguity refers to a sentence that has two meanings (either the sentence belongs to sentiment or not). For example: the sentence "*The old man and the boats*". The word "old" can be used as a noun (countable) as well as an adjective (negative meaning). "Man" can be also be used as a noun as well as a verb (meaning taking charge of).

Complexity of semantic information occurs when the sentence has different semantic values (positive and negative). For example: "*the colorless green car*", the word "*colorless*" denotes a negative sentiment and the word "*green*" represents a positive sentiment. This sort of conflicting sentiment in a sentence makes it all the more difficult to provide an accurate sentiment reading.

University of Malaya

In order to address this concern, many NLP tools have been continuing to be developed. Sentiment analysis is a part of NLP. There are many sentiment analysis tools available in market, however each tool features are different from one another. So, each sentiment analysis tool has its own advantages and limitations.

2.3 Sentiment lexicon

Sentiment lexicon is a lexical resource for sentiment analysis which has a collection of lexical units (in the form of database) with semantic orientation (Ahire, 2014; Heerschop et al., 2011). Sentiment lexicon can be defined as a set of tuples in a form of lexical unit or sentiment. Lexical unit can be in a category of word, phrases, word senses and so on. Sentiment can be determined in several forms such as (Ahire, 2014):

i) General categorization of either positive, negative or neutral (no sentiment).

ii) A categorization between positive or negative value. For example: strongly positive, mildly positive, neutral, mildly negative, and strongly negative.

iii) Assigned a strength value such as (-2, -1, +1, +2). This will be useful to determine the strength of a document by calculating the average value of the total scores of sentiment present in a document.

In most lexicon-based analysis, the lexicon unit will be in the form of word which has specific strength value (Heerschop et al., 2011). These words are Part of Speech (POS) that can be categorized into adjectives, verbs, nouns and adverbs.

At earlier stage in sentiment analysis research, the sentiment detection only focused on the presence of adjectives or adjective phrases in a sentence or document (Hu & Liu 2004; Taboada et al., 2006; Hatzivassiloglou & McKeown 1997). Further studies found that besides adjectives, other POS words could increase the semantic polarity (Polanyi & Zaenen, 2006). More studies were conducted that began to merge different forms of grammar such as adjectives were merged with adverbs (Benamara et al., 2007), verbs

University of Malaya

and adjectives (Kim & Hovy, 2004), adjectives, verbs, and adverbs (Subrahmanian & Reforgiato, 2008) in order to improve sentiment detection.

2.4 Approaches in lexicon creation

There are two methods used to create a lexicon: manual creation and automated creation.

2.4.1 Manual lexicon creation

Manual lexicon creation is a human effort. The lexicon will be identified by analyzing the structure of the word's sentiments (Ahire, 2014). Generally, manual lexicon creation can be divided into two stages (Kotelnikov et al., 2016):

1) Generate sentiment-bearing words list

2) Assigning sentiment label (positive/negative/neutral) for the identified word.

Generation of sentiment bearing words list is actually to create a collection of words that have the capability to bear sentiment. It mostly involves adjective and including other POS words that can give sentiment based on the occurrence of the word in a sentence. Second, sentiment label such as negative, positive or neutral will be given for each sentiment bearing word in the word list.

The list of words can be obtained from any dictionary or corpus. Involvement of multiple annotators to identify the sentiment label would increase the robustness of manual lexicon creation. However, there is no any complexity involved such as computational or algorithmic works. The correctness of this approach is guaranteed because it involves human judgment to create the lexicon (Ahire, 2014). However, the lexicon creation requires long timeline to be built (Kotelnikov et al., 2016). There are many research that have applied this approach such as Taboada et al. (2011) who built the lexicon manually by involving both stages (as per above), Mohammad and Turney (2013) used crowdsourcing method to create the list of word-emotion and word-polarity

University of Malaya

while Amiri et al. (2015) created a Persian word list manually before it was annotated by several human annotators with the aid of web interface.

2.4.2 Automated lexicon creation

The automated lexicon creation involves huge lexical creation with less human effort (Oliveira et. al, 2016). There are many techniques to create the automated lexicon, however one of the most common techniques is creating a set of seed words with the sentiment orientation and expand the seed set by using an existing lexical resource (Ahire, 2014). Another way of creating automated lexicon is by using bootstrapping method that creates lexicon list based on the patterns from corpus without the use of lexicon (Banea et al., 2016). The advantage of this approach is its ability to produce a huge collection of lexicon in a minimal timeline (Ahire, 2014). Nevertheless, its inefficient sentiment label algorithm may cause less accuracy on assigning sentiment label for lexicon (Ahire, 2014).

There are several studies conducted based on this approach such as Baccianella et al. (2010) created a lexicon automatically (SentiWordNet) by assigning scores to all WordNet synsets, Turney and Littman (2003) created a lexicon by assigning positive and negative semantic scores for each word from Pointwise mutual information (PMI), and Hu and Liu (2004) extracted opinion words from large corpus of customer reviews that related to product features.

2.5 Lexicon modifiers

Lexicon modifiers are features that are responsible for adjusting the sentiment strength of a lexicon for a specific context (Muhammad et al., 2015). Generally, lexicon modifiers can be divided into two categories: Lexical valence shifters and non-lexical valence (Muhammad et al., 2015).

University of Malaya

2.5.1 Lexical valence shifters

Lexical valence shifters are the dictionary recognizable words that have capability to increase, decrease polarity of sentiment of specific context. The main types of lexical valence shifters are intensifiers and negation. In this section there are sub-sections that describes about intensifiers and negation.

2.5.1.1 Intensifiers

The characteristic of intensifiers is to increase or decrease sentiment. According to Quirk et al. (1985) intensifiers can be classified into two types: amplifier and downtoner. Amplifier is to increase sentiment such as "*very*", "*extremely*", and "*highly*". Downtoner is to decrease sentiment such as "*slightly*" and "*somewhat*". The calculation of polarity for intensified context will vary in different research.

Generally, most research assign general values for intensifiers (Kennedy et al., 2006; Polanyi et al., 2006). However, several studies have modeled intensification by assigning each intensifying word with a different percentage value (Taboada et al., 2011; Muhammad et al., 2015). Examples of intensifier sentences are:

Amplifier sentence - "*I am very happy with the results*"

Downtoners sentence - "*I am slightly disappointed with your performance*"

2.5.1.2 Negation

Negation is an important linguistic component that has the ability to change a text polarity (Dadvar et al., 2011). The characteristic of negation is to reverse the polarity of a sentiment which is also known as switch negation (Sauri, 2008). According to Muhammad et al. (2015), negation can be considered as a diminisher due to its ability to reduce the sentiment polarity. Among the examples of negation words are "no, not, nobody, none, nothing and never". The presence of negation not only changes the sentiment of the neighborhood (adjective/adverb) word, but also the sentiment of the

University of Malaya

whole sentence by looking into certain verbs, past determiners and copulas (Taboada et al., 2011). For example:

Normal negation sentence: "*I am not happy with your statement*" (not negated happy)

Implicit negation sentence: "*Nobody gives a good review of this movie*". (nobody negated good)

2.5.2 Non-lexical valence shifters

Non-lexical valence or non-lexical modifiers are features that only appear in informal text communication. Social media is one of the main sources where this form of negation is abundantly available. These modifiers can be identified in the form of a sequence of repeating letters, capitalization and presence of emoticons (Muhammad et al., 2015). In this section there are sub-sections that describes about capitalization, repeated letters and emoticons.

2.5.2.1 Capitalization

The main purpose of capitalization in informal text communication is to emphasis sentiment or to express emotion. According to Korkontzelosa et al. (2016), capital letters present in informal communication indicates anger that stresses the significance of the content. There are also studies that show capitalization acting as an intensifier that increases the strength of the sentiment (Taboada et al., 2011; Muhammad et al., 2015; Paltogloua et al., 2010). However, in informal text communication not all the capitalization would emphasis sentiment. Some texts may belong to abbreviation or acronyms (Muhammad et al., 2015). Examples of capitalization that emphasis sentiment are:

Capitalization positive sentence – "*Your performance was GREAT*"

Capitalization negative sentence – "*I DON'T LIKE YOUR CUSTOMER SERVICE*"

University of Malaya

2.5.2.2 Repeated letters

Repeating letters is another way to emphasize an emotion or sentiment in social media (Brody & Diakopoulos, 2011; Muhammad et al., 2015; Paltogloua et al., 2010). According to Ghorbel and Jacot (2011), misspelled word that has repetition of letters, can express a kind of stress and intonation. Several studies have integrated this feature and assigned a general value to intensify the sentiment polarity of a phrase or sentence (Brody & Diakopoulos, 2011; Muhammad et al., 2015; Thelwall et al. 2012). The repeated letter phenomena mostly occurred in POS words. Examples of repeated letter sentence are:

Repeated letter positive sentence – "*I am happpppppyyyyyyy*"

Repeated letter negative sentence – "*I am dissssappointed*"

For the positive sentence the word "*happy*" has many repeated characters as "*happpppppyyyyyyy*", this indicate that the sentiment value of "*happy*" word was intensified and the positive scores will be higher than the original value. Same condition applied for negative sentence, where the negative scores of the repeated letter word will be higher than the original value.

2.5.2.3 Emoticons

Emoticons are also used to express emotions especially in online media (Wang et al., 2015). Emoticon is present in a sequence of typographical symbols that resemble facial expression (Hogenboom et al., 2013). For example: ":-(" (a sad face), ":-)" (a happy face) , ":'(" (a crying face) and many more. According to Wang et al. (2015), different people have different opinions towards the emoticon sentiment. For example, emoticons that represents states of being annoyed and uneasy could indicate negative sentiment to some people, but it may indicate no sentiment (neutral) for others. So a generalized sentiment value should be determined for each emoticon which may help to perform sentiment polarity calculation for a specific context. Many studies have been

University of Malaya

conducted in sentiment analysis that show the importance of emoticons (Hogenboom et al., 2013; Elgamal, 2016; Muhammad et al., 2015; Wang et al., 2015).

In addition to the non-lexical modifiers discussed above, there is another feature that plays an equally important role to emphasize a sentiment, which is the exclamation symbol (!). This symbol has the ability to intensify a specific context. Acknowledging the importance of this symbol, most of studies treat the exclamation as an intensifier and the sentiment polarity is calculated accordingly (Taboada et al., 2011; Muhammad et al., 2015; Paltogloua et al., 2010; Thelwall et al. 2012).

2.6 Review on lexicon based approach studies.

In this section, existing lexicon based sentiment analysis tools are discussed. Among the tools discussed are Semantic Orientation Calculator (SO-CAL), SentiStrength 2, Linguistic Inquiry and Word Count (LIWC), SentiHealth-Cancer (SHC-pt) and SmartSA. Besides, a brief explanation will be given for other recent studies that are relevant to this research.

2.6.1 Semantic Orientation Calculator (SO-CAL)

The sentiment orientation calculator (SO-CAL) is a tool to analyze semantic orientation of individual words, developed for the English language. Taboada et al. (2011) developed this tool with the aim to perform in depth analysis toward the semantic orientation of sentiment words and contextual valence shifters. However, this tool was not incorporated with linguistic analysis techniques that responsible to disambiguate meaning of a sentence, for example: POS tag. The main function of this tool is to extract sentiment bearing words such as adjectives, adverbs, nouns and verb and assign semantic scores for each word by considering valence shifters such as negation and intensification. Then a final score will be calculated for the document or sentence.

University of Malaya

Taboada et al. (2011) has created semantic orientation dictionaries which has 5000 lexical words with their orientation. All the dictionaries in SO-CAL were created manually, as Taboada et al. (2011) believed manual creation with human annotation may improve the accuracy of the final results. The score range for the SO-CAL dictionary is between -5(most negative) and +5 (most positive). Zero score denotes as neutral sentiment (no sentiment detected). The dictionaries created contain 2,252 adjective, 1,142 nouns, 903 verbs, and 745 adverbs. In addition, features such as intensification, negation and irrealis blocking were incorporated in SO-CAL as rules to modify the sentiment orientation (SO) scores.

2.6.1.1 Intensification

Taboada et al. (2011) has identified that assigning a general value for the intensification sentiment score calculation does not cover a wide range of intensifiers such as words like "*extraordinarily*", "*tremendously*". Below is the examples show the calculation of sentiment score with general value of intensification.

Intensification General Value = 1

Score for "*happy*" = 2

Amplifier sentence = "*I am very happy*"   Score = 2+1 = 3

Downtoner sentence = "*I am slightly happy*" Score = 2-1 = 1

In SO-CAL, it has a separate dictionary for intensifier which stores all the intensifier words and each word has its own percentage value. For example, the percentage value for "most" is 100% and "sleazy" is -30%. The sentiment score will be calculated based on the percentage value of the intensifier word.

University of Malaya

Below are the examples that shows the calculation of sentiment score for intensification in SO-CAL.

Example 1:

Score for "*excellent*" word =5

Percentage value for "*most*" = 100%

Amplifier sentence = "*most excellent*"

Total sentiment score = 5 × (100% +100%) = 10.

Example 2:

Score for "*sleazy*" word =-3

Percentage value for "*somewhat*" = -30%

Downtoners sentence = "*somewhat sleazy*"

Total sentiment score = −3 × (100% +(− 30%)) = −2.1.

For intensifier calculation, every percentage value of intensifier word will be sum with 100%. This 100% indicating total percentage value of sentiment word (positive or negative word). After the addition, the value will be multiplied with the sentiment scores. This result will give a final score for the intensified sentiment word. Above examples shows the calculation for the positive (example 1) and negative (example 2) intensifiers.

Other than intensifier words, SO-CAL has also incorporated other non-lexical valence such as capitalization and exclamation symbol. As discussed in the literature above, this is significant because it able to emphasis the strength of a phrase or sentence. By the presence of capitalization or exclamation symbol on sentiment bearing word, the total score of sentence / phrases will be multiplied by two.

Example 1 (capitalization):

Normal sentence = "*I love my school*". Original score is 3.

Capitalization sentence = "*I LOVE my school*". Intensified score is 3x2=6.

University of Malaya

Example 2 (exclamation):

Normal sentence = "*I hate reading*". Original score is -3.

Exclamation sentence = "*I hate reading!"* Intensified score is -3x2= -6.

In example 1, it shows the original semantic score for "*love*" is 3. After capitalization the sematic score of "*love*" has intensified to 6. Similarly, for example 2 the original semantic score for "*hate*" is – 3. After the present of exclamation mark the semantic score of "*hate*" has intensified to -6.

2.6.1.2 Negation

Generally, most sentiment analysis tool may apply switch negation, which eventually reverses the polarity value of a sentiment bearing word. However, Taboada et al. (2011) has explained switch negation does not work well for every situation. For example: the phrase "*not excellent*" is partly positive than the phrase "*not good*" and the phrase "*not atrocious*" is more positive than the word "*good*". By considering this, a numerical shift has been applied in SO-CAL to deal with this kind of negation which is known as shift negation. SO-CAL also fixed the negation number to four that will be used for shift negation calculation. Below are the examples of shift negation calculation:

Example 1:

Score for word "*good*" =3

Sentence = "*not good*" = 3-4= -1

Example 2:

Score for word "*excellent*" =5

Sentence = "*not excellent*" = 5-4= 1

In example 1, it shows that the word "*good*" negated by deducting the sentiment score with shift negation score four and the final score turns into -1, which means the sentence has been negated and it turns into negative sentiment for the whole phrase. However, when the same calculation applied in example 2, the final scores does not turn

University of Malaya

into negative value, because the degree level of positive sentiment word (excellent) is higher than the shift negation value. So it shows the sentiment for "not excellent" is actually more positive than the "*not good*" phrase.

Negation can occur in any part of a phrase/sentence; either next to a POS word or non-sentiment words. Negation that occurs next to non-sentiment words usually negates the whole sentence/phrases. This can prove to be a challenge in negation calculation.

Sentence 1: "*Nobody is happy with the changes made*" (nobody negates happy)

Sentence 2: "*None of this is bad*" (none negates bad)

To encounter this situation, SO-CAL uses two options to detect the presence of negation in a phrase or sentence. First, to look backwards until the boundary marker has been hit and second is to look backward until it reaches the negation clauses.

2.6.1.3 Irrealis Blocking

Some sentences/phrases with the sentiment bearing word may become unreliable due to the presence of irrealis markers or also known as conditional markers. SO-CAL ignores unreliable sentence for sentiment score calculation and automatically assigns a sentiment score of zero for the irrealis marker sentence. There are different types of irrealis markers used in SO-CAL such as "*could*", "*would*", "*any*", "*anything*", "*expect*", "*if*", "*doubt*", "*questions*" and "*quoted word or phrases*". Example of irrealis sentences are:

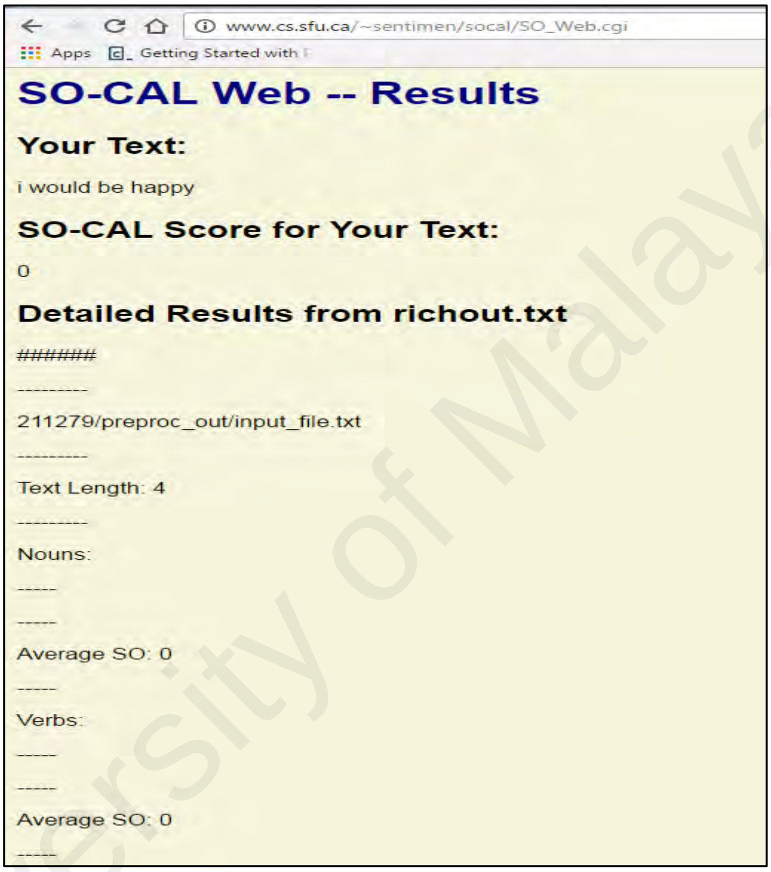
Sentence 1: "*But for youngsters, this program could be one of the best of the festival season*" – (Sentiment score from 3 to 0)

Sentence 2: "*This should have been a great activity*" – (Sentiment Score from 5 to 0)

Besides these features, SO-CAL also applied text-level features that are able to reduce positive bias issues in sentiment detection. According to Kennedy and Inkpen (2006), in some situations the classification of sentiment tends to show as positive (will be called as positive bias) even though the actual meaning belonging to a negative

University of Malaya

sentiment. Thus to overcome this, SO-CAL has incorporated text level features as negation weighting and repetition. Negation weighting increases the negative sentiment scores by 50%, which ultimately reduces inherent positive bias. Repetition is to reduce repetition of same sentiment word in a sentence. The SO-CAL tool is freely available in online (http://www.cs.sfu.ca/~sentimen/socal/).



SO-CAL Web -- Results

Your Text:

i would be happy

SO-CAL Score for Your Text:

0

Detailed Results from richout.txt

```
#######
---------
211279/preproc_out/input_file.txt
---------
Text Length: 4
---------
Nouns:
-----
-----
Average SO: 0
-----
Verbs:
-----
-----
Average SO: 0
-----
```

Figure 2.1: Sample of SO-CAL result page (http://www.cs.sfu.ca/~sentimen/socal/SO_Web.cgi)

Figure 2.1 shows the sample of SO-CAL result page. In this page, it shows the text that was analyzed, overall scores, average scores for all the part of speech words, detection of irrealis marker and lastly scores by sentence. These structures could provide users a good understanding of the analysis result.

University of Malaya

By reviewing SO-CAL tool, it shows a solid function human annotated dictionaries incorporated with this tool that able to provide an effective semantic orientation scores. Besides, this tool is able to produce an overall result without having restriction on the length of document and also able to produce sentence level scores. According to Taboade et al (2011), SO-CAL performance is robust across different domains. However, the major problem of this tool is, it is unable to process typo words, which may become as major threat in analyzing social media text. Besides, due to inability to process typo words, there are no non lexical modifiers (especially repetition of letters) that were incorporated with this tool. SO-CAL also does not incorporate with POS, hence accuracy of classification may be affected.

In this research, SO-CAL will be used as a main reference tool to develop the LexiPro-SM. Moreover, this tool plays an important role in achieving the second objective of this research where an evaluation will be performed by comparing SO-CAL and LexiPro-SM scoring results.

2.6.2 SentiStrength 2

SentiStrength is another lexicon based sentiment analysis tool that is freely available in online (http://sentistrength.wlv.ac.uk/). SentiStrength was initially developed to analyze social media text that was written in English although it was not expanded to include multiple other languages such as Finnish, German, Dutch, Spanish, Italian, Russian and etc. The updated version of this tool is known as SentiStrength2 which incorporated advanced features that improves the accuracy of sentiment detection.

The characteristic of this tool is it give dual sentiment scores (positive and negative) for the sentiment analysis. For example: if the sentiment score is “3, 4”, it means that the phrase has moderate positive sentiment and high negative sentiment. According to Norman et al., (2011), the same text may evoke different sentiments in people. The dictionary used in SentiStrength 2 is made up of 2310 sentiment words which have been

University of Malaya

obtained from Linguistic inquiry and word count (LIWC), the General Inquirer list, and some from other sources during the testing. The polarity scores range for SentiStrength 2 is 1 to 5 for positive sentiment and -1 to -5 for negative sentiment. Although both SentiStrength2 and SO-CAL share many similarities, the former tool is specially designed for social media analysis. Apart from textual dictionary, SentiStrength2 has also included a list of emoticons with human annotated sentiment scores.

In order to improve the accuracy of sentiment detection, SentiStrength2 has incorporated a number of features that can deal with special cases, such as an idiom list, same word with different sentiment, spelling correction algorithm, non-lexical valence modifier- repeated letters, booster word – intensifier, negation, emoticon list and exclamation mark (Thelwall et al., 2012). These features will be further discussed in the following sub-sections.

University of Malaya

2.6.2.1 An Idiom list

An idiom list is a list of phrases that represents word senses that may contain some form of sentiment. For example:

"*Couch potato*" is indicative of a lazy person which may reflect a negative sentiment in sentiment analysis.

2.6.2.2 Same word with different sentiment

Identifying a correct sentiment for words that may belong to more than one sentiment (positive, negative, or neutral). For example, the word "*like*" can belong to a positive sentiment (e.g, "*I like the service*") or can belong to neutral if used as a comparator (e.g, "*It looks like a cat*").

2.6.2.3 Spelling correction algorithm

To detect typos those contain repeated letters and delete the repeated letters. If a word is not found in the English dictionary, it will create the word into the dictionary (e.g., likke --> like).

2.6.2.4 Non-lexical valence modifier- repeated letters

If there are two or more repeated letters in a word, it will boost the sentiment words by 1. For example: if the sentiment score for "*happy*" is 2, the sentiment score for "*happpppyyy*" would be 2+1=3.

2.6.2.5 Booster word – Intensifier

If the phrase contains intensifier word, such as "*very*" (amplifier) or "*somewhat*" (downtoner), the strength of the phrase will be increased/ decreased by 1.

2.6.2.6 Negation

A negation word list used to invert sentiment of the phrase. For example: "I am not sad" – This phrase has a negative bearing word (sad), however the negation word (not) inverts the phrase sentiment to a positive sentiment. Besides, the negation sentiment will be ignored if the phrase belongs to a question.

2.6.2.7 Emoticon list

As discussed above, Sentristrength 2 also incorporated with emoticon list (manually annotated by humans). The emoticon contributes scores as +1 or 1. So, the presence of emoticon in a text may increase the sentiment score.

University of Malaya

2.6.2.8 Exclamation Mark

The minimum positive strength for phrases with exclamation marks is 2, unless it is a negative emotion. Repeated letters with exclamation marks would increase the strength of the phrase immediately by adding 1.

By reviewing this SentiStrength2 tool, it shows that it is well equipped to analyze social media data. However, it displays a limitation when tested against six social web data sets (BBC Forum posts, Digg.com posts, MySpace comments, Runners World forum posts, Twitter posts and YouTube comments) (Thelwall et al., 2012). The performance of this tool was below average due to its incapability to deal with ambiguity in sentences as it is not incorporated with Part of Speech (POS) tagging. This is because social media contains informal text that does not rely on standard grammar, so unambiguous techniques would not help to improve the performance of this tool (Thelwall et al., 2012).

Furthermore, this tool is only able to process short text with a maximum of 100 characters. Hence, this tool would not be suitable for every social media data set especially Facebook posts (possess large size of text). Moreover, some features in this tool, is not fully utilized to extract the sentiment. For example, Booster word/ intensifier has general value as one, however there are many intensifier words that have different level of strength like how it has been explained in SO-CAL tool. Similarly, a general value given for the repeated letters does not define the exact strength of the sentiment. Finally, this tool not featured to detect sarcasm (express negative sentiment with positive word using sarcastic tone) and irony (express negative sentiment with positive word using funny tone) sentiments.

University of Malaya

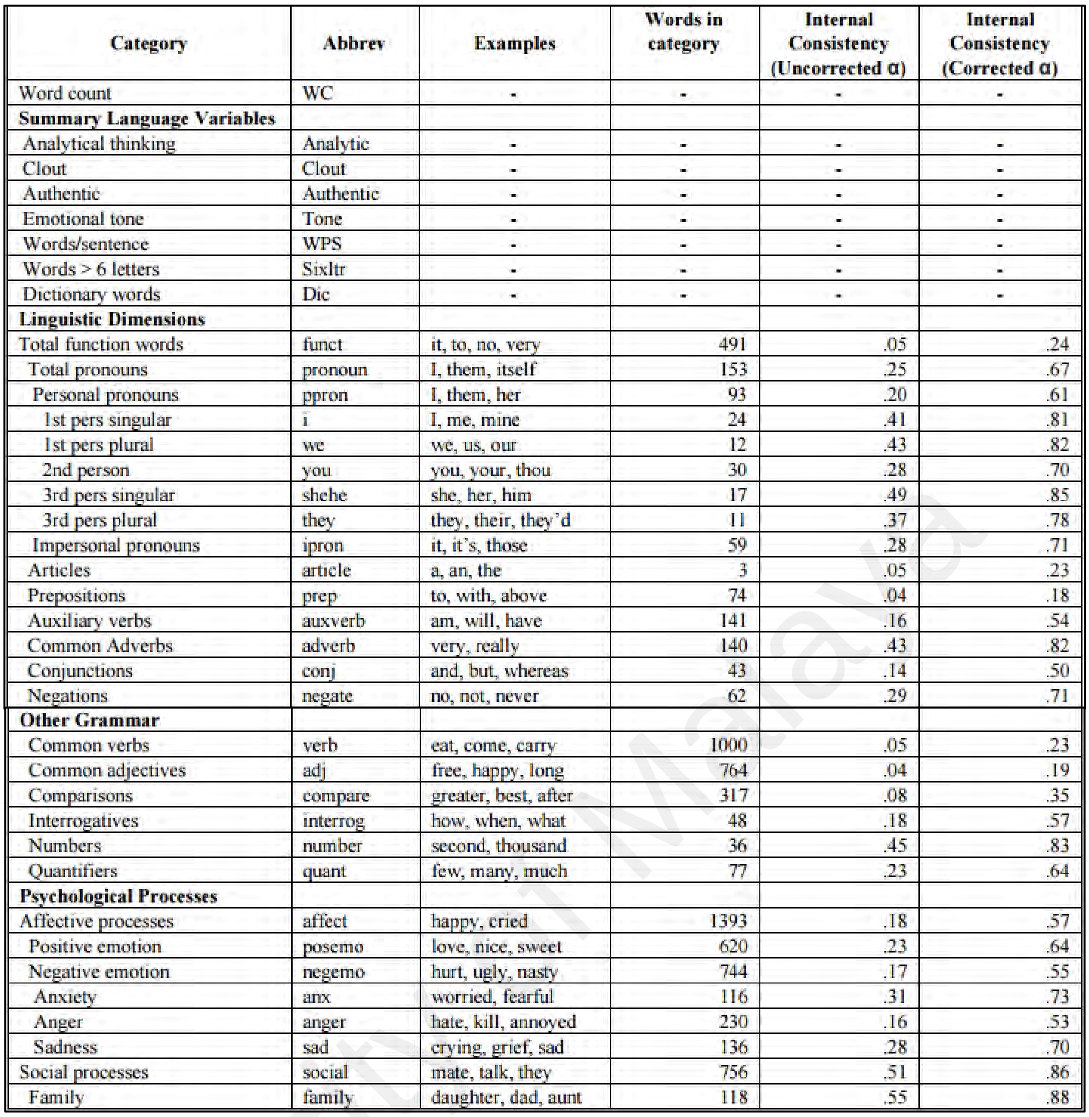
2.6.3 Linguistic inquiry and word count (LIWC)

Linguistic inquiry and word count (LIWC) is a commercialized product that is used for text analysis. LIWC is packaged as a standalone software that is able to work offline (Pennebaker et al., 2015). The main function of this tool is to calculate the degree of use words in pre-established categories defined within LIWC itself and produce a report that shows the percentage of words per text for each category (up to 80 categories). The heart of this tool is the use of LIWC dictionary. Originally, LIWC dictionary was built with English lexicons (Pennebaker et al., 2007), however now it is available in other languages and with independent language dictionaries such as Chinese, Arabic, Spanish, Dutch, French, German, Italian, Russian and Turkish. Besides, the English LIWC dictionary has been used in other studies such as SentiStrength (Thelwall et al, 2010) and Kim et al. (2012), uses LIWC dictionary to classify anonymous texts.

Generally, LIWC is designed to detect conscious (aware thoughts) and unconscious psychological (unaware thoughts) experts within a text and is widely used by psychologist, sociologists, linguists and computer scientists to perform psycho-linguistic analysis (Hutto & Gilbert, 2014; Pennebaker et al., 2007). The first version of LIWC dictionary was built from corpus analysis composed of 4500 words (Pennebaker et al., 2007), however now the LIWC2015 dictionary is updated with 6400 words that include word stems, and emoticons. LIWC2015 is the latest software where the dictionary has been upgraded by adding punctuation, emoticons, numbers, abbreviation and short phrases (Pennebaker et al., 2015). These additions allow to process social media data that is mostly obtained from Twitter and Facebook posts.

The main categories in LIWC that relate to sentiment analysis are Posemo (positive emotion) and Negemo (negative emotion). However, there are other categories that able to analyze emotions such as affect, anger, sad, and etc. Figure 2.2 shows the LIWC2015 categories list.

University of Malaya

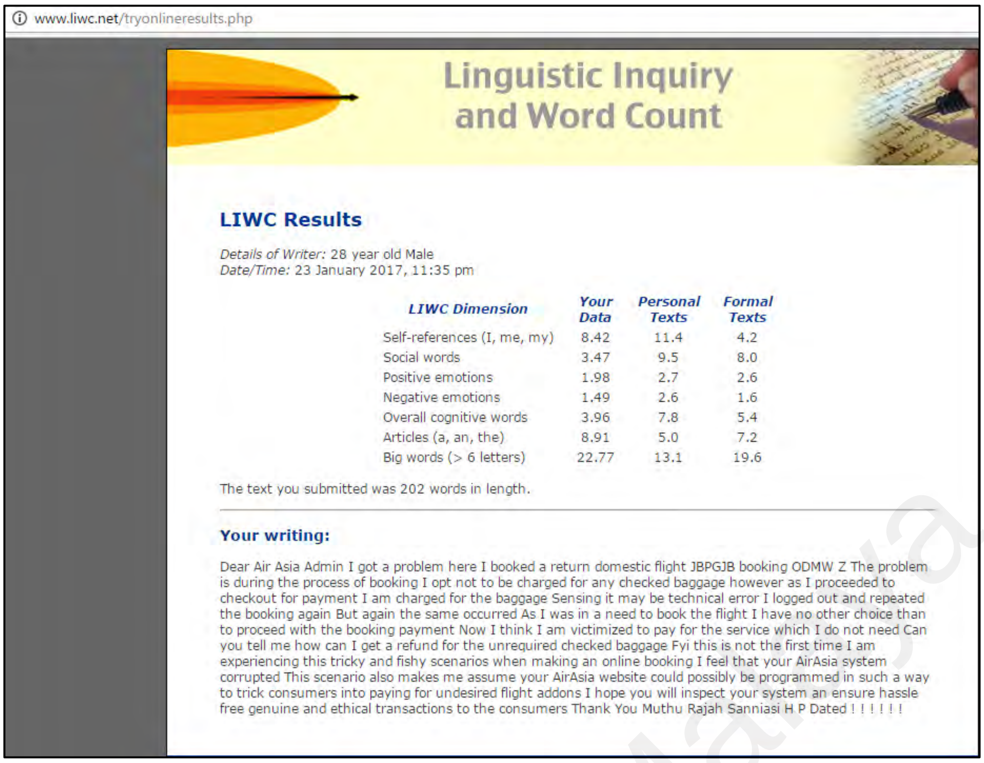


Category	Abbrev	Examples	Words in category	Internal Consistency (Uncorrected α)	Internal Consistency (Corrected α)
Word count	WC	-	-	-	-
Summary Language Variables					
Analytical thinking	Analytic	-	-	-	-
Clout	Clout	-	-	-	-
Authentic	Authentic	-	-	-	-
Emotional tone	Tone	-	-	-	-
Words/sentence	WPS	-	-	-	-
Words > 6 letters	Sixltr	-	-	-	-
Dictionary words	Dic	-	-	-	-
Linguistic Dimensions					
Total function words	funct	it, to, no, very	491	.05	.24
Total pronouns	pronoun	I, them, itself	153	.25	.67
Personal pronouns	ppron	I, them, her	93	.20	.61
1st pers singular	i	I, me, mine	24	.41	.81
1st pers plural	we	we, us, our	12	.43	.82
2nd person	you	you, your, thou	30	.28	.70
3rd pers singular	shehe	she, her, him	17	.49	.85
3rd pers plural	they	they, their, they'd	11	.37	.78
Impersonal pronouns	ipron	it, it's, those	59	.28	.71
Articles	article	a, an, the	3	.05	.23
Prepositions	prep	to, with, above	74	.04	.18
Auxiliary verbs	auxverb	am, will, have	141	.16	.54
Common Adverbs	adverb	very, really	140	.43	.82
Conjunctions	conj	and, but, whereas	43	.14	.50
Negations	negate	no, not, never	62	.29	.71
Other Grammar					
Common verbs	verb	eat, come, carry	1000	.05	.23
Common adjectives	adj	free, happy, long	764	.04	.19
Comparisons	compare	greater, best, after	317	.08	.35
Interrogatives	interrog	how, when, what	48	.18	.57
Numbers	number	second, thousand	36	.45	.83
Quantifiers	quant	few, many, much	77	.23	.64
Psychological Processes					
Affective processes	affect	happy, cried	1393	.18	.57
Positive emotion	posemo	love, nice, sweet	620	.23	.64
Negative emotion	negemo	hurt, ugly, nasty	744	.17	.55
Anxiety	anx	worried, fearful	116	.31	.73
Anger	anger	hate, kill, annoyed	230	.16	.53
Sadness	sad	crying, grief, sad	136	.28	.70
Social processes	social	mate, talk, they	756	.51	.86
Family	family	daughter, dad, aunt	118	.55	.88

Figure 2.2: LIWC2015 category list, (Pennebaker et al., 2015)

The LIWC will start reading all words in a given text and count the number of word that match each category. Then, the percentage of the count is calculated for each category of the dictionary. For example, a text with 2000 words will be analyzed with LIWC dictionary. At this point, it may find 160 pronouns and 90 negative emotion words used. So, it would calculate the percentage as 8.0% pronouns and 4.5% negative emotion words. In LIWC dictionary each entry is defined in more than one category. For example, the word "cried" belongs to five categories: negative emotion, sadness, affect, past focus and verbs. Hence, while analyzing a text, if the word "cried" is found, these five categories scores will be incremented concurrently.

University of Malaya


www.liwc.net/tryonlineresults.php
Linguistic Inquiry and Word Count

LIWC Results

Details of Writer: 28 year old Male
Date/Time: 23 January 2017, 11:35 pm

LIWC Dimension	Your Data	Personal Texts	Formal Texts
Self-references (I, me, my)	8.42	11.4	4.2
Social words	3.47	9.5	8.0
Positive emotions	1.98	2.7	2.6
Negative emotions	1.49	2.6	1.6
Overall cognitive words	3.96	7.8	5.4
Articles (a, an, the)	8.91	5.0	7.2
Big words (> 6 letters)	22.77	13.1	19.6

The text you submitted was 202 words in length.

Your writing:

Dear Air Asia Admin I got a problem here I booked a return domestic flight JBPGJB booking ODMW Z The problem is during the process of booking I opt not to be charged for any checked baggage however as I proceeded to checkout for payment I am charged for the baggage Sensing it may be technical error I logged out and repeated the booking again But again the same occurred As I was in a need to book the flight I have no other choice than to proceed with the booking payment Now I think I am victimized to pay for the service which I do not need Can you tell me how can I get a refund for the unrequired checked baggage Fyi this is not the first time I am experiencing this tricky and fishy scenarios when making an online booking I feel that your AirAsia system corrupted This scenario also makes me assume your AirAsia website could possibly be programmed in such a way to trick consumers into paying for undesired flight addons I hope you will inspect your system an ensure hassle free genuine and ethical transactions to the consumers Thank You Muthu Rajah Sanniasi H P Dated ! ! ! ! ! !

Figure 2.3: Sample of LIWC test result page
(http://www.liwc.net/tryonline.php)

Figure 2.3 shows a sample of result page of LIWC analysis. This result comprises the percentage value for each category (such as self-references, social words, positive emotion, negative emotion and etc.), length of the text and shows the text that used to for the analysis.

By reviewing LIWC, it shows this tool is only responsible to calculate percentage of the word count for each category. However, the incorporation of psychological process categories can be used as a part of emotion evaluation. LIWC is able to process a large text file and the latest version of LIWC2015 can be used to analyze social media data that mainly comprised with informal texts (acronym, punctuation, emoticons and numbers). When analyzing LIWC in term of sentiment analysis technique, this tool does not incorporate with many features such as negation, intensification, lexical valence, non-lexical valence, POS tags, sarcasm and irony. So it shows LIWC would not provide effective results towards the sentiment analysis process.

University of Malaya

2.6.4 SentiHealth-Cancer (SHC-pt)

SentiHealth-Cancer (SHC-pt) is a sentiment analysis tool is built to analyze the emotional state of cancer patients in Brazil via social media (Rodrigues et al., 2016). The main objective of this tool is to improve sentiment detection by detecting positive, negative and neutral messages from online cancer communities. SHC-pt was developed in Portuguese language and focused only on patients diagnosed with cancer. SHC-pt has analyzed online cancer patient posts that collected from two different cancer communities from Facebook. Besides, a comparative study with the existing system has been performed to identify the effectiveness of SHC-pt (Rodrigues et al., 2016). The polarity range for this tool is from −5 (most negative) to 5 (most positive). The features that incorporated with SHT-pt are as follows:

2.6.4.1 Dictionary set

Four different sets of dictionaries were used in SHC-pt tool: “dictionary.txt”, “emoticon.txt”, “hashtags. txt” and “ngrams.txt”. Every dictionary text file has the sentiment item list, where each row contains one sentiment item together with its sentiment score.

Four different sets of dictionaries were used in SHC-pt tool: “dictionary.txt”, “emoticon.txt”, “hashtags. txt” and “ngrams.txt”. Every dictionary text file has the sentiment item list, where each row contains one sentiment item together with its sentiment score. The “dictionary.txt” file is listed with Portuguese lexicons and its sentiment scores that are built based on SentiStrength dictionary. The “emoticon.txt” file is listed with 124 textual emoticons and its sentiment scores that were taken from SentiStrength tool. In addition, this emoticons list has also included question symbol with the score of zero. The “hashtags.txt” file is listed with hashtags that are commonly used in cancer groups on social networks, such as “#obrigadodoador” and “#obrigadodeus”.The “ngrams.txt” file listed with 86 n-grams and each line has four

University of Malaya

information: n-gram, its sentimental score, priority (1-yes or 0-no) and variation (1-yes or 0-no). For example, "*happy*: 4:1:1" which means n-gram "*happy*" with sentimental strength +4, it is priority and variations is considered (happiness, unhappiness, happily). This n-grams identification logic was embedded in SHC-pt tool.

2.6.4.2 Emoticons and hashtags

If an emoticon or hash tag in the dictionary file is contained in the sentence, other sentiment items will be disregarded, where the sentiment scores of the sentence will be calculated based on emoticon or hashtag only. Otherwise, priority will be given for emoticon and hashtag will be disregarded. For example:

If score for emoticon ":)" = 2 and

score for hashtag "#thankyougod" = 3 and

score for "*good*" =1

Example 1 (sentence with emoticon and lexicon):

"This is good deal :)"

Total score= 1-1(disregard lexicon) + 2 (emoticon score) =2.

Example 2 (sentence with emoticon and hashtag):

I achieved my target :) #thankyougod"

Total score= 3-3(disregard hashtag) + 2 (emoticon score) =2.

In example 1, there were two sentiment items present in the sentence: "*good*" lexicon and ":)" emoticon. So the score for lexicon disregarded and only counting the total score based on emoticon score. The same priority will be given for hashtag if a sentence contains hashtag and lexicon only. However, in example 2, it shows that priority only will be given for emoticon if the sentence contains both emoticon and hashtag.

University of Malaya

2.6.4.3 Question mark

A question sentence will be treated as a no sentiment sentence. Thus, the question symbol has been included in "emotions.txt" file with the score of zero. Since priority is given for emoticon, the sentence will be given score as zero. For example:

If score for "*good*"=1

Question mark sentence - "*Is this a good deal?*"

Total score= 1-1(disregard lexicon) + 0 (emoticon score) =0.

The example above shows, although the word "*good*" has sentiment score, it has been disregarded due to the presence of question symbol.

2.6.4.4 Exclamation mark, capitalized word and repeated vowels

If the sentence contains exclamation mark the sentiment scores will be calculated by doubling up the score value (multiply by 2). The same calculation will be performed if the sentence has uppercase case or word with repeated vowel, for example: "KIND" and "Kiiiiind".

2.6.4.5 N-gram

N-gram is a sequence of sentence where the n indicates the size of the sentence (Kok & Brouwer, 2011; Aisopos et al., 2016; Awachate & Kshirsagar, 2016). For example, unigram (n-gram size is 1), bigram (n-gram size is 2) and so on. Some n-gram may have the same word in "dictionary.txt" file and in "n-gram.txt" file (between unigram and bigram). This may cause confliction when performing the sentiment score calculation. Thus in SHC-pt tool, priority is given to unigram. The unigram will be removed from the sentence after the sentiment score is calculated. In some cases, priority will be given for the bigger n-gram if the removal of unigram could affect the strength of the bigger n-gram. For example; "*fight and win cancer*", if the unigram "*cancer*" is removed, it would reduce the strength of the bigger n-gram where the sentence becomes as "*fight and win*".

University of Malaya

In SHC-pt, the score calculation for a document is performed based on priority sentence. Sentences that have emoticon, hashtag, exclamation mark, capitalized word or some n-gram are known as priority sentences. Priority sentence will be given importance to calculate the sentiment score for each sentence in the document. If the document has more positive sentence than negative sentence, the document will be assigned positive scores and vice versa.

By reviewing the features included in SHC-pt tool, it shows that SHC-pt performs well compared to other general purpose tools such as Textalytics, SentiStrength, AlchemyAPI and Semantria (Rodrigues et al., 2016). SHC pt is only designed with specific context for cancer (Rodrigues et al., 2016). However, it is unable to support misspelled word, slang, irony, sarcasm and other indirect expression in sentence. Furthermore, this tool does not incorporate other important features such as intensifiers and negation, which are derived in sentiment polarity determination.

University of Malaya

2.6.5 SmartSA

SmartSA is a lexicon based sentiment analysis tool that specially designed to analyze social media data responsible to capture contextual polarity from the interaction of two perspective: local (textual neighborhood) and global context (global genre) (Muhammad et al., 2015). SmartSA introduced hybridized lexicon created with the combination of SentiWordNet (SWN) dictionary and generated domain lexicons. The polarity range of SmartSA is as per SWN (-1 to 1). This is because most lexicons were obtained from SWN dictionary that already have specific polarity value for each lexicons.

Figure 2.4 shows the overall process of SmartSA and following sub-sections explained the major activities involved to analyze the data.

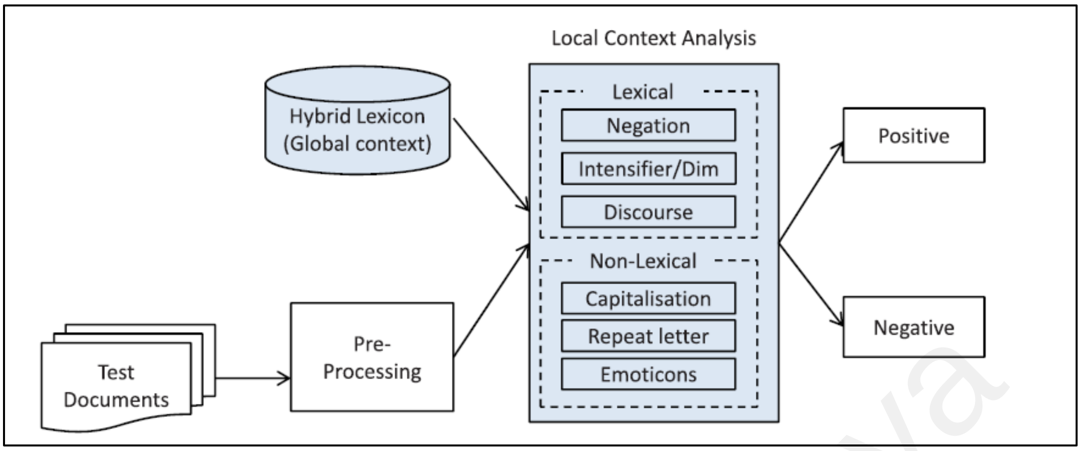


Figure 2.4: SmartSA- Overall process (Muhammad et al.,2015).

2.6.5.1 Pre-processing

This is a prior activity for sentiment classification, where the input text will be transformed into unit terms together with its information. In pre-processing, tokenization and POS tagging will be performed by using TweetNLP (Gimpel et al., 2011). Each token is converted into dictionary by using lemmatization (Manning et al., 2014) after extracting scores from SWN by using the word lemma.

2.6.5.2 Global context

SWN is a sole sentiment lexicon for SmartSA, where it is only contributing general sentiment bearing words towards the analysis. This provides limitation as the domain-specific terms are being ignored. To overcome this limitation Muhammad et al. (2015) introduced hybridized dictionary: combination of SWN dictionary and the sentiment context extracted from target domain (Twitter, Digg and MySpace). Figure 2.5 shows the generation of hybridization dictionary.

University of Malaya

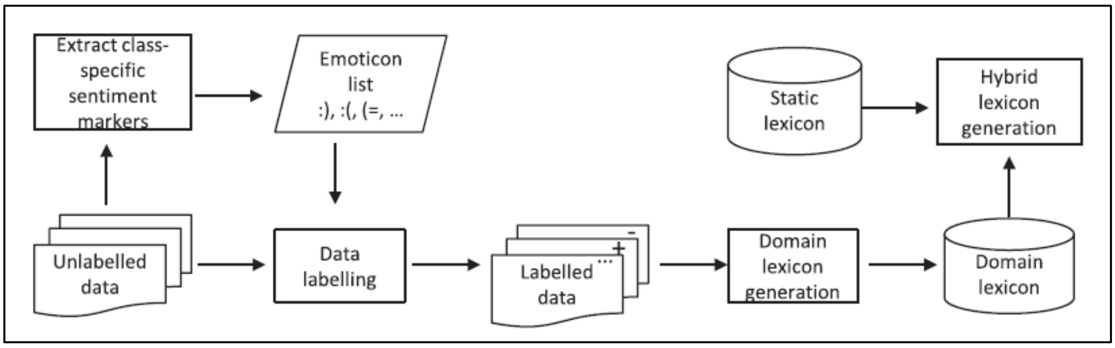

Extract class-specific sentiment markers
Emoticon list :), :(, (=, ...
Static lexicon
Hybrid lexicon generation
Unlabelled data
Data labelling
Labelled data
+
-
...
Domain lexicon generation
Domain lexicon

Figure 2.5: Process of hybridization dictionary generation (Muhammad et al., 2015)

First, the sentiment context is identified from the domain data and labeling is done using distant supervision approach to generate domain lexicon. Next, generate hybridized lexicon by combining domain lexicon and its sentiment scores with existing SWN dictionary.

2.6.5.3 Local context

Local context are the features that can alter the meaning of a phrase. This context was defined as lexical valence shifter that can be divided into two categories: lexical and non-lexical valence. The lexical valence shifters are recognizable words which are intensification/diminishing and Negation. Whilst, the non-lexical valence shifters are the informal texts that mostly present in social media data. The non-lexical valence shifters can appear in the form of a sequence of repeated letters, capitalization and emoticons.

Below are the explanations given for the lexical valence shifters in SmartSA:

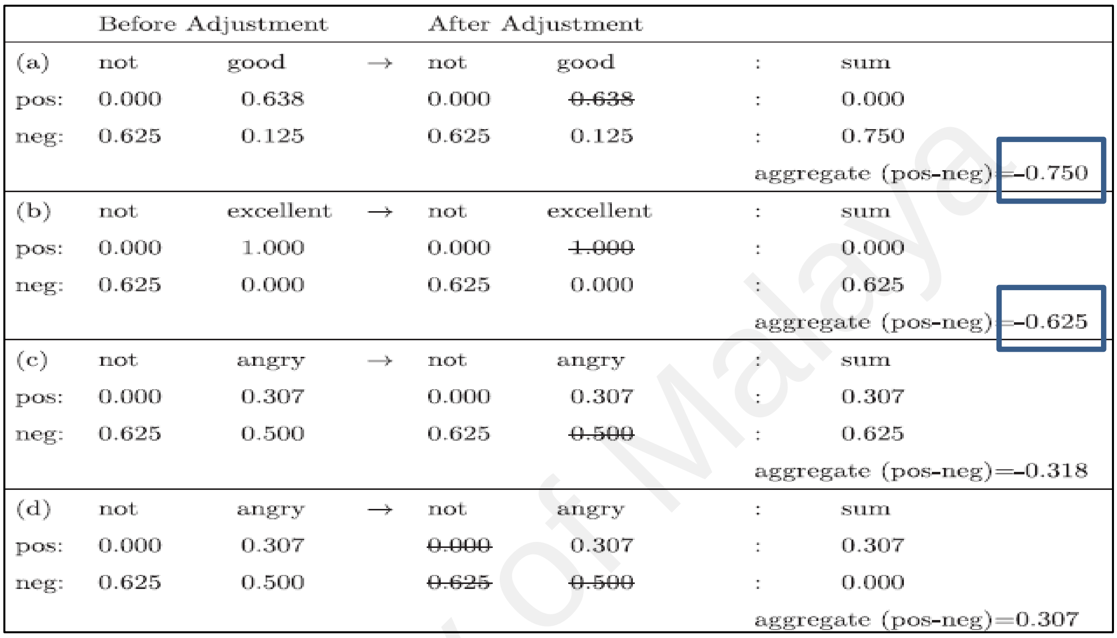
(a) **Intensification/diminishing**

SmartSA has its own dictionary for words intensification where each word has its own percentage score. This is similar to the way SO-CAL deals with intensification words.

University of Malaya

(b) **Negation**

In SmartSA, negation will be calculated by applying shift negation. SmartSA has the ability to calculate sentiment score according to the strength of the lexicon present. For example: the phrase “not good” is more negative than the phrase “not excellent”, (see figure 2.6).



	Before Adjustment			After Adjustment			
(a)	not	good	→	not	good	:	sum
pos:	0.000	0.638		0.000	~~0.638~~	:	0.000
neg:	0.625	0.125		0.625	0.125	:	0.750
							aggregate (pos-neg)=–0.750
(b)	not	excellent	→	not	excellent	:	sum
pos:	0.000	1.000		0.000	~~1.000~~	:	0.000
neg:	0.625	0.000		0.625	0.000	:	0.625
							aggregate (pos-neg)=–0.625
(c)	not	angry	→	not	angry	:	sum
pos:	0.000	0.307		0.000	0.307	:	0.307
neg:	0.625	0.500		0.625	~~0.500~~	:	0.625
							aggregate (pos-neg)=–0.318
(d)	not	angry	→	not	angry	:	sum
pos:	0.000	0.307		~~0.000~~	0.307	:	0.307
neg:	0.625	0.500		~~0.625~~	~~0.500~~	:	0.000
							aggregate (pos-neg)=0.307

Figure 2.6: Negation calculations for SmartSA (Muhammad et al., 2015)

Figure 2.6 shows the calculations for negation in SmartSA. In Sentiwordnet (SWN) dictionary (same dictionary used by SmartSA), each lexicon has two values which belong to positive and negative score. The positive score for negation is zero. An adjustment will be made by giving priority for the type of lexicon present in negation. If positive lexicon present in negation the positive scores of lexicon will be disregarded and calculates the sum of negative scores of both negation and lexicon. The calculation shows in figure 2.6 (a) and (b). If negative lexicon present in negation, the negative score of lexicon will be disregarded and calculates the sum of positive score of lexicon and negative score of negation (e.g. Figure 2.6(c)). If total calculation gives negative scores, then the adjustment will be changed by giving final score same as per positive score of the lexicon (e.g. Figure 2.6 (d)).

University of Malaya

Below are the explanations given for the non-lexical valence shifters in SmartSA:

(a) Repeated letters

In SmartSA, when repeated letters are detected, the letters will categorically be reduced one at a time and checked with SWN. This process will be performed a maximum of two times. If the word is present in SWN, the polarity score will be calculated by assigning intensification value for the word "very". For example: "*I am not happpy*", score for happy word is 2, percentage for the "very" word is 25% so the score will be 2 (original score of "*happy*" word) + [25% (percentage value of "*very*" word) * 2 (original score of "*happy*" word)] = 2 + 0.5 = 2.5.

(b) Capitalization

SmartSA introduced capitalization as an intensification in sentiment detection. The capitalization detection will take place if the whole sentence or phrase is not being capitalized. This is because if the whole sentence is capitalized, it may indicative of a writing style. Similarly, an intensification value for the word "very" will be used for sentiment score calculation for the repeated letters that are successfully detected. For example:

"*I am UPSET with you!"* changed to *"I am very upset with you*"

(c) Emotions

The emoticons provided in SmartSA are those that are provided by Thelwall et al. (2010). The scores for the emoticons detection has been set in a simple way, where if found one or more positive emoticons in a sentence, the score will be given as 1. The same will be applied to negative emoticons. For this tool the emotions context has been restricted to sentence level. This is because sentiment has the ability to change from one sentence to another (Andreevskaia et al., 2007).

University of Malaya

By reviewing SmartSA tool, it has improved classification accuracy and it performs better than SentiStrength tool as well (Muhammad et al., 2015). Incorporation of hybridized dictionary in SmartSA enhances the performance across social media platforms such as Twitter, Digg and MySpace (Muhammad et al., 2015). However, there is restriction on non-lexical valences, where a general value has been given if the sentence is detected with repeated letter or emoticons. This needs to be improved as length of sentiment item (such as word, punctuation mark etc.) influences a significant change towards the sentiment strength (Brody & Diakopoulos, 2011). Furthermore, SmartSA also has compatibility problems when analyzing different context of data, which will be investigated in further studies.

2.7 Other lexicon-based approach studies

In this section, the discussion will focus on relevant studies that were recently conducted to improve the lexicon based sentiment analysis.

2.7.1 Lexicon‑based sentiment analysis for social media analytics

In Jurek et al., (2015), a new algorithm of lexicon-based approach was proposed in order to perform real-time analysis on Twitter data. The lexicon dictionary used in this study was generated manually with the reference of SentiWordNet (SWN) dictionary. There are 6300 sentiment lexicons integrated in this dictionary. Besides, the polarity range of the sentiment lexicon is −100 (most negative) to 100 (most positive). There are two important components merged in this new algorithm: sentiment normalization and evidence-based combination function. These components are responsible to estimate the intensity of sentiment and also to support classification process that involve mixed sentiments (positive and negative sentiments).

Sentiment normalization is used to combine the average sentiment of a sentence and the number of words used to calculate the average. Besides, evidence-based combination is used to determine the sentiment of a sentence if it contains mixed

University of Malaya

sentiment, where a formula is used to identify the higher weight in sentiment (between negative and positive) and finalized the sentiment score by taking the highest weight sentiment.

Moreover, this algorithm has also incorporated intensifiers and negation. Intensifiers in this algorithm are divided into three group downtoners, weak amplifiers and strong amplifiers where it will be calculated in percentage value, similarly done with SO-CAL and SmartSA. Negation in this algorithm does not use switch method, however a new calculation was proposed for negation that involves negation dictionary that contain specific scores for each negation word. The evaluation of this algorithm was performed with Stanford Twitter test set and Internet Movie Database (IMDB) data set, and with results showing that the two new components improved the performance of standard lexicon based analysis (Jurek et al., 2015). However, this algorithm has weakness as the performance is lower to process long documents compared with short messages, such as tweets.

2.7.2 Lexical-based sentiment analysis – verb, adverb and negation

In Shamsudin et al. (2016), a study conducted to introduce lexical based sentiment analysis on Facebook comments in the Malay language. The main contribution of this research is the enhancement of the term counting method into term counting average. Term counting (TC) is a method to calculate the overall average of sentiment without looking into the type of sentiment (positive or negative). Term counting average (TCAvg) is a method to obtain average value by calculating the difference between positive and negative average values of a sentence (calculates the positive and negative average separately). Besides this study incorporated POS tags that contains, verb, adverb and negation. The lexicon dictionary of these studies was built by combining adjectives from Wordnet Bahasa and Indonesian lexicon dictionary.

University of Malaya

Moreover, additional dictionaries were created for negation, adverb and verb. Adverb and verb dictionary built with scores that manually assigned. Then, negation for this study is using a switch method, where it reverses the polarity of a sentiment. For example if the neighborhood sentiment value is -3 its will become +3 and vice versa. This proposed method was tested with Facebook comments that manually verified by human judgments. From the evaluation between TC and TCAvg method, TC shows a better performance in adjective and TCAvg shows a significant improvement with the incorporation of POS such as verb and adverb (Shamsudin et al., 2016). However, this study does not incorporate intensifiers which can cause major problems in sentiment analysis.

2.7.3 Data mining

Mehto and Indras (2016), proposed a lexicon-based approach with an additional feature, called aspect catalogue. The mechanism of this sentiment analysis works by identifying the presence of keyword in aspect catalogue. Once identified, it will continue the process with sentiment detection. The purpose of aspect catalogue is to find the degree of importance towards the features of a product. The aspect catalogue was manually created and the context used for this analysis is mobile phone reviews.

The weightage for the aspect features have been determined based on the frequency of feature appearance in mobile phone reviews. The lexicon dictionary used for this study is SentiWordNet (SWN). By adding aspect catalogue in sentiment detection, it helps to achieve more accurate ratings (Mehto and Indras ,2016). However, it is unable to process informal text (e.g repeated letter, capitalization, emoticons, acronym) in the form of substantial amount of expressions, this brings about limitations to this proposed tool. Another limitation is that it requires huge human effort to build the aspect catalogue.

University of Malaya

2.8 Summary

In this chapter, an introduction has been given for natural language processing (NLP) and the problems faced in NLP such as level of ambiguity and complexity of semantic information. Then the term of sentiment lexicon has been explained. There are two approaches for lexicon creation; manual lexicon creation and automated lexicon creation that have been explained by providing example of studies. Furthermore, the necessity of lexicon modifiers in lexicon based sentiment analysis has been discussed. Finally, a detailed explanation given on existing lexicon based sentiment analysis system: semantic orientation calculator (SO-CAL), SentiStrength, Linguistic Inquiry and Word Count (LIWC), SentiHealth-Cancer (SHC-pt) and SmartSA. Overall, each sentiment analysis system has its own advantages and disadvantages.

From the literature studies, it shows a majority of the systems are built to cater for the English language except SHC-pt that only supports the Portuguese language. Muhammad et al. (2015) showed that SmartSA is able to achieve better accuracy compared to SentiStrength. SentiStrength can perform better than SO-CAL, however the limitation to the number of characters in SentiStrength causes a concern.

Although SO-CAL is proven as a robust tool across different domains, the performance of this tool may not be satisfactory when it comes to analyzing social media data. This is because informal text used on social media requires non-lexical modifiers to produce an effective result and SO-CAL does not have the feature to analyze typo or misspelled words. Besides, Rodrigues et al. (2016) showed that SHC-pt has a better accuracy result than SentiStrength when analyzing context based data in the Portuguese language. LIWC is a text analysis tool, but it was not designed with sentiment features like other tools. However, LIWC can be used as a reference tool which can identify the overall percentage of positive and negative words in a document.

Besides the comparison made among these five tools, a similar deficiency has been identified from all the existing systems. That is implementation of non-lexical

University of Malaya

modifiers. Although some tools have already incorporated non-lexical modifies, but still they were not fully expanded. This can pose a threat when defining the strength of the sentiment. The focus of this research is to improve the non-lexical modifiers (especially repetition letter, capitalization and exclamation). This could provide a better outcome from sentiment analysis technique applied on social media data.

University of Malaya

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

A methodology implies a set of steps used to conduct a research. Likewise, the steps taken to complete this research will be stated and discussed within this chapter. The following sections will provide a detailed explanation on system methodology deployed in this research, system requirements to develop the proposed scoring mechanism, data collection, data cleaning, data analysis and evaluation planning.

3.2 System methodology

Rapid Application Development (RAD) is known as modern SDLC (System Development Life Cycle) because it involves iterative process which is commonly adopted for rapid software application development. This methodology has been adopted to develop the LexiPro-SM as it is ideal for the development of web-based application development. Figure 3.1 shows a pictorial explanation of the phases involved in RAD methodology.

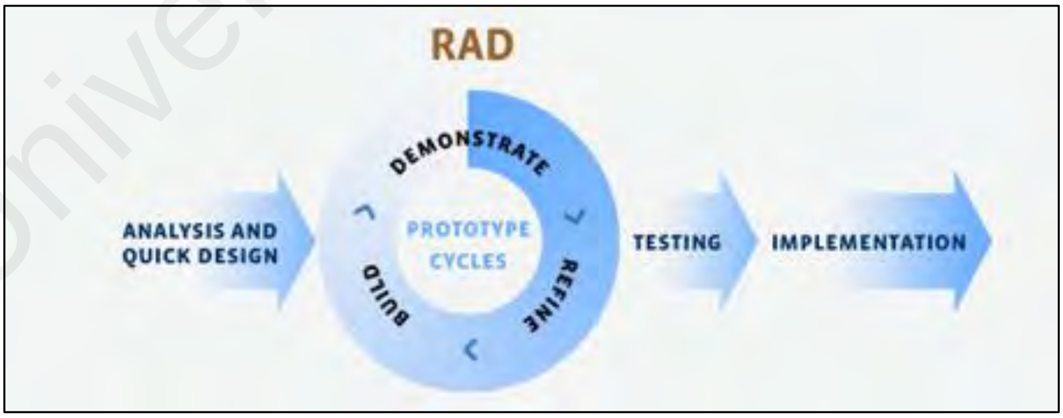


Figure 3.1: RAD Methodology Phases (http://www.testingexcellence.com/rapid-application-development-rad/)

University of Malaya

i) Phase One: Analysis and quick design

This phase is known as quick design where literature review performed and analyzed existing sentiment analysis tools to gather system requirements upon which a basic design flow is finalized. The design comprises of screen flows and layouts for essential parts of the system.

ii) Phase Two: Construction prototype

In this stage, the LexiPro-SM will be developed in an iterative cycle development where the testing, requirement refining and reconstruction will be continued until the development is completed. The data model will be converted into a functional model. It is necessary to deliver all iterations on time while scheduling other functionalities for the development after the initial release. Once the initial prototype has been developed, it will be tested using test data that prepared during quick design phase. Upon completion of prototype test phase, a review will be conducted to define the effectiveness of the initial prototype. After the review, the next iteration requirements will be determined.

iii) Phase Three: Testing

During this stage the test focused on the completed scoring mechanism rather than the prototypes iterations. A user performed testing on the developed LexiPro-SM.

iv) Phase Four: Implementation

In this stage, the LexiPro-SM will be used to process live collected data. Then an evaluation was performed to ensure the effectiveness of the LexiPro-SM in live data.

University of Malaya

3.3 System requirements

In this section, a brief explanation is given on the main tools used to develop the LexiPro-SM mechanism. The tools are PHP programming language, Apache server, phpMyadmin database and MySQL.

i. **PHP programming language**

PHP is a scripting language that is widely used as an open source programming language which mostly supports web development and can be embedded into HTML. It is known as an easy programming language and suitable for use as text based analysis as PHP possess useful features for text processing. Furthermore, previous experience using this programming language helped shorten the learning curve when it came to developing LexiPro-SM. The PHP program coding was done using Notepad++.

ii. **Apache Server**

Apache server is an open source web server that creates interconnection between PHP programming language and phpMyadmin database. This local server supports offline web-based application development.

iii. **phpMyadmin and MySQL**

phpMyadmin is a free database that interconnects PHP programming using MySQL query language. In this research, all the data activity will be handled in phpMyadmin database.

University of Malaya

3.4 Operational Framework

In the process of gathering all the information needed, the researcher had built an operational framework. The purpose of this framework is to provide an overview of all the tasks carried out throughout the entire research. The operational framework main activities involved in research initiation, literature review, design, data collection, data cleaning, LexiPro-SM development, testing and report writing. Figure 3.2 illustrates the operational framework.

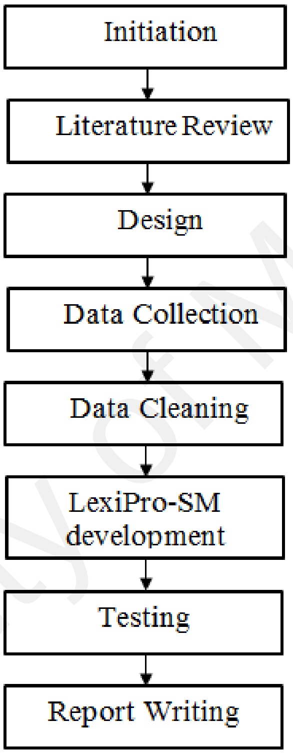


Figure 3.2: Operational Framework

University of Malaya

Figure 3.2 shows overall work flow in this research. First, is initiation that involves discussions with a supervisor to brainstorm ideas. Second is literature review, where research was initiated by studying existing systems relating to this research and to identify the objective, research question and project scope. Third, the design framework for data collection, data cleaning and data analysis (LexiPro-SM) was carried out. Fourth, perform data collection from Facebook airline pages and stored it in a database. Fifth, perform data cleaning on collected data from Facebook and store the clean data in database. Sixth, involves development of the scoring mechanism with new features. Seventh, perform testing on the developed scoring mechanism. Lastly, perform report writing by compiling all the activities in this research. Following sections discuss data collection, data cleaning and data analysis which are important in LexiPro-SM sentiment analysis.

3.5 Data collection

Facebook is the main source of data collection (comments). This is because Facebook, unlike Twitter has the capability to store large amounts of data in a single post and there is no restriction to extract the data from public pages of Facebook (Rastogi et al., 2014; Troussas et al., 2013). Data has been collected from the official pages in Facebook that belong to two major airlines in Malaysia, which will be referred to Airline A and Airline B respectively.

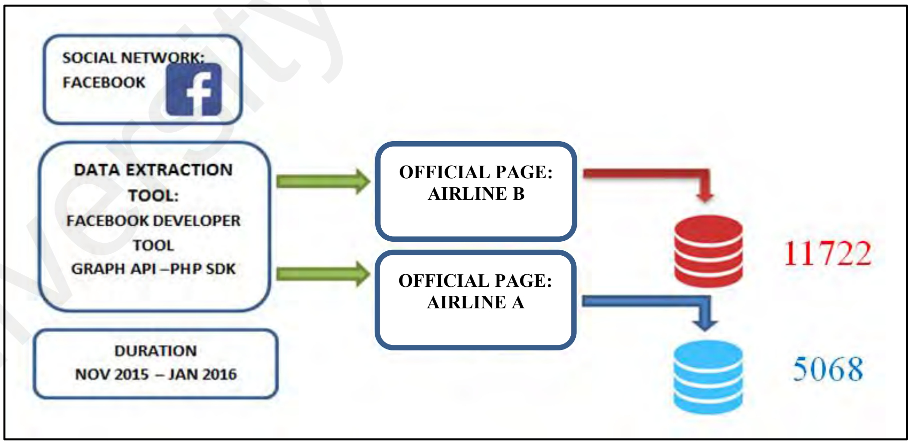

SOCIAL NETWORK: FACEBOOK
DATA EXTRACTION TOOL: FACEBOOK DEVELOPER TOOL GRAPH API –PHP SDK
DURATION NOV 2015 – JAN 2016
OFFICIAL PAGE: AIRLINE B
OFFICIAL PAGE: AIRLINE A
11722
5068

Figure 3.3: Data Collection

Figure 3.3 shows the data collection process, where all the data for this research were collected with the aid of Facebook developer tool which is known as Graph API. Graph API is a low-level HTTP-based API tool that is used to retrieve or publish data (Weaver and Tarjan, 2012). Graph API has many software development kits (SDKs) with

University of Malaya

different languages such as iOS, Android, JavaScript, PHP SDKs and etc. Besides, each SDKs has their own API reference which will be used to access the Facebook data. Thus, for this research the data will be collected by using PHP SDKs which can be used in PHP environment. In order to access the Facebook data with API references, some important information must be included in the scripting work. The information is application id, application secret and the Facebook page id (the page id of Airline A and Airline B).

For this research, three months' worth of comments were collected from the posts between November 2015 and January 2016. The reason for choosing this particular time frame is because this duration falls within the school holidays and festival seasons in Malaysia and is considered peak season for the airline industry with the assumption that this peak season may lead to an increase in comments in the airlines official Facebook pages. When performing the data extraction, all the posts and comments that were posted by the administrator (of the airline page) and empty data were neglected.

The Table 3.1 shows the total data collected from each official page of the airline in Facebook for the duration of three months.

Table 3.1: Raw data count for each airline

Duration: Nov 2014- Jan 2016	**Airline A**	**Airline B**
Total Data	5068	11722

All data collected from Facebook were stored in a database where each airline has its own table to store its raw data. The information stored in a raw data table is post id (unique) of the comment, date of the comment posted, and the comment.

University of Malaya

3.6 Data cleaning

Data cleaning is the pre-processing method that involves operations to remove noises in raw data (Seerat & Azam, 2012; Rahm & Do, 2000). Noises are the meaningless data or irrelevant information for a data analysis and they vary in every type of data analysis (Xiong et al., 2006). Thus, the identification of noise elements is important for data analysis (Zafarani et al., 2014). The main idea of this research is to perform data analysis towardS the text collected from Facebook which can be referred to as text analysis. Facebook is one of the social media that contain informal text and also contain different types of irrelevant information. However, the determination of noises is fully dependent on the criteria of the data is required for data analysis.

In this research, the data analysis is fully focused on the text (alphabets) and only one punctuation symbol which is exclamation mark (!). Other than that, everything can be categorized under irrelevant data. The elements of noises that was identified in this research are:

a) Hashtag

b) URL link

c) Emoticon

d) Numeric characters

e) Every symbol except exclamation mark

f) Special characters

f) Malay words

g) Irrelevant phrases (e.g greetings, words contain no and etc).

The data cleaning process starts with the removal of elements which are used for navigation purposes: Hashtag (e.g #merdeka, #iloveu and etc.) and URL link (e.g https://www.facebook.com/, https://www.google.com/ and etc.). Then, the removing

University of Malaya

process continued with emoticon (e.g "B|","o.O","(y)","O.o",":P","=D" and etc.). Although sentiment analysis techniques can be used to analyze Hashtag and emoticons (Qadir & Riloff, 2013; Koto & Adriani, 2015), in this research they will be treated as noises. This is because, the main focus of this research is to identify sentiment from the misspelled word and improve the scoring mechanism with the lexicon modifiers. The cleaning process continued by removing all the numeric characters, special characters (e.g "த" ,"泰" ," لمي" and etc.) and symbols (e.g. "?", ".", "=" and etc.) except the exclamation mark (!). The reason for using exclamation mark symbol as a relevant data in this research is because this symbol is an important element in sentiment analysis that can be used to emphasis the strength of the sentiment (Taboada et al., 2011; Muhammad et al., 2015; Paltogloua et al., 2010; Thelwall et al. 2012).

Another important cleaning stage in this phase is removing Malay words from the raw data, because the importance of this research is only given to processing data that relate to the English language. Malay words will be removed by searching the existence of the words in the Malay dictionary. The Malay dictionary used in this cleaning method was obtained from the PHP SpellCheck dictionary that has GPL license from MySpell. However, the cleaning process does not remove the alphabets that are not related to the English language. This is because; irrelevant alphabets in the form of misspelled words are the main source of data used in this research.

Finally, the data cleaning process will be applied on irrelevant phrases. Through observations from data collection, many comments were found containing greetings such as "*good morning*", "*happy new year*", and "*thank you* (at the end of text)" which are not important to be used in data analysis and it may contribute to false results. Similarly, most of the raw data contain acronym for the word "*number*" as "*no*". The word "*no*" is a negation word that play a big role in determining the sentiment of a sentence, so it also could cause faulty results during the data analysis process. Therefore, all the acronym for the word "number" (no) have been removed. For

University of Malaya

example: the phrase "booking no" will be replaced by "booking". Below is an example of raw data and clean data for the same comment.

Raw data: "*I have login into your website http://www.airline.com/ and it does not have information as provided in facebook #airline page. unable to find accurate info !!! :(terima kasih 1234 米尔*"

Clean data: "*I have login into your website and it does not have information as provided in facebook page unable to find accurate result ! ! !*"

Table 3.2 shows the total count of the data for each airline after performing the data cleaning process.

Table 3.2: Total clean data for Airline A and Airline B airlines

Total clean data	**Airline A**	**Airline B**
	4291	10189

3.7 Data analysis

This section is the most significant phase in this research, where the enhancement of scoring mechanism is implemented. Therefore, the discussion in this section will be based on the incorporation of dictionaries, intensifiers, negation and non-lexicon modifiers such as capitalization, repetition letters and exclamation mark. The following sub-sections will provide an in-depth explanation of the LexiPro-SM development.

University of Malaya

3.7.1 Type of dictionary used in LexiPro-SM

In this scoring mechanism, there were five types of dictionary used to analyze the data. The dictionaries are: English dictionary and Malay dictionary, lexicon dictionary, intensifier dictionary and acronym dictionary. Following are the explanation given for the four types of dictionaries:

i. **English dictionary and Malay dictionary**

This dictionary is a collection of English words which will be used to check the spelling of the word whether it is correct or not. Similarly, Malay dictionary also has a collection of Malay words with a same structure, because both English and Malay dictionaries were obtained from the same source. These dictionaries were obtained from the PHP SpellCheck dictionary that has GPL license from MySpell.

ii. **Lexicon dictionary**

The lexicon dictionary was taken from SO-CAL, where the lexicons were hand-ranked with a scale of +5 to -5. It contains 2,252 adjective entries, 1,142 nouns, 903 verbs, and 745 adverbs (Taboada et al., 2011).

iii. **Intensifier dictionary**

This dictionary also was taken from the SO-CAL, where the intensifier word labeled with a percentage value. It has 177 words such as (very, extremely, somewhat and etc.) (Taboada et al., 2011).

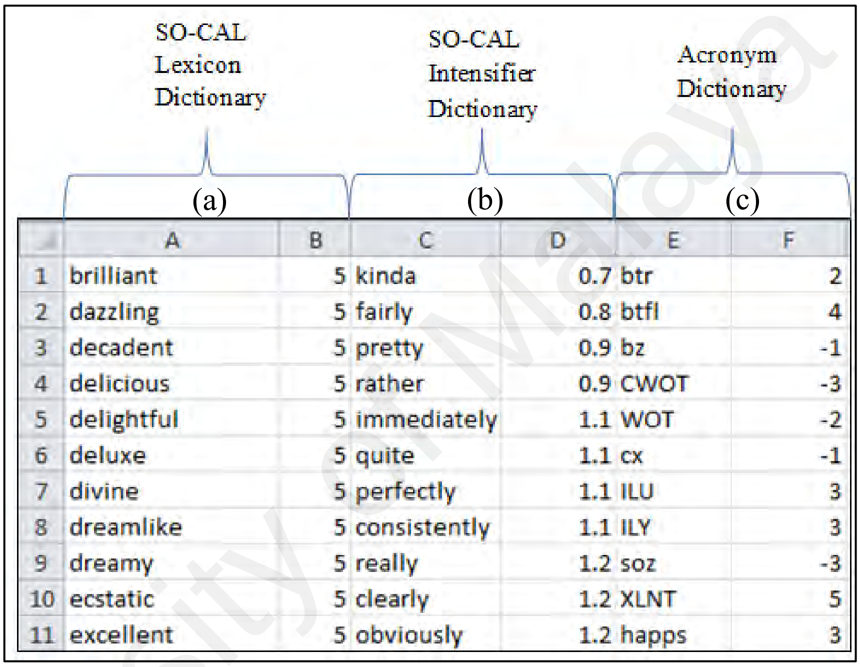
iv. **Acronym dictionary**

This dictionary is a small collection of short form words that were frequently used in informal conversation. For example: "*gud*" (good), "*hpy*" (happy) and etc. It was manually created by referring to the actual scores of the abbreviation in lexicon dictionary.

University of Malaya

3.7.2 Dictionary generation for LexiPro-SM

In LexiPro-SM all the dictionaries were obtained from other sources such as SO-CAL and PHP SpellCheck dictionary except acronym dictionary was manually created. However, in LexiPro-SM three dictionaries were combined into one file (excel.csv format used), which are SO-CAL lexicon dictionary, SO-CAL intensifier dictionary and acronym dictionary.

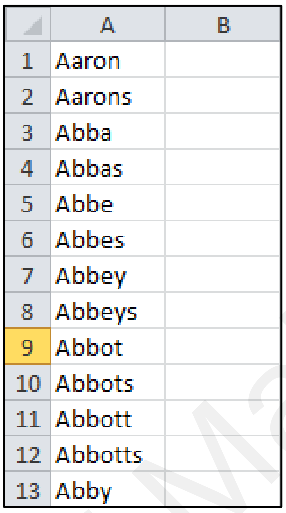


	A	B	C	D	E	F
1	brilliant	5	kinda	0.7	btr	2
2	dazzling	5	fairly	0.8	btfl	4
3	decadent	5	pretty	0.9	bz	-1
4	delicious	5	rather	0.9	CWOT	-3
5	delightful	5	immediately	1.1	WOT	-2
6	deluxe	5	quite	1.1	cx	-1
7	divine	5	perfectly	1.1	ILU	3
8	dreamlike	5	consistently	1.1	ILY	3
9	dreamy	5	really	1.2	soz	-3
10	ecstatic	5	clearly	1.2	XLNT	5
11	excellent	5	obviously	1.2	happs	3

Figure 3.4: LexiPro-SM dictionaries “lexiprodict.csv”

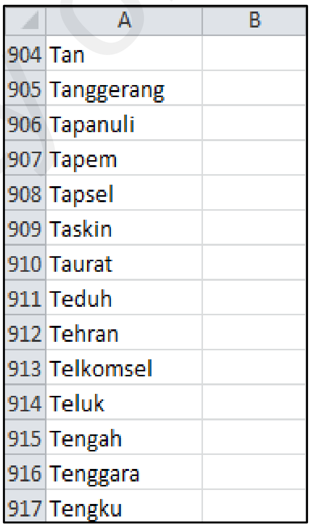
University of Malaya

Figure 3.4 is the combination of three types of dictionaries where (a) belongs to SO-CAL lexicon dictionary, (b) belongs to intensifier dictionary and (c) belongs to acronym dictionary. Each dictionary has two columns where first column contain the sentiment words and second column is contain the score for the sentiment words. For example in Figure 3.6 (c), column E contains the sentiment words (acronym words) and column F contain the scores for the words. The reason for the combination of dictionaries in one file is because in LexiPro-SM the dictionaries will be read in “.csv” format and it is being read, each dictionary will be loaded in to an array concurrently. If each file is read separately it will take time to load the dictionary into each array. However, the language dictionaries will be treated separately as both dictionaries have a huge collection of words. Figure 3.5 is the interface of English dictionary where the file name is “engdict.csv” and Figure 3.6 is the interface of Malay dictionary where the file name is “malaydict.csv”.



	A	B
1	Aaron	
2	Aarons	
3	Abba	
4	Abbas	
5	Abbe	
6	Abbes	
7	Abbey	
8	Abbeys	
9	Abbot	
10	Abbots	
11	Abbott	
12	Abbotts	
13	Abby	

Figure 3.5: English dictionary “engdict.csv”



	A	B
904	Tan	
905	Tanggerang	
906	Tapanuli	
907	Tapem	
908	Tapsel	
909	Taskin	
910	Taurat	
911	Teduh	
912	Tehran	
913	Telkomsel	
914	Teluk	
915	Tengah	
916	Tenggara	
917	Tengku	

Figure 3.6: Malay dictionary “malaydict.csv”

University of Malaya

3.7.3 Examples of sentiment calculation in LexiPro-SM

LexiPro-SM is the lexicon-based sentiment analysis tool that is able to identify the sentiment present in a text and calculates the total scores for the sentiment present in a text. Below are examples that show the calculation for three types of sentiments which are positive, negative and neutral.

Example 1 (positive sentiment):

If score for the lexicons "good" =3 and "delicious" =5

Positive sentence = "*The restaurant providing a good customer service and the food served was delicious*"

Total score= (3 [score for the word "good"] + 5 [score for the word "delicious"]) / 2 (number of lexicon present in a text) = 8/2 =4.

Example 2 (negative sentiment):

If score for the lexicons "*stupid*" = -4 and "hate" = -4

Negative sentence = "*You are stupid and I hate you*"

Total score= (-4 [score for the word "stupid"] + -4 [score for the word "hate"])/2 (number of lexicon present in a text) = -8/2 = -4.

Example 3 (neutral sentiment):

Negative sentence = "*Your payment has been processed*"

Total score= 0. (no sentiment detected)

Above examples shows the calculation for sentiment detection in a sentence where example 1 is for positive sentiment, example 2 is for negative sentiment and example 3 is for neutral sentiment. The calculation of total scores is the average value of the sum of sentiments score over the number of sentiment present in a sentence like in examples

University of Malaya

1 and 2. However for neutral detection, the score will be given as zero if there is no sentiment word present in a text/phrase.

3.7.4 Features incorporation in LexiPro-SM

The score calculation for LexiPro-SM has embedded with six features: intensifier, negation, repeated letters, capitalization, exclamation mark and auto word correction. The explanation given for these five features as follows:

a) Intensifier

Intensifier calculation for this LexiPro-SM follows the SO-CAL intensifier calculation (Chapter 2), where each intensifier has its own percentage value that describes the strength of the intensifier word. For example:

If score for "*happy*" word =2 and

Percentage value for "*really*" = 15%

Amplifier sentence = *"really happy"*

Total sentiment score = 2 (score for "happy")× (100% +15%) = 2.3.

In social media, the informal text can appear in many forms to express people's emotion such as capitalization, repeated letter or punctuation mark. Similarly, intensifier word also can appear twice or more to emphasize the sentiment of the word. For example: "*I am very very happy*". This shows that the person is happier than very happy (Bhaskar et al., 2014). Hence, a calculation is included to count for text/sentence that has two intensifier words with the neighborhood. For example:

University of Malaya

If score for "*happy*" word =2 and

Percentage value for "*very*" = 25%

Amplifier sentence = *"I am very very happy"*

Total sentiment score = (2 (score for happy)× (100% +25%)) x (100% +25%)) = 2.88.

The example above shows the similar calculation for intensifier. However, additional percentage value is added in this calculation due to two intensifier words present in the same sentence.

The first calculation will be calculated for "*very happy*" and followed by a second calculation for the additional word "*very*". Finally, the value obtained from the first calculation will be multiplied with value of the second calculation to get a sentiment score for "*very very happy*"".

b) Negation

Negation in this scoring mechanism can be divided into three categories which are: switch negation, intensifier negation and sentence negation.

i. Switch negation

Switch negation is the common negation method which will reverse the polarity sentiment of a word. For example:

If score for "*happy*" word =2

Switch negation = *"I am not happy"* (not negated happy)

Total sentiment score = 2 x (-1) =-2.

Above example shows the sentiment of "*happy*" word changing from positive to negative with the presence of negation.

University of Malaya

ii. Intensifier negation

Intensifier negation is the presence of negation on next to intensifier. This type of negation does not reverse the polarity of a sentiment but it can reduce the intensified lexicon score. For example:

If score for "*happy*" word =2 and

Percentage value for "*very*" = 25%

Intensifier negation sentence = *"not very happy"*

Total sentiment score = (2 × (100% +25%))/2= 1.25.

According to Collins advanced English dictionary, the use of "*not very*" with an adjective or adverb gives meaning as not completely true or true only to a small degree. This shows the presence of negation in intensifiers will reduce the degree level of sentiment. By considering this, division by two has been finalized for the intensifier negation calculation, which can reduce the degree level of intensified sentiment. The example above shows that the score of "*happy*" word (2) is multiplied with intensifier value (100% +25%). Then the value from the multiplication divided by two. Thus 1.25 is the final score given for this sentence.

iii. Sentence negation

This negation is a bit complicated than the above negations, because this type of negation is present in a sentence but not next to lexicons or intensifier (adjectives or adverb). So, there are possibilities of sentence sentiment to turns into positive sentiment if the actual meaning belongs to negative sentiment. This is because negation words only will be

University of Malaya

calculated if it present next to lexicons or intensifier. By considering this, a general value has been assigned for the sentence negation which is -4. This value was taken from SO-CAL (used for negation calculation) (Taboada et al, 2011) and also finalized by conducted internal testing. For example:

If score for "*good*" word =3

Sentence negation = *"I can't find any good deals"*

Total sentiment score = 3-4(general value for sentence negation) = -1.

The example above shows the sentence having positive lexicon (good) while no negative lexicon is found in this sentence. The sentiment can be calculated as positive polarity, however it does not give the actual meaning of the sentence that is completely described in a negative way. Thus, the general value for negation is assigned (-4) for the sentence calculation and the final scores turn into negative (-1) which defines the actual meaning of the sentence. This calculation is aimed to reduce the degree level of sentiment in a sentence.

University of Malaya

c) Repeated letter

This part is the most important process where the contribution is made in this research. This research is mainly focused on analyzing misspelled words with repeated letters, where the scoring mechanism will identify the lexicon/intensifier in repeated letters' words by following two methods which are looping and stemming methods (as discussed previously). Below are examples showing the score calculation for repetition letters in a word:

Example 1:

If the score for word “*Happy*” is 2,

Value for single repetition letter = 0.5

Sentence A: “*I’m happpppy with your service*”

Score for letter repetition = 3 * 0.5 =1.5

Total Score = 2+ 1.5 = 3.5

Example 2:

If the score for word “*Happy*” is 2, and

Value for single repetition letter = 0.5

Sentence B: “*I’m happppppppy with your service*”

Score for letter repetition = 6 * 0.5 =3.0

Total Score = 2+ 3.0 = 5.0

University of Malaya

The above examples show the difference in number of repetition of letter between two lexicons (that has same meaning) would produce different weight for the sentiment value.

In this research the value for single repetition letter is given as 0.5. This is because, repetition letter is a part of intensification which is responsible to increase or decrease the weight of a sentiment (Muhammad et al., 2015). Generally, intensifier in SO-CAL has different percentage values for each word. However, a percentage value cannot be defined for repetition of letter words as the repetition of letters can occur in any kind of words in a document. A standard value is needed to be assigned to each repetition letter. Thus, 0.5 score has been finalized after performed many internal testing. The sentiment value of 0.5 score will be changed according to the sentiment value of lexicon. For example: if negative lexicon has repeated letters, then the score for each repetition letter is -0.5 and vice versa.

d) Capitalization

For LexiPro-SM, the capitalization score calculation will be the same as SO-CAL calculation, where the capitalized lexicon score will be multiplied by two. For example:

If the score for “*happy*” word =2

Score for “*great*” word =4

Capitalized sentence = *“HAPPY to work with a great person like you”*

Total sentiment score = ((2x2) +4)/2=4.

Above example shows, the sentence contain two positive lexicons where one lexicon is capitalized (“*HAPPY*”) and other is not capitalized (“*great*”). So the capitalized lexicon score will be multiplied by two and that value will sum up with “*great*” score. Finally, the total score will be divided by two.

This division is due to the number of lexicon present in a sentence where it is giving an average value for the whole sentence.

University of Malaya

e) Exclamation Mark

As discussed in Chapter 2, exclamation mark is a part of the intensifier of a sentence. In SO-CAL, the exclamation mark calculation gives a strong strength for the sentiment by multiplying the lexicon value with 2 while it does not give importance to the number of exclamation marks present in a sentence. However, in LexiPro-SM development importance was given to the number of exclamation marks that present in a sentence and a general value score will be assigned for each exclamation mark.

This general value score follows as per repetition letter (as discussed previously). This is because both features are similar in terms of length of character, which can improve the strength of a sentence. Below are the examples that show exclamation mark calculation for SO-CAL and LexiPro-SM.

Example 1 (SO-CAL):

If the score for "*happy*" word =2

Sentence with exclamation mark = *"I am HAPPY with your service!!!!!!!!!!"*

Total score calculation= 2 [score for "happy" word] x 2 [exclamation intensification] = 4.

Example 2 (LexiPro SM):

If the score for "*happy*" word =2

Sentence with exclamation mark = *"I am HAPPY with your service!!!!!!!!!!"*

Total score calculation= 2 [score for "*happy*" word] + (0.5 [exclamation intensification] x 10 [number of exclamation mark]) = 2+5=7.

Above examples shows the same sentence but the total scores are different between SO-CAL and LexiPro-SM. This is because, in LexiPro attention is given to the number of exclamation marks, however in SO-CAL only focuses on the presence of exclamation mark. In LexiPro-SM the number of exclamation marks counted and multiplied with 0.5 score (finalized the score with many internal testing). So the exclamation score will be sum together with the average score of the lexicons present in a sentence forming the total score for the calculation. The sentiment of 0.5 score will be changed according to the sentiment that has highest average score in a sentence. If the sentiment for the

University of Malaya

highest average score is negative, then the score for each exclamation will be changed to -0.5 and vice versa.

f) Auto word correction by PHP spell check

Incorporation of word correction algorithm in LexiPro-SM is one of the efforts to make an improvement in social media text analysis. This feature was incorporated in LexiPro-SM as a means to study the pro and cons of word correction algorithm in social media text analysis. Thus, below is the example of calculation for auto word correction feature:

Auto word correction sentence = *“You are doing a horible job”*

Lexicon detected = “*horrible*”

If before score for “*horrible*” word =-3

Total sentiment score = -3.

The example above shows the sentiment calculation for the sentence “*You are doing a horible job*”, where the misspelled word “*horible*” was corrected automatically as “horrible” and a sentiment score calculated for the sentence is “-3” instead of zero sentiment score.

University of Malaya

3.7.5 Features comparison: LexiPro-SM vs SO-CAL

Table 3.3 shows the features comparison between LexiPro-SM and SO-CAL. In LexiPro-SM only one feature was solely inherited from SO-CAL which is capitalization. Improvement was made on two features which are intensifier and negation. The improvement in intensifier calculation is dealing with sentiment word that has repeated intensifier words in a sentence such as "*I am very very happy*". This is because in SO-CAL, the intensifier feature is only giving scores for a sentiment word that has one intensifier word in a sentence. The improvement in negation calculation was made by merging additional methods (switch negation and intensifier negation) with shift negation method. However, the shift negation method in LexiPro-SM will be called as sentence negation as this method only applied for negation that present in a sentence but not next to lexicons or intensifier (adjectives or adverb) such as *"I can't find any good deals"*. Besides, one feature in LexiPro-SM is not inherited from SO-CAL, that is exclamation mark. For exclamation mark calculation, attention given to the number of exclamation marks but in SO-CAL it only focuses on the presence of exclamation mark. The modification or improvement made on these features is to enhance the degree level of the sentiment presents in a sentence. Lastly, there are two new features included in LexiPro-SM but not incorporated with SO-CAL, which are repeated letter and auto word correction.

University of Malaya

Table 3.3: Features comparison: LexiPro-SM vs SO-CAL

Features	LexiPro-SM	SO-CAL
Capitalization	Solely inherited from SO-CAL	The value of capitalized sentiment word will be multiplied by 2
Intensifier	Inherited from SO-CAL. However, improvement made on calculation by dealing with repeated intensifier word.	Calculation made for only one intensifier word for a sentiment word
Negation	Using three types of negation: Switch, Intensifier and Sentence (inherited from SO-CAL).	Shift negation that using specific value (as 4/-4) for negation calculation.
Exclamation Marks	Attention is given to the number of exclamation marks	Focuses on the presence of exclamation mark
Repeated Letter	Giving scores for the sentiment word that contains repeated letter	Not applicable
Auto word correction by PHP spell check	Auto-correct misspelled word that relevant to sentiment word and gives scores	Not applicable

University of Malaya

In LexiPro-SM the score calculation can be divided into three stages:

At first stage of score calculation, the calculation involves incorporation of all features except exclamation mark. Besides, an individual average score will be calculated for the positive and negative sentiment.

In second stage of score calculation, a comparison will be made between average score of positive and negative sentiments. If the average score of negative is higher than positive score, the exclamation scores will be sum together with average score of negative and formed a final score for the negative sentiment and vice versa.

Once the individual final score is formed for both positive and negative sentiments, the difference will be calculated between both sentiments. This difference value will be assigned as a total sentiment score for a text/sentence.

The last stage of this score calculation is categorization of the total sentiment score into a polarity group from -5 (most negative) to +5 (most positive); zero value is assigned to neutral. This categorization will be used to analyze the overall results of analysis in Chapter 5.

There is another function working together in a parallel manner with the scoring process. The purpose of this function is to categorize each text/sentence into four groups that represent the airline industry services. The four groups are “customer service”, “price”, “pre-flight”, and “facility”. However, the text/sentence that do not meet the criteria of any groups (mentioned above) will be categorized under “other” group.

The main aim of this service categorization is to help the service management in the airline industry to narrow down their attention to services that require improvements. The process of categorization will be worked by checking each chunked word into four difference arrays that contain a list of keyword that is related to each category. For example: if the word present in the “price” array, then the sentence will be categorized under the “price” group.

Once the text/sentence has been analyzed, this information will be stored in a phpMyadmin database, where the table contains post id, cleaned text, total sentiment score, polarity group and service categories such as customer service, price, preflight, facility and others. This table will save all the processed data for final results of the data analysis. Results and evaluations will be discussed further in Chapter 5.

University of Malaya

3.8 Evaluation planning

In this section, the evaluation planning and effectiveness of LexiPro-SM is discussed. This planning is also important because the second objective of this research is to evaluate the enhanced scoring mechanism by comparing it with an existing system, where SO-CAL will be used to perform result comparison with LexiPro-SM. Therefore, the planning for the evaluation follows as per the steps below:

i) Analyze all clean data using LexiPro-SM and SO-CAL

ii) Select 300 data randomly from the total processed sample which has both tool scores

iii) Perform human expert analysis

iv) Generalize the human experts results

v) Perform evaluation measures with evaluation metrics

Above steps shows the method to conduct evaluation on LexiPro-SM. First, sentiment analysis will be performed for the finalized clean data by using LexiPro-SM and SO-CAL and the results (sentiment scores) will be recorded. Next, 300 randomly selected data will be sent for human expert analysis. The human experts (linguistic experts) consulted for verification purpose are from education, linguistic and the information technology background. The expert role is to manually analyze posts and determine the sentiment without the help of any tools. For example if a data belongs to a positive sentiment, the human expert will classify the data as “positive” instead of assigning score for it. Once the human experts are done with the analysis, all three sets of results will be collected and comparison made between the results. A finalized result will be generated (assigned most define sentiment) based on the three results. The finalized human expert result will be used as trained data to perform evaluation measures by using the evaluation metrics (discussed in following section). Hence, from the evaluation a measurement of the effectiveness of LexiPro-SM can be determined.

University of Malaya

3.9 Evaluation Metrics

Evaluation metrics are the methods that are used to measure the capabilities and effectiveness of a tool (Moraes et al., 2013; Yeet al., 2009). In this research, there are four evaluation metrics applied, which are accuracy, precision, recall and F1-Score measure. Following are the explanation for each of the evaluation metrics:

a) Accuracy

Accuracy is to determine the overall performance of a sentiment analysis (SA) tool where it defines how accurate the results produced.

Accuracy = total count well classified (positive + negative + neutral) / total data

Accuracy of a tool can be identified by calculating the division of: the total count of data that match the human expert results by the total data of human expert results (which is 300 data set used for this research).

b) Precision

Precision measures how many of human expert data are correctly matched with the total data of a particular sentiment generated by an SA tool.

Precision = well classified count / (well classified count + bad classified count)

In this calculation, the precision will be calculated by dividing the total count of data that match the human expert results (only for a single type of data) by the total count of data that is classified by a tool (same type of data).

For example:

$$\text{Positive precision} = \frac{\text{Total positive data that match with human expert result}}{\text{Total positive data that produced by a tool}}$$

This example shows the precision formula for one type of data which is positive.

University of Malaya

c) Recall

Recall measures how many posts using SA tool correctly matches the total data of a particular sentiment determined by human expert data.

$$\boxed{\text{Recall} \quad = \quad \text{well classified count / human expert analyzed count}}$$

In this calculation the recall will be calculated by dividing the total count of data that match the human expert results (only for one type of data) by the total count of human expert results data (same type of data):

$$\boxed{\text{Positive recall} = \frac{\text{Total positive data that match with human expert result}}{\text{Total positive data of human expert result}}}$$

This example shows the recall formula for one type of data which is positive.

d) F1-Score measure

Tools that have high precision value tend to have low recall and vice versa. Thus, F1-Score measure used to calculate the combination of precision and recall to produce a final result that defines the effectiveness of a SA tool.

$$\boxed{\text{F1-Score} = \quad 2\ (\text{Recall x Precision}) / (\text{Recall} + \text{Precision})}$$

The calculation will be made by multiplying the value obtain from multiplication between recall and precision, by two, and then the resulting value is divided by the sum of recall and precession.

Additionally, the performance of the proposed system with SO-CAL was tested for its significant difference using pairwise comparisons. SPSS version 23 was used for this purpose, with an alpha level of 0.05.

University of Malaya

3.10 Summary

In summary, chapter 3 discussed the planning and steps involved in building the proposed scoring mechanism of LexiPro-SM. The system methodology deployed in this research is Rapid Application Development (RAD). The programing language used to develop LexiPro-SM is PHP. The operational framework of this research was also discussed.

Data collection was performed by collecting data from Facebook pages of two major airlines (Airline A and Airline B) with the aid of Graph API (PHP SDKs) and stored in database. Then, data cleaning was performed by removing all the noises that is irrelevant to this research and where the clean data will only contain alphabets and exclamation mark. Finally, the clean data will be used for data analysis activity in order to determine the sentiment score and categorized score based on the polarity range.

Explanations given for the features will incorporated in LexiPro-SM such as dictionaries, negation, intensifier, exclamation mark, capitalization, repeated letter and auto word corrector.

Furthermore, a parallel activity was performed together with the scoring process, where the data was categorized into four groups of airline services (customer service, price, pre-flight and facilities) which would be helpful for airline service improvement. The irrelevant data will be grouped under "other" category. All raw data, clean data and processed data will be stored in a database consisting of separate tables for each airline.

Finally, evaluation planning was discussed, together with the use of evaluation metrics to determine the effectiveness of LexiPro-SM.

University of Malaya

CHAPTER 4: IMPLEMENTATION AND TESTING

4.1 Introduction

This chapter describes the activities involved in the system design phase of LexiPro-SM, followed by system implementation. Besides, an explanation is given on the testing performed with LexiPro-SM, identifying the effectiveness as compared against the existing system. Finally, a description is given on the web-based portal for LexiPro-SM analysis.

4.2 System design

System design is a process to design the characteristic of a system defined by methods, functions or procedures (Waldo, 2006). System design also aid in understanding the overall process from a single page view. In this research, there are three types of designs to showcase the overall process involved in LexiPro-SM, which are Use case diagram, Entity-relationship diagram (ERD) and Data flow diagram (DFD). The following sub-sections were explained each of the design with a diagram.

University of Malaya

4.2.1 Use case diagram

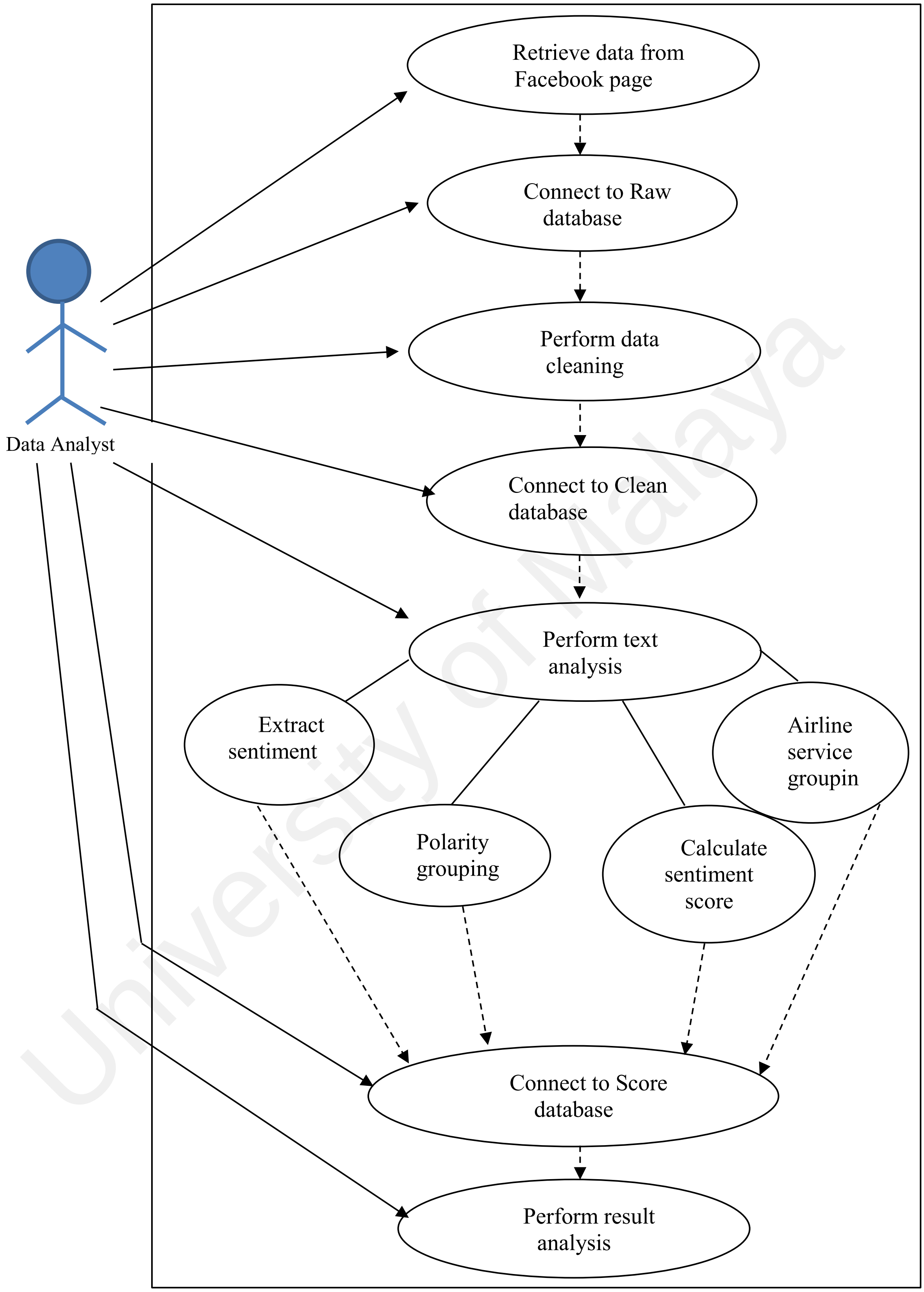


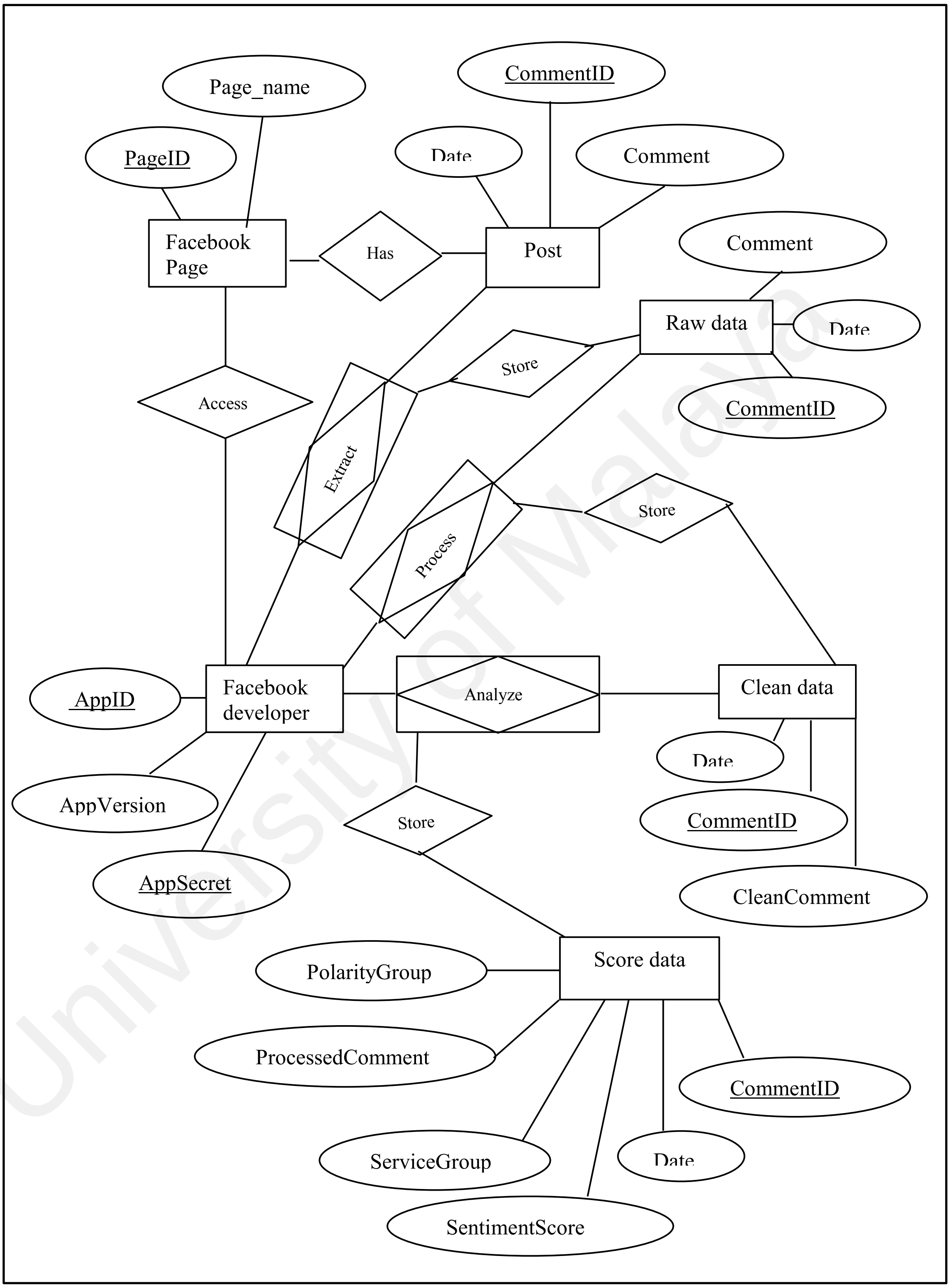
Figure 4.1: Use Case Diagram for LexiPro-SM

University of Malaya

Figure 4.1 shows the activities of an actor in LexiPRO-SM sentiment analysis. This actor is addressed as a data analyst who is responsible to collect data and perform sentiment analysis. The process starts with the data analyst retrieving data from a Facebook page and storing it in a raw database. Then data cleaning is performed on the raw data and a clean database is used to store the processed data. Next, the data analyst performs text analysis on the clean data where it involves extracting sentiment, calculating sentiment score, polarity grouping and airline service grouping. The processed data from the text analysis will be stored in a score database. Finally, the data analyst analyzes the score data in the score database to identify the overall results of the data analysis.

University of Malaya

4.2.2 Entity-relationship diagram (ERD)



University of Malaya

Figure 4.2: Entity Relationship Diagram (ERD) for LexiPRO-SM

Figure 4.2 shows the relationship between entity and database. The relationship starts from Facebook developer who uses AppID, AppSecret and AppVersion to access the Facebook page that belongs to a specific PageID . The Facebook page contains posts where each post has its own CommentId, date and comment. With Facebook developer access, the posts were extracted and stored in raw database that has CommentId, date and Comment columns. Cleaning was performed on the data from the raw database and stored the processed data in clean database that contain CommentId, date and CleanComment columns. Finally, sentiment analysis was performed on the data from the clean database and the results stored in the score database that has PolarityGroup, ProcessedComment, ServiceGroup, SentimentScore, Date, CommentID as columns.

University of Malaya

4.2.3 Data flow diagram (DFD)

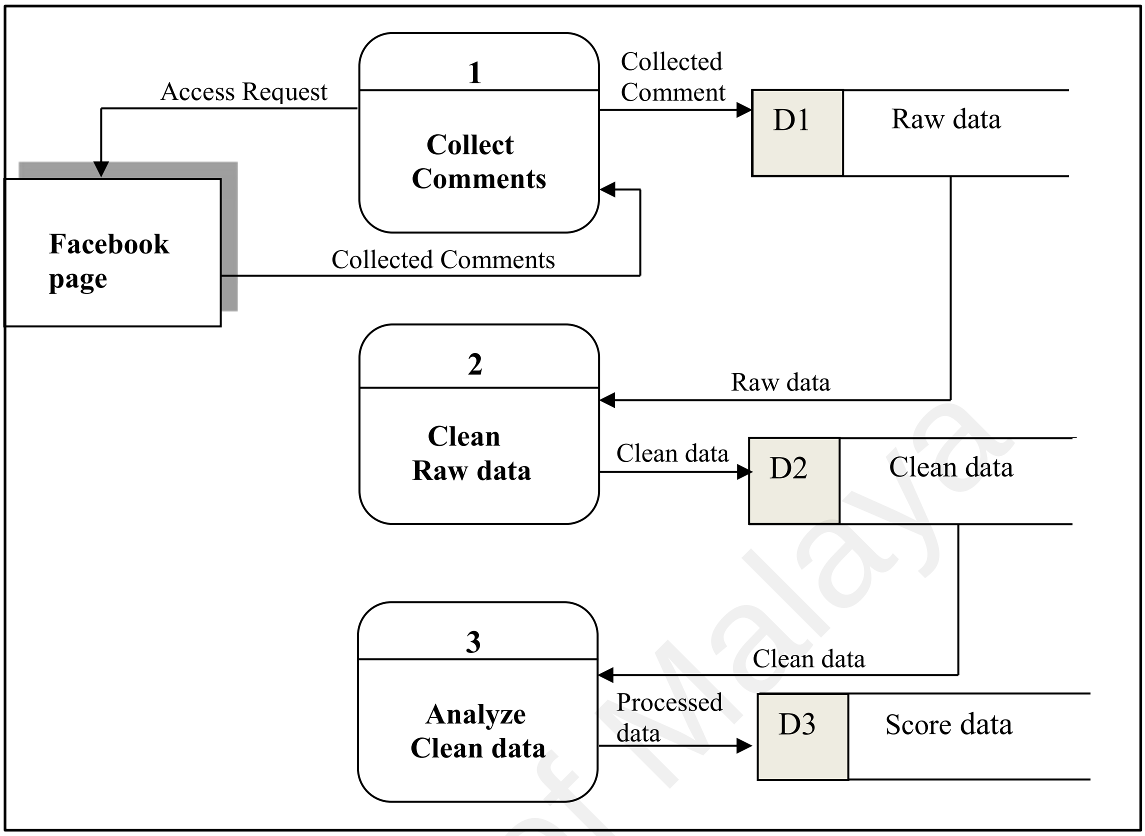


Figure 4.3: Data flow diagram (ERD) for LexiPRO-SM

Figure 4.3 shows how data flows between the process and database. The crucial processes involved in this sentiment analysis are collecting comments, raw data cleaning and analyzing the clean data. Three different types of databases were incorporated in this analysis, namely raw database, clean database, and score database. The activity of data flow starts with comments collecting process, where a request is sent to gain access into the Facebook page. Once the access is granted, the comments from the Facebook page will be extracted and stored in a raw database. Then the data cleaning process is carried out, where the raw data will undergo a cleaning process and the clean data will be stored in the clean database. Finally, the process continues with the analysis of clean data, this is to identify the sentiments present in the data and give scores for each data. Once the clean data is analyzed, the resulting data is stored in the score database.

University of Malaya

4.3 System implementation

In this section, the implementation of LexiPRO-SM is discussed. The implementation phase is divided into three parts which are data collection, data cleaning and data analysis. The following sub-sections provide a detailed explanation of the implementation.

4.3.1 Overall framework of LexiPro-SM analysis

Figure 4.4 shows the overall framework of LexiPro-SM analysis. This starts from selecting an unprocessed clean text from the clean data database and chunk it into words. Then, each word will be analyzed individually and the given scores will be averaged after the last chunked word in the text has been processed.

Once the text has been chunked, the word will be checked in English dictionary. If the word is present in English dictionary it will be treated as a correct word else will be treated as a misspelled word. The correct word will then be checked in the lexicon dictionary and the misspelled word will be checked in acronym dictionary.

If the correct word is present in the lexicon dictionary, the sentiment score will be given for the word. Scores calculated for each word will be stored in an array which will be used for the total score calculation on the end of data analysis. If the correct word is not present in the lexicon dictionary, it will be treated as non-lexicon word which will bear neutral value as zero.

For each misspelled word present in the acronym dictionary, a score will be given for the word and stored in the same array as discussed above. However, if the word is not present in the acronym dictionary, it will be brought to the next level of checking. In this checking, the misspelled word will be checked as it contains repetition letters or not. If repetition of letter is found, it will be analyzed by repetition letter methods (looping and stemming method) else the misspelled word will be checked under auto word correction program.

University of Malaya

Two methods were applied on repetition letters analysis which are looping and stemming methods. The looping method in this research is defined as doing loop check on the repetition letter word by removing one repeated letter from the word at a time and the new word will be checked in lexicon dictionary. If it is present in lexicon dictionary, scores will be given (with the sum of repetition letters score) else the looping will be continued until the last repetition letter and no score will be given if there is no lexicon identified until the end of the looping.

The stemming method in this research is defined as replacing all repetition letters into one letter. For example: if the repetition letters word is "bessstttt" it will be replaced as "best". Then, the stemmed word will be checked in lexicon dictionary. If it is present, scores will be given (with the sum of repetition letters score) or else no score will be given.

Looping method is only applied for the repetition letters word that has only one type of repeated letter, for example the word "badddddd" has only one type of repeated letter: "d". However, the stemming method is applied for the repetition letters word that has more than one type of repeated letter for example the word "besssttt" which has two types of repeated letters: "s" and "t".

Looping method is not suitable for the word that has more than one type of repeated letter. This is because, the presence of more than one type of repeated letters will increase the checking logics which indirectly increase the looping complexity. This can lead to more time consumption to process a text which results in poor performance. Figure 3.10 illustrate the algorithm used for repeated letter analysis.

Besides, if the misspelled word does not contain repetition letters, this word will be checked under an auto word correction program known as PHP SpellCheck. This is a free source program developed by Digital River Inc. and it is available online (http://www.phpspellcheck.com/). The auto word correction algorithm works by replacing the misspelled word with the most relevant word. For example, the word

University of Malaya

“fantstic” will be replaced to “fantastic”. Then the replaced word will be checked in lexicon dictionary, if it is present in lexicon dictionary, scores will be given or else the misspelled word will be treated as non-lexicon word with a neutral value. The incorporation of auto word correction is one of the attempts in this research, to identify the meaning hidden in misspelled word which would help to improve the strength of a sentiment.

University of Malaya

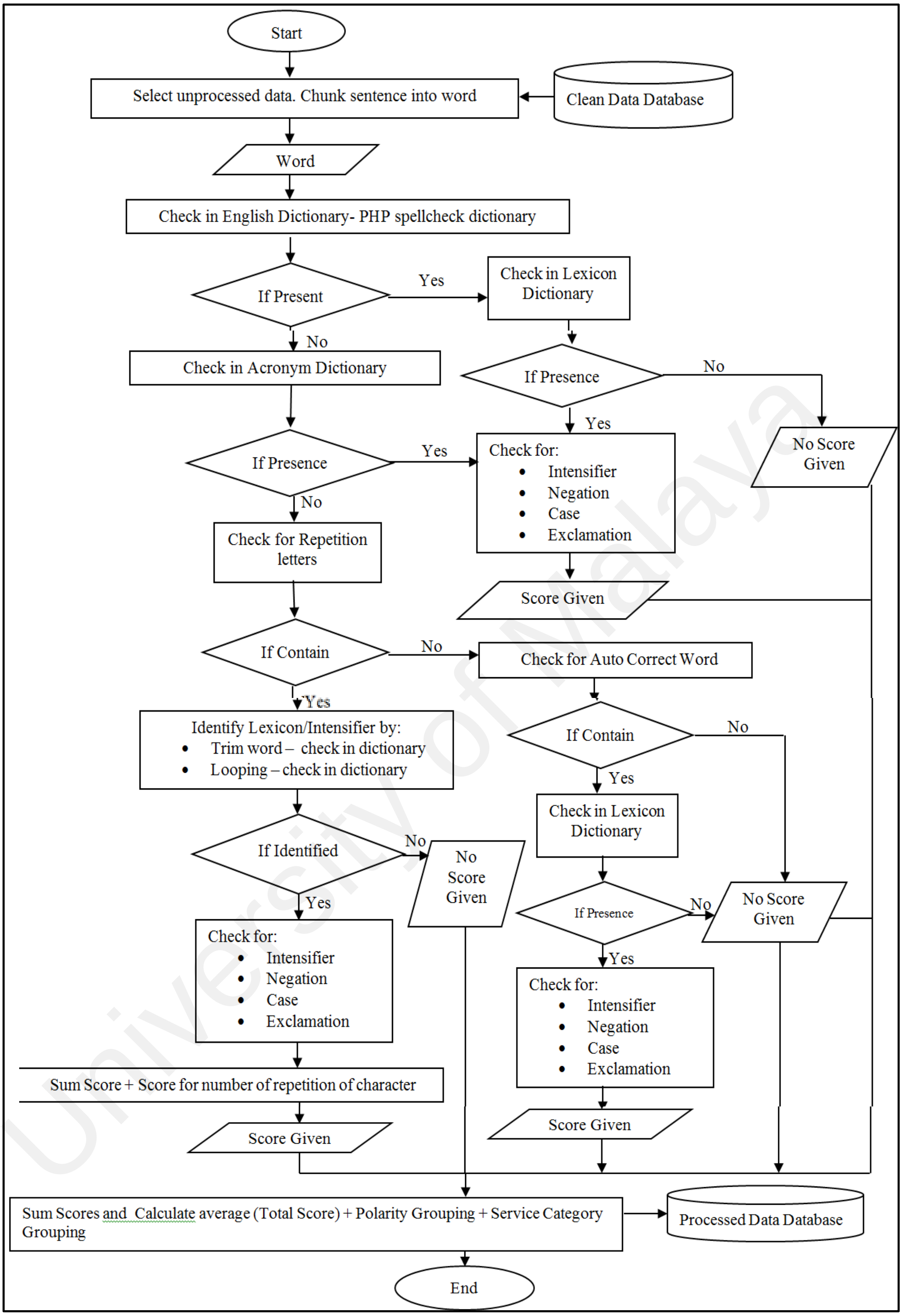

Start
Select unprocessed data. Chunk sentence into word
Clean Data Database
Word
Check in English Dictionary- PHP spellcheck dictionary
If Present
Yes
Check in Lexicon Dictionary
No
Check in Acronym Dictionary
If Presence
No
No Score Given
Yes
If Presence
Yes
Check for:
Intensifier
Negation
Case
Exclamation
No
Check for Repetition letters
Score Given
If Contain
No
Check for Auto Correct Word
Yes
Identify Lexicon/Intensifier by:
Trim word – check in dictionary
Looping – check in dictionary
If Contain
No
Yes
Check in Lexicon Dictionary
If Identified
No
No Score Given
Yes
If Presence
No
No Score Given
Check for:
Intensifier
Negation
Case
Exclamation
Yes
Check for:
Intensifier
Negation
Case
Exclamation
Sum Score + Score for number of repetition of character
Score Given
Score Given
Sum Scores and Calculate average (Total Score) + Polarity Grouping + Service Category Grouping
Processed Data Database
End

Figure 4.4: Overall framework of Data Analysis

University of Malaya

```
1. Start
2. Chunk sentence into word
3. Loop start-each word
   3.1  Check word in English dictionary
   3.2  If present
   3.3  Give score
   3.4  Else
   3.5  Check for repeated letters
   3.6  If have
   3.7  Count for type of repeated letter
        3.7.1   if count =1
        3.7.2   Use loop method and check word in lexicon dictionary
                3.7.2.1 If present
                3.7.2.2 Calculate score with length of repeated letter
                3.7.2.3 Else
                3.7.2.4 Give no score
        3.7.3   Else
        3.7.4   Use stem method and check word in lexicon dictionary
                3.7.4.1 If present
                3.7.4.2 Calculate score with length of repeated letter
                3.7.4.3 Else
                3.7.4.4 Give no score
   3.8  Else
   3.9  Check word in auto word corrector
   3.10 If have
   3.11 Give score
   3.12 Else
   3.13 Give no score
4. End loop
5. Calculate total sentiment score for sentence
6. Store in database
7. End
```

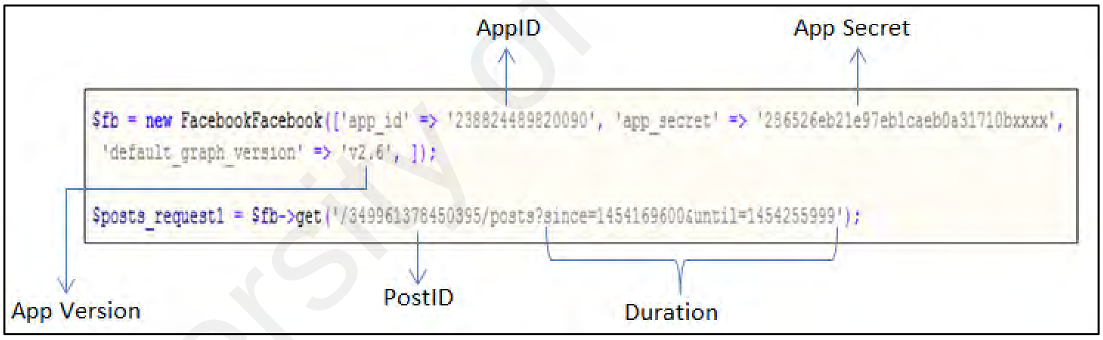
Figure 4.5: Repeated letter algorithms

Figure 4.5 is showing the algorithm to detect repeated letters in a word and assigned a score if the word present in lexicon dictionary.

University of Malaya

4.3.2 Implementation of data collection

As discussed in Chapter 3, all the data for this research were collected from Facebook with the aid of Graph API PHP SDKs. In order to extract data from Facebook, the user should have access to the Facebook developer page and it can be obtained by completing a simple registration. This Facebook developer page contains three types of details that will be used as a credential to access the Facebook data via the scripting work. The details are API Version, AppID and App Secret. Besides the three details, another detail must be included in scripting work that give access to the data of a particular page. This detail will be called PageID. Each group or individual page has its own PageID which is unique to other PageIDs.



```
$fb = new FacebookFacebook(['app_id' => '238824489820090', 'app_secret' => '286526eb21e97eb1caeb0a31710bxxxx',
 'default_graph_version' => 'v2.6', ]);

$posts_request1 = $fb->get('/349961378450395/posts?since=1454169600&until=1454255999');
```

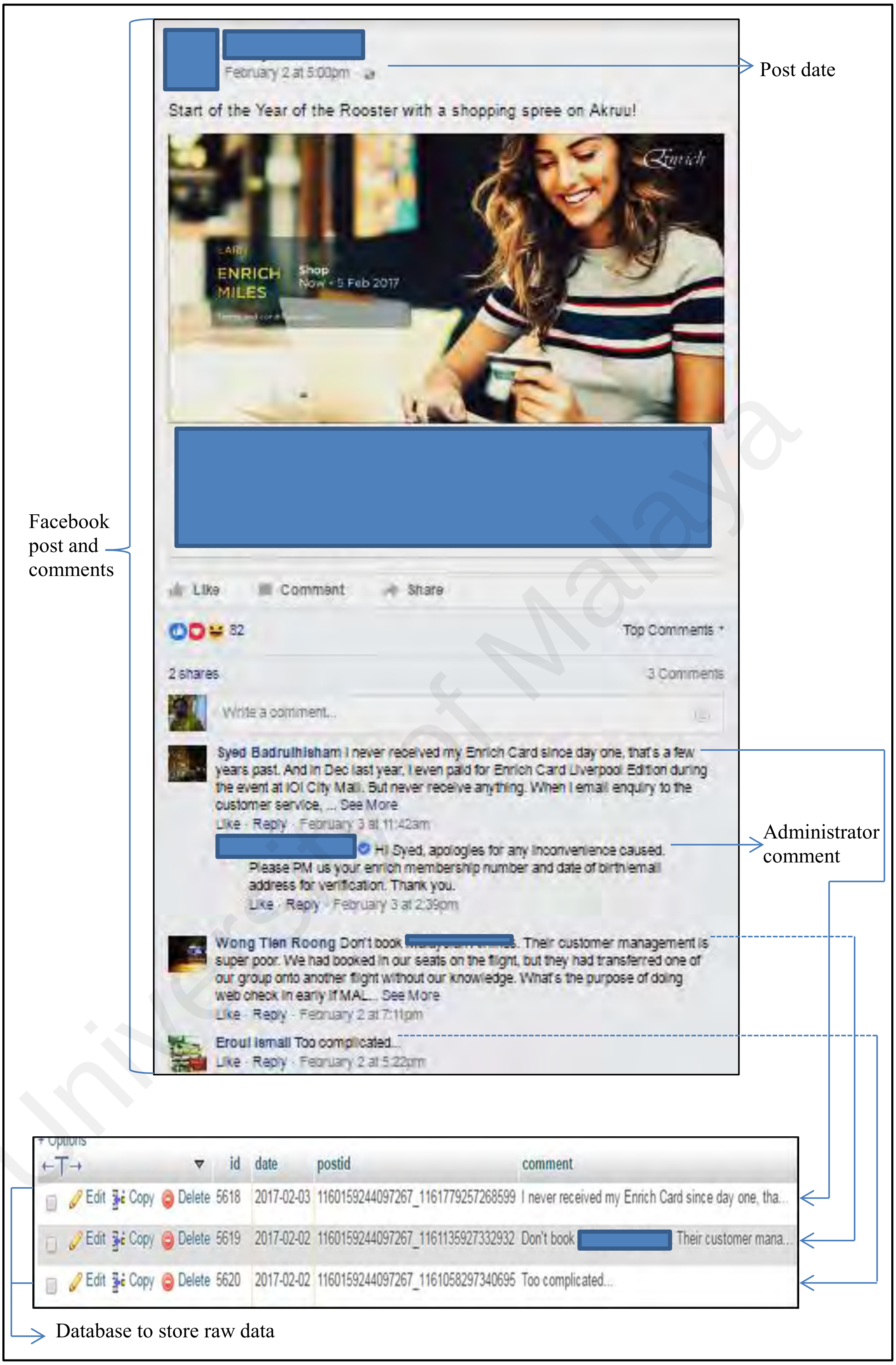
Figure 4.6: Important API References to access Facebook data

Figure 4.6 shows the API references used in scripting work to access the Facebook data. Besides, it has also included information as “Duration”. This used to extract data in a particular duration, where the duration will be in the form of timestamp. For example:

“$fb->get('/349961378450395/posts?since=1454169600&until=1454255999')”,

This section involves obtaining the posts of Airline A page from 30th of January 2016 (4.00 pm) to 31st of January 2016 (3.59 pm).

University of Malaya



[redacted]

[redacted]

February 2 at 5:00pm

Start of the Year of the Rooster with a shopping spree on Akruu!

[redacted]

Like Comment Share

82 Top Comments

2 shares 3 Comments

Write a comment...

Syed Badrulhisham I never received my Enrich Card since day one, that's a few years past. And in Dec last year, I even paid for Enrich Card Liverpool Edition during the event at IOI City Mall. But never receive anything. When I email enquiry to the customer service, ... See More

Like · Reply · February 3 at 11:42am

[redacted] Hi Syed, apologies for any inconvenience caused. Please PM us your enrich membership number and date of birth/email address for verification. Thank you.

Like · Reply · February 3 at 2:39pm

Wong Tien Roong Don't book [redacted]. Their customer management is super poor. We had booked in our seats on the flight, but they had transferred one of our group onto another flight without our knowledge. What's the purpose of doing web check in early if MAL... See More

Like · Reply · February 2 at 7:11pm

Eroul Ismail Too complicated...

Like · Reply · February 2 at 5:22pm

+ Options

←T→	▼	id	date	postid	comment
□	Edit Copy Delete	5618	2017-02-03	1160159244097267_1161779257268599	I never received my Enrich Card since day one, tha...
□	Edit Copy Delete	5619	2017-02-02	1160159244097267_1161135927332932	Don't book [redacted] Their customer mana...
□	Edit Copy Delete	5620	2017-02-02	1160159244097267_1161058297340695	Too complicated...

Post date

Facebook post and comments

Administrator comment

Database to store raw data

Figure 4.7: Sample of Facebook post and data storing

University of Malaya

Figure 4.7 shows an example of post's image that was taken from Airline A page and the data storing in database. In Facebook, each page has its own administrator who is managing the posts, pictures and other features of their page. An administrator is someone who is able to post information and they can participate in the comment conversation as well. However, a user is only allowed to give comments under the post. These comments will be extracted (along with comment ID and comment date) and used for sentiment analysis. The post date will be used in API reference to setup the duration. In some comment conversation, there were comments posted by the administrator. These comments are irrelevant for data analysis as the research only required public's opinion towards the organization. Thus, during the data extraction, the administrator's comments will be ignored and only public comments will be stored in database.

The extraction of the Facebook comments is linked to database, where all comments were grabbed from the Facebook page, and directly stored in database. These comments will be called as "raw data". The raw data was collected and stored separately for each airline, where the table name for Airline A is "rawdataA" and for Airline B is "rawdataB". The table of raw data has three important columns, which are "postid" (to store comment's id), "date" (to store the date of the comment) and "comment" (to store the comment).

University of Malaya

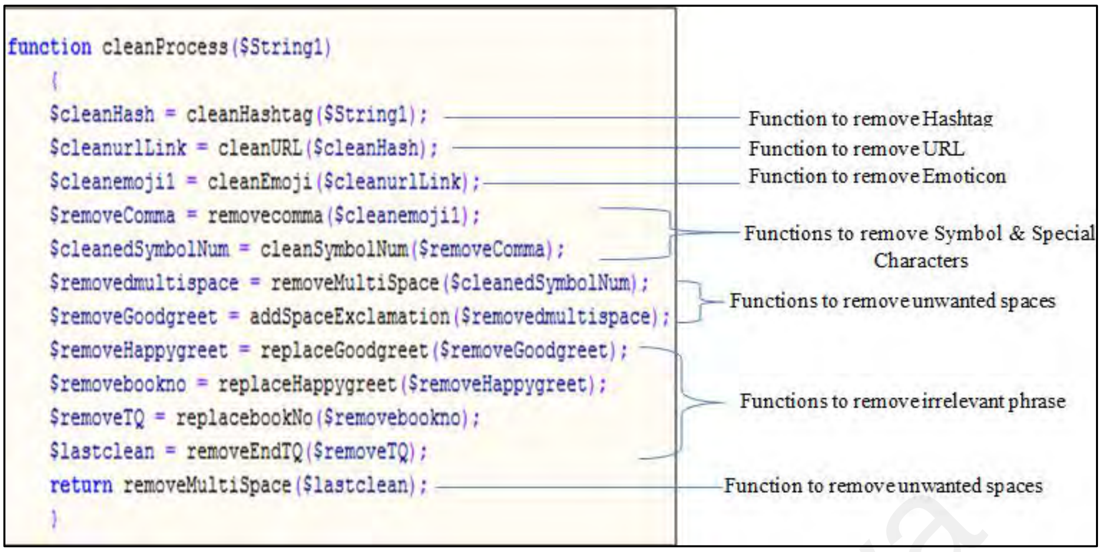
4.3.3 Implementation of data cleaning

As discussed in Chapter 3, data cleaning is the process to remove irrelevant data (which is known as noises) from the raw data. The noises were determined based on the research requirements. The noises in this research were narrowed down to URL link, emoticon, numeric characters, special characters, Malay words and irrelevant phrases and every symbol except exclamation mark. The scripting work for data cleaning process was built on a function basis, which means each noises has its own function to process the data. So the raw data will be passing through all the functions and finally will obtain the clean data as an output. As mentioned in previous chapter, the clean data for this research will be in a form of alphabets or alphabets with exclamation mark symbol.

Figure 4.8 shows all the functions to remove the noises that were included in one main function which is referred to as “cleanProcess” function. This function will call the raw data and it will pass through all the cleaning functions and finally return data that remain alphabets and exclamation mark.

Besides the “cleanProcess” function, other functions were included to remove unwanted spaces in data. This functions are applied so removing noises in data will result in more space that may otherwise increase the size of the data. Thus, these functions reduce the size of data by removing unwanted spaces.

University of Malaya



```
function cleanProcess($String1)
    {
    $cleanHash = cleanHashtag($String1);
    $cleanurlLink = cleanURL($cleanHash);
    $cleanemoji1 = cleanEmoji($cleanurlLink);
    $removeComma = removecomma($cleanemoji1);
    $cleanedSymbolNum = cleanSymbolNum($removeComma);
    $removedmultispace = removeMultiSpace($cleanedSymbolNum);
    $removeGoodgreet = addSpaceExclamation($removedmultispace);
    $removeHappygreet = replaceGoodgreet($removeGoodgreet);
    $removebookno = replaceHappygreet($removeHappygreet);
    $removeTQ = replacebookNo($removebookno);
    $lastclean = removeEndTQ($removeTQ);
    return removeMultiSpace($lastclean);
    }
```

Function to remove Hashtag
Function to remove URL
Function to remove Emoticon
Functions to remove Symbol & Special Characters
Functions to remove unwanted spaces
Functions to remove irrelevant phrase
Function to remove unwanted spaces

Figure 4.8: Functions for data cleaning

University of Malaya

Once the “cleanProcess” function has processed the raw data, it will return processed data that contain alphabet and exclamation mark. However, this processed data will pass through another stage of cleaning which is the removing of Malay language. This stage can be divided in to two parts, where the first part the whole data is removed /deleted if the count of Malay word exceeds 50% of the total count of words (in a data). Second, if the Malay word count less than 50% of the total count of words (in a data), all the Malay words will be removed from the data. This is because through observation during data extraction, the data that has more than 50% Malay words may not be relevant to the data analysis (example 2). Thus, the whole data will be removed and concentration will be given to data that has less than 50% Malay words. Below are the two samples of data that will be used to show the cleaning process, where raw data 1 contains all the noises and Malay words that are less than 50% count. Then, raw data 2 contain Malay words more than 50% count.

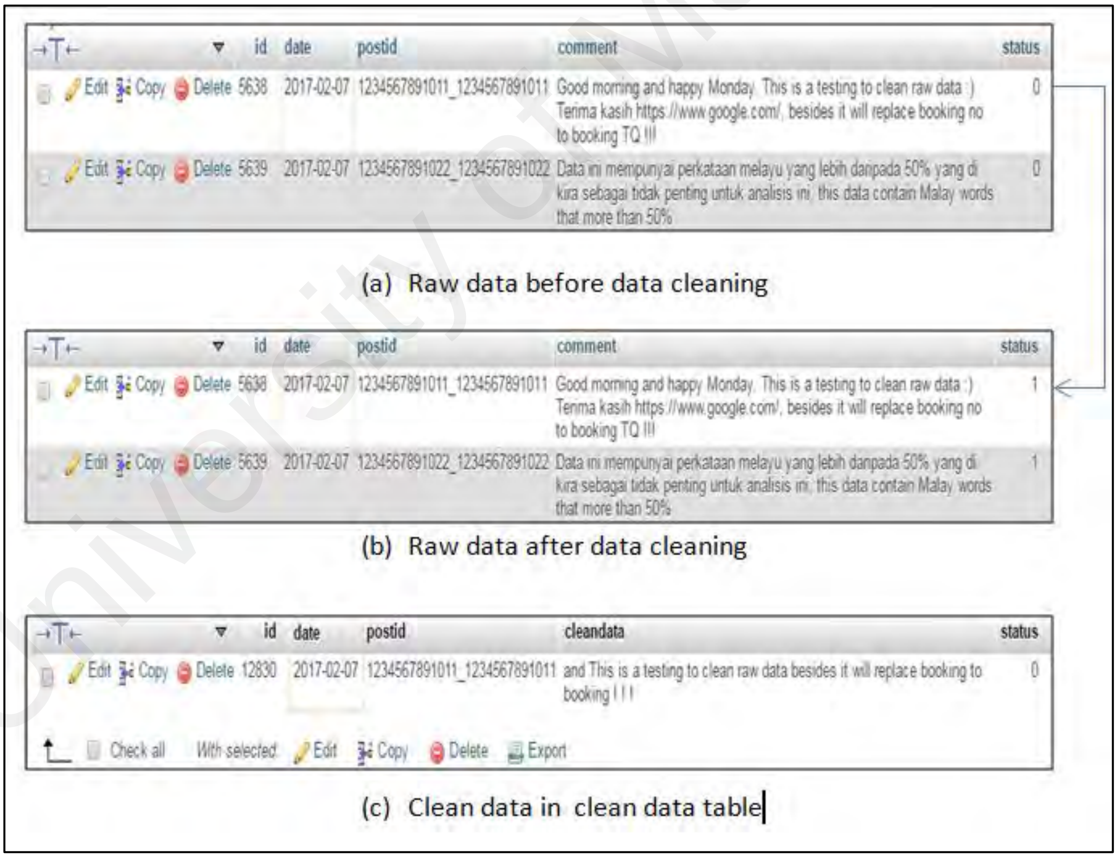
For example:

Example 1 (raw data 1):

"Good morning and happy Monday. This is a testing to clean raw data :) Terima kasih https://www.google.com/, besides it will replace booking no to booking TQ !!!"

Example 2 (raw data 2):

"*Data ini mempunyai perkataan melayu yang lebih daripada 50% yang di kira sebagai tidak penting untuk analisis ini, this data contain Malay words that more than 50%"*



→T←	id	date	postid	comment	status
□ Edit Copy Delete	5638	2017-02-07	1234567891011_1234567891011	Good morning and happy Monday. This is a testing to clean raw data :) Terima kasih https://www.google.com/, besides it will replace booking no to booking TQ !!!	0
□ Edit Copy Delete	5639	2017-02-07	1234567891022_1234567891022	Data ini mempunyai perkataan melayu yang lebih daripada 50% yang di kira sebagai tidak penting untuk analisis ini, this data contain Malay words that more than 50%	0

(a) Raw data before data cleaning

→T←	id	date	postid	comment	status
□ Edit Copy Delete	5638	2017-02-07	1234567891011_1234567891011	Good morning and happy Monday. This is a testing to clean raw data :) Terima kasih https://www.google.com/, besides it will replace booking no to booking TQ !!!	1
□ Edit Copy Delete	5639	2017-02-07	1234567891022_1234567891022	Data ini mempunyai perkataan melayu yang lebih daripada 50% yang di kira sebagai tidak penting untuk analisis ini, this data contain Malay words that more than 50%	1

(b) Raw data after data cleaning

→T←	id	date	postid	cleandata	status
□ Edit Copy Delete	12830	2017-02-07	1234567891011_1234567891011	and This is a testing to clean raw data besides it will replace booking to booking ! ! !	0

(c) Clean data in clean data table

Figure 4.9: Example of data cleaning results

In Figure 4.9(a) and (b) are raw data table, then (c) is clean data table. The two raw data sample were inserted into raw data table and data cleaning was performed. In the cleaning process, once the process completed the clean data will be inserted into clean

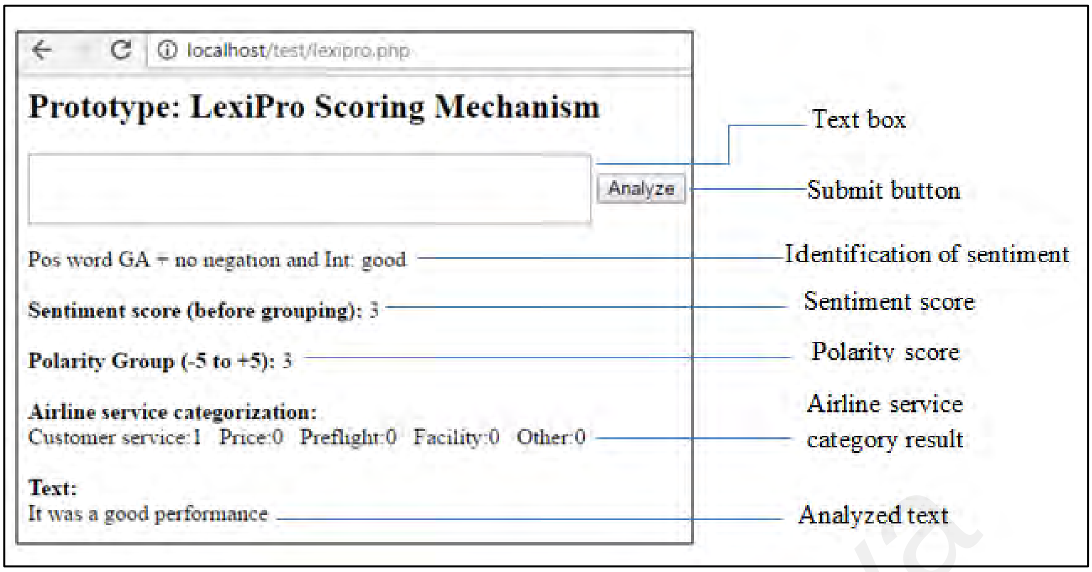
University of Malaya

data table and the data in raw data table will be updated as processed by changing the "status" value from 0 to 1 asp Figure 4.9 (b). The "status" column is used for internal checking purpose, this to avoid data duplication during the data insertion process, where 0 indicates not processed and 1 indicates as processed. So from the Figure 4.9 above it shows that all the noises were removed from raw data 1 and inserted into clean data table as ("and This is a testing to clean raw data besides it will replace booking to booking ! ! !") (see Figure 4.9(c)). Then the raw data 2 was ignored as it contains Malay words greater than 50%. The following figures show the interface of clean data table for both airlines where "cleandataA"for Airline A and "cleandataB" is for Airline B.

4.3.4 Implementation of data analysis

In this research, no system was developed to perform the sentiment analysis. However, it is only involves scripting work that is linked to database where it is able to process data and store the sentiment scores in database. In order to visualize the implementation of data analysis towards the enhanced scoring mechanism, a web-based prototype was developed with PHP language. This prototype will be called LexiPro-SM prototype. The function of LexiPro-SM prototype is to perform single data analysis and shows the sentiment results as an output of the analysis. The main purpose to develop this prototype is to identify the effectiveness of LexiPRO-SM by doing testing and making comparison with an existing system.

University of Malaya



localhost/test/lexipro.php

Prototype: LexiPro Scoring Mechanism

Pos word GA + no negation and Int: good

Sentiment score (before grouping): 3

Polarity Group (-5 to +5): 3

Airline service categorization:
Customer service:1 Price:0 Preflight:0 Facility:0 Other:0

Text:
It was a good performance

Figure 4.10: Interface of LexiPro-SM prototype

Figure 4.10 shows a simple design of LexiPro-SM prototype. The text will be analyzed by placing a clean text in the textbox and click the "Analyze" button. Once the "Analyze" button is clicked, it will populate the sentiment result in three categories, which are sentiment score, polarity group result and airline service category result. The sentiment score represents the total sentiment scores that is calculated for a text.

Polarity group is the categorization of sentiment score in a scale of (-5 to +5) whereby -5 is the most negative, +5 is most positive and zero represent neutral value. This polarity group result will be used in data evaluation in Chapter 5. The final result is for airline service categorization. This is to categorize the text into five groups which are customer service, price, preflight, facility and others. The result for airline categorization is indicates as "0 or 1", where "0" is not in category and "1" is in category. Besides there are other information added to the output of the analysis, which is the identification of sentiments that shows the details of lexicon present in a text such as "*Pos word + no negation and Int: good*" and the text that was used to analyze to obtain the sentiment result (figure 4.10).

University of Malaya

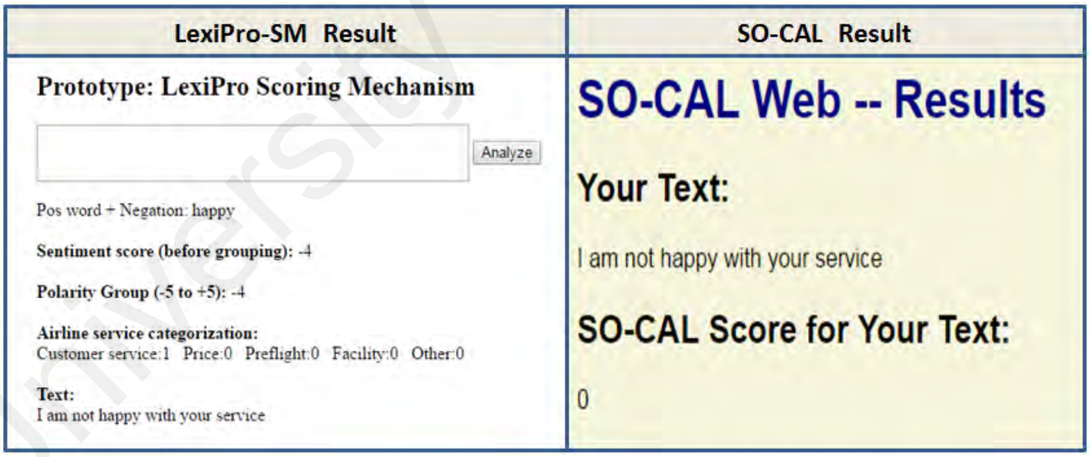
4.4 System testing

In this section, testing was performed on LexiPro-SM prototype and SO-CAL, where the importance were given to the key features that were improved in LexiPro-SM such as negation, repeated letter, exclamation mark and auto word correction. During the testing, same data was used for both tools and comparison were made with the results. By doing a comparison, the effectiveness of the LexiPro-SM can be identified. Following are the testing on the key features as mentioned above:

4.4.1 Testing on negation

As discussed in Chapter 3, the negation calculation in LexiPro-SM can be divided into three categories which are switch negation, intensifier negation and sentence negation. Below are the corresponding tests made for each negation as follows:

a) Switch negation



LexiPro-SM Result

Prototype: LexiPro Scoring Mechanism

Pos word + Negation: happy

Sentiment score (before grouping): -4

Polarity Group (-5 to +5): -4

Airline service categorization:
Customer service:1 Price:0 Preflight:0 Facility:0 Other:0

Text:
I am not happy with your service

SO-CAL Result

SO-CAL Web -- Results

Your Text:

I am not happy with your service

SO-CAL Score for Your Text:

0

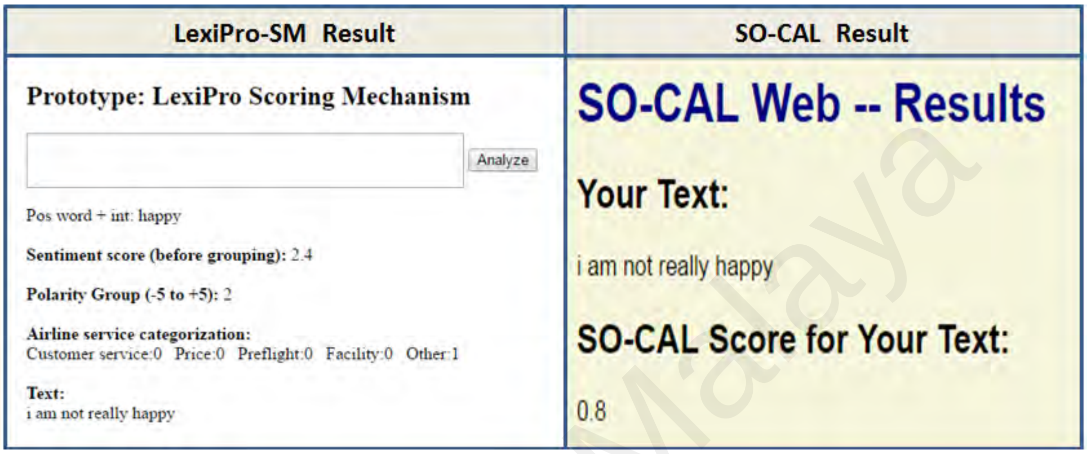
Figure 4.11: Testing results for switch negation

Figure 4.11 shows the result for the negation calculation where LexiPro-SM applied switch negation but in SO-CAL it did not. The score obtained for the sentence “I am not happy with your service” are LexiPro-SM “-4” and SO-CAL “0” respectively. It is shows LexiPro-SM giving a correct sentiment value while SO-CAL did not. In SO-CAL the negation calculation will be calculated by assigning a standard value as “-4” (see Chapter 2). Thus, for some situation it may give an

University of Malaya

inaccurate result as per Figure 4.11 above. However, by applying switch negation method in LexiPro-SM, it improves the sentiment score calculation by detecting a correct sentiment.

b) Intensifier negation



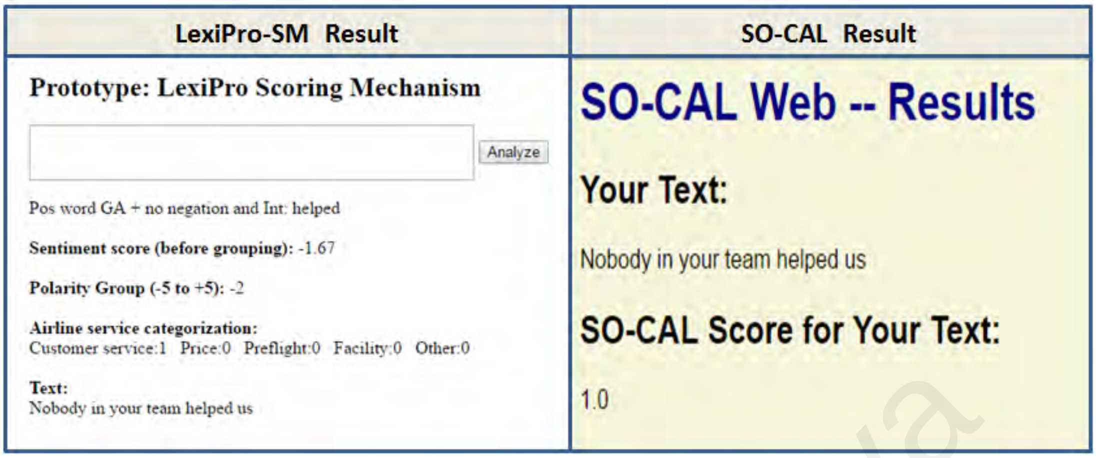
LexiPro-SM Result	SO-CAL Result
Prototype: LexiPro Scoring Mechanism Pos word + int: happy **Sentiment score (before grouping):** 2.4 **Polarity Group (-5 to +5):** 2 **Airline service categorization:** Customer service:0 Price:0 Preflight:0 Facility:0 Other:1 **Text:** i am not really happy	**SO-CAL Web -- Results** **Your Text:** i am not really happy **SO-CAL Score for Your Text:** 0.8

Figure 4.12: Testing results for intensifier negation

Figure 4.12 shows intensifier negation calculation, where the function of intensifier negation reduces the strength of the sentiment but not reverse the sentiment value. The score obtained for the sentence “i am not really happy” are, LexiPro-SM “2.4” and SO-CAL "0.8” respectively. Both results remain in positive sentiment, but SO-CAL gives a lower score closer to the neutral value. LexiPro-SM maintain its results as an average positive value that shows it follows the intensifier negation condition. So overall, LexiPro-SM gives a better score than SO-CAL.

University of Malaya

c) Sentence negation



LexiPro-SM Result	SO-CAL Result
Prototype: LexiPro Scoring Mechanism Analyze Pos word GA + no negation and Int: helped **Sentiment score (before grouping):** -1.67 **Polarity Group (-5 to +5):** -2 **Airline service categorization:** Customer service:1 Price:0 Preflight:0 Facility:0 Other:0 **Text:** Nobody in your team helped us	**SO-CAL Web -- Results** **Your Text:** Nobody in your team helped us **SO-CAL Score for Your Text:** 1.0

Figure 4.13: Testing result for sentence negation

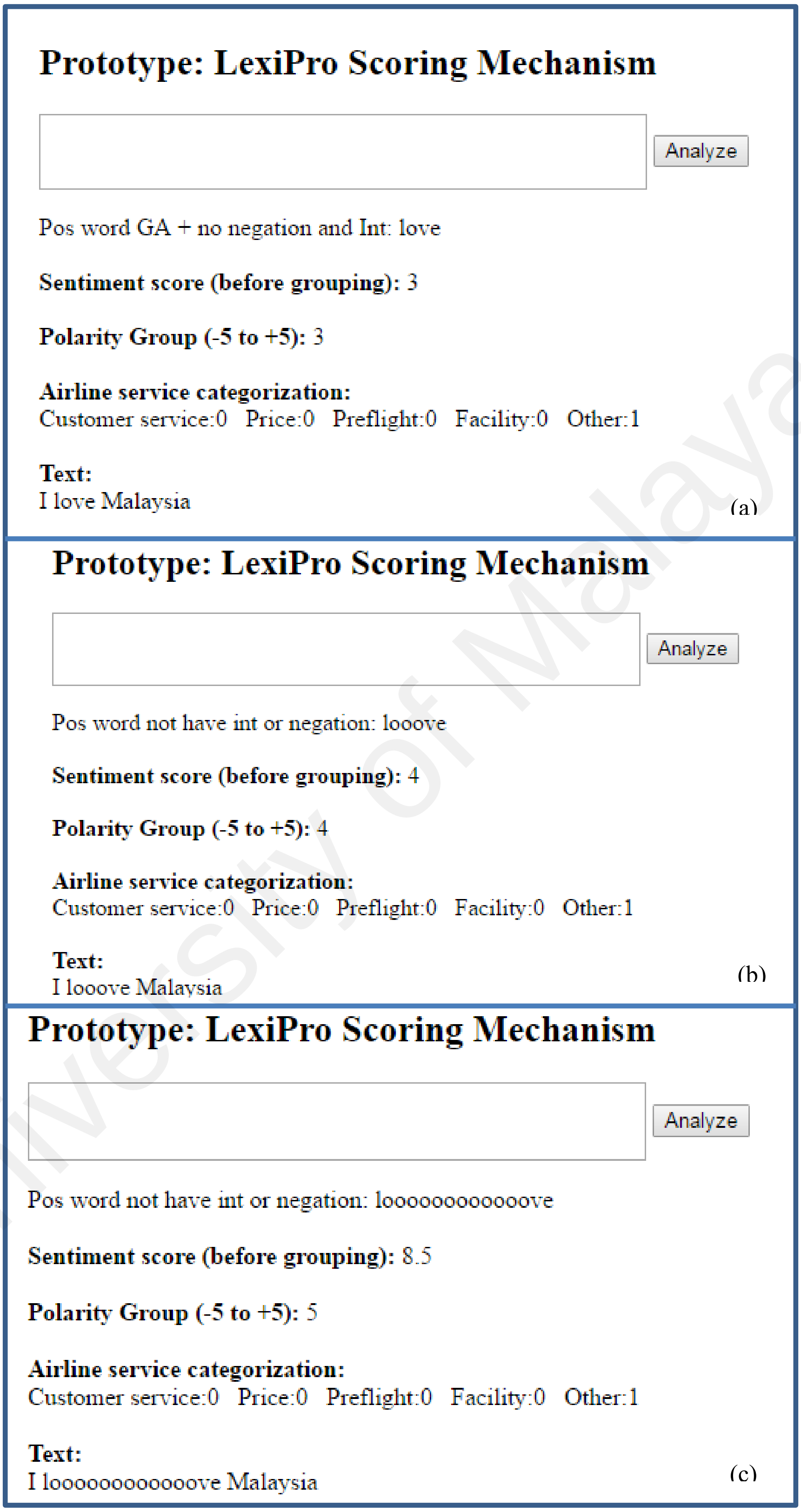
Figure 4.13 shows sentence negation calculation results, where the calculation made for the negation that present next to a non-sentiment word. The scores obtained for the sentence “Nobody in your team helped us” are, for LexiPro-SM “-1.67” and SO-CAL “1.0” respectively. From the results, it shows that LexiPro-SM is able to calculate a correct sentiment score, in this case a negative sentiment while SO-CAL does not.

4.4.2 Testing on repeated letter

This section is the major contribution to this research, where improvements made by identifying lexicon in misspelled words that have repeated letters calculate scores based on the length of the repeated letters.

University of Malaya

University of Malaya



Prototype: LexiPro Scoring Mechanism

Pos word GA + no negation and Int: love

Sentiment score (before grouping): 3

Polarity Group (-5 to +5): 3

Airline service categorization:
Customer service:0 Price:0 Preflight:0 Facility:0 Other:1

Text:
I love Malaysia

Prototype: LexiPro Scoring Mechanism

Pos word not have int or negation: looove

Sentiment score (before grouping): 4

Polarity Group (-5 to +5): 4

Airline service categorization:
Customer service:0 Price:0 Preflight:0 Facility:0 Other:1

Text:
I looove Malaysia

Prototype: LexiPro Scoring Mechanism

Pos word not have int or negation: looooooooooooove

Sentiment score (before grouping): 8.5

Polarity Group (-5 to +5): 5

Airline service categorization:
Customer service:0 Price:0 Preflight:0 Facility:0 Other:1

Text:
I looooooooooooove Malaysia

Figure 4.14: Testing results for repeated letter

Figure 4.14 shows the results of sentiment score calculation for the lexicon word that contain repeated letters (in the form of misspelled word). There are three samples of results, which are 4.14(a) normal sentiment score for lexicon word "*love*" (general sentiment score for love is +3), 4.14(b) and 4.14(c) scores for the repeated letters word which have different lengths of repetition letters. All the sentences used for this calculation has the same meaning as "*I love Malaysia*", however the number of repeated letters "o" in the lexicon word "love" are different in 4.14(b) and 4.14(c), thus giving it different sentiment scores. Overall the tests show that the contribution to this research was achieved and LexiPro-SM is proven to improve the sentiment calculation for repeated letters, while SO-CAL does not. This is because SO-CAL does not have the feature to process misspelled words or the repetition of letters. Table 4.1 shows the testing results of real data that has repeated letters.

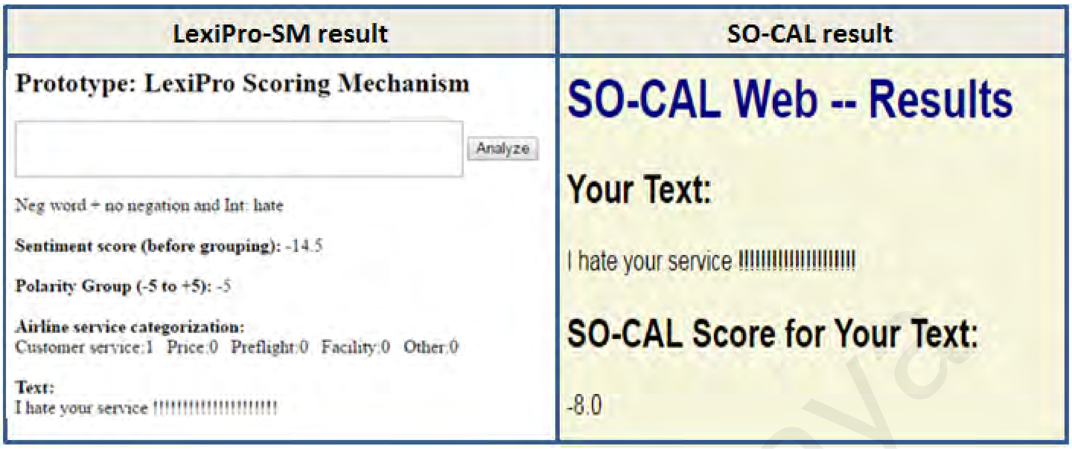
Table 4.1: Results of real data that has repeated letters

PostID	Comment	LexiPro-SM Score	SO-CAL Score
10153831209492387_10153831306292387	Loveee u AA	4	0
10153731962257387_10153732347432387	your website is slowwwwww	-3.5	0
10153847271892387_10153856004057387	Hi yes I saw Super cheapppp	2.95	0
10153746184497387_10153746359392387	Im waitinnnnnggggg	-6.5	0
818701548243040_823267847786410	Speechless woww	2.5	0
831486256964569_835771293202732	Wowwww	3.5	0
838723519574176_838924016220793	I luvv MH	3.5	0
868063803306814_868940789885782	The bestt u all crew	5.5	0

From the results, (Table 4.1) shows that LexiPro-SM is able to produce sentiment scores for repeated letters, whereas SO-CAL is unable to do so. Thus, LexiPro-SM seems to be more effective than SO-CAL in processing misspelled words with repeated letters.

University of Malaya

4.4.3 Testing on exclamation mark

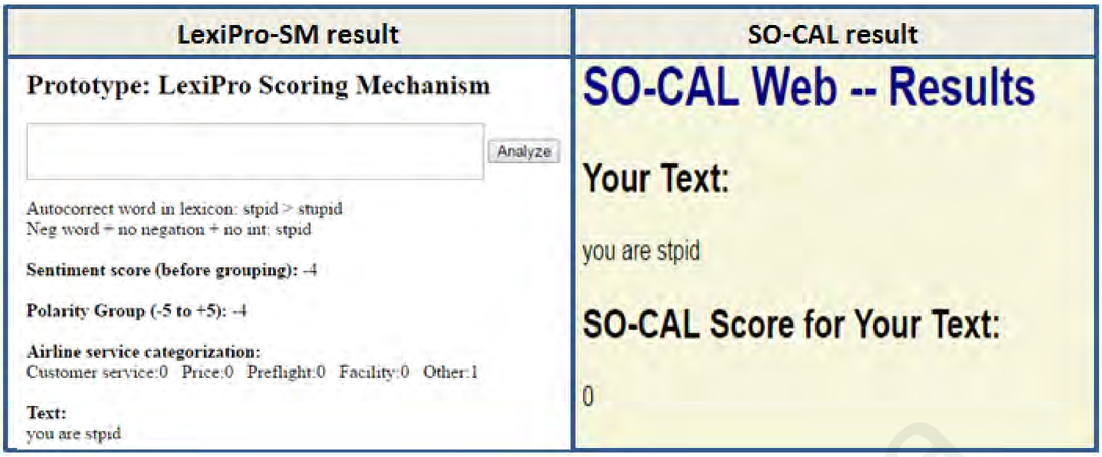


LexiPro-SM result	SO-CAL result
Prototype: LexiPro Scoring Mechanism Analyze Neg word + no negation and Int: hate **Sentiment score (before grouping):** -14.5 **Polarity Group (-5 to +5):** -5 **Airline service categorization:** Customer service:1 Price:0 Preflight:0 Facility:0 Other:0 **Text:** I hate your service !!!!!!!!!!!!!!!!!!!!!	**SO-CAL Web -- Results** **Your Text:** I hate your service !!!!!!!!!!!!!!!!!!!!! **SO-CAL Score for Your Text:** -8.0

Figure 4.15: Testing results for exclamation mark sentiment score calculation

Figure 4.15 shows the testing results of sentiment score calculation for exclamation mark. The scores obtained for the sentence "*I hate your service !!!!!!!!!!!!!!!!!!!!!*" from LexiPro-SM is "-14.5" and SO-CAL is "-8.0". From the results, it shows that LexiPro-SM is able to determine the strength of a sentiment based on the length of repeated exclamation marks, however SO-CAL only calculate scores for the presence of exclamation marks in a sentence (see Chapter 2). Hence, the enhancement made on LexiPro-SM has improved the sentiment score calculation for exclamation marks.

4.4.4 Testing on auto word correction



Prototype: LexiPro Scoring Mechanism

Autocorrect word in lexicon: stpid > stupid
Neg word = no negation + no int: stpid

Sentiment score (before grouping): -4

Polarity Group (-5 to +5): -4

Airline service categorization:
Customer service:0 Price:0 Preflight:0 Facility:0 Other:1

Text:
you are stpid

SO-CAL Web -- Results

Your Text:

you are stpid

SO-CAL Score for Your Text:

0

Figure 4.16: Testing results for auto correction word sentiment score calculation

Figure 4.16 shows the results from analyzing misspelled words with phpspellcheck auto correcter. The score obtained for the sentence "you are stpid" from LexiPro-SM is "-4" and SO-CAL is "0" respectively. From the results, it shows the implementation of phpspecheck auto corrector in LexiPro-SM being able to detect the correct spelling for the word "stpid" as "stupid". However, SO-CAL is not incorporated with a word correction feature. Therefore, the implementation of this feature in LexiPro-SM can produce better results than SO-CAL.

Table 4.2 shows the total processed data, where for Airline A is 4291 and Airline B is 10189. There is huge difference between the processed data count of both airlines. Thus, for further analysis these counts were equalized to 4000 by random data selection from each airline.

Table 4.2: Total processed data count

	Airline A	Airline B
Total processed data	4291	10189

University of Malaya

4.5 Web-based portal for LexiPro-SM analysis report

A web-based portal was created to present the analysis results of LexiPro-SM in a graphical visualization. In this portal, each airline has own page to view the results obtained from the LexiPro-SM analysis. The airline page contains charts that represent the results in three different forms, which are overall result chart, polarity group chart and sub-services chart.

The overall result chart is a pie chart that contains total counts of positive and negative scores (in percentage). This helps the airline management to get an overall idea of customer opinion of the services provided.

The polarity group chart is a bar chart that represents the count (in percentage) of polarity results in eleven categories which are (-5 to +5). The categories indicate the strength of sentiment present in a text or comment, where -5 and -4 indicate as most negative scores which are in red bars. On the other hand, +4 and +5 indicate most positive scores represented in in green bars. Zero polarity is represented by a yellow bar. The purpose of this polarity group is to identify the emotion level of the customer. For example: if the bar chart has high green bars compared to other positive bars, it indicates that the level of customer satisfaction is very high.

The third set of results is a bar chart for sub-services, in which the results are categorized into five groups, which are customer service, price, preflight, facilities and others (Liau & Tan, 2014). The others group is to show the sentiment results that do not fit the four sub-services, for example the phrase "*I am happy*", this indicates a positive sentiment, however it is not related to any particular sub-services. Thus, it will be categorized as "other". The results for each group are presented in dual bars where red color indicates negative scores and green color indicates positives scores. The count of negative and positive scores was presented as a percentage.

University of Malaya

The customer service group is related to staff performance, such as the call center, online supports (email, e-form) etc. The price group is for comments related to payments, charges, fares, promotion, ticketing etc. The pre-flight group is related to the services that are provided before boarding such as the check-in process, luggage, boarding pass, flight selection etc., whereas the facilities group is for the facilities that are provided in and out of the flight, such as airport environment, food and beverage served in flight, special class service in flight, medical supports etc. The sub-service chart will be useful for the airline management to monitor the services provided for the customer, where management can narrow down their attention to a particular service if the chart shows a high percentage of negative scores (i.e. indicates a poor performance). Furthermore, these results may help the airline management to determine their marketing strategy without performing any manual activity such as questionnaires and interviews which could involve high expenses and consume long hours.

In addition, the airline page was also built with a feature which allows the airline management to view the top five comments with the most negative and most positive scores (Figure 4.20). The purpose of this function is to show the five real comments that achieved the highest scores. This would be helpful for the management to understand the emotion of the customer by reading their comments and would be helpful to maintain the reputation of the organization by solving problems immediately (for negative comments) or encouraging customers with additional services such as free tickets and promotions (for positive comments).

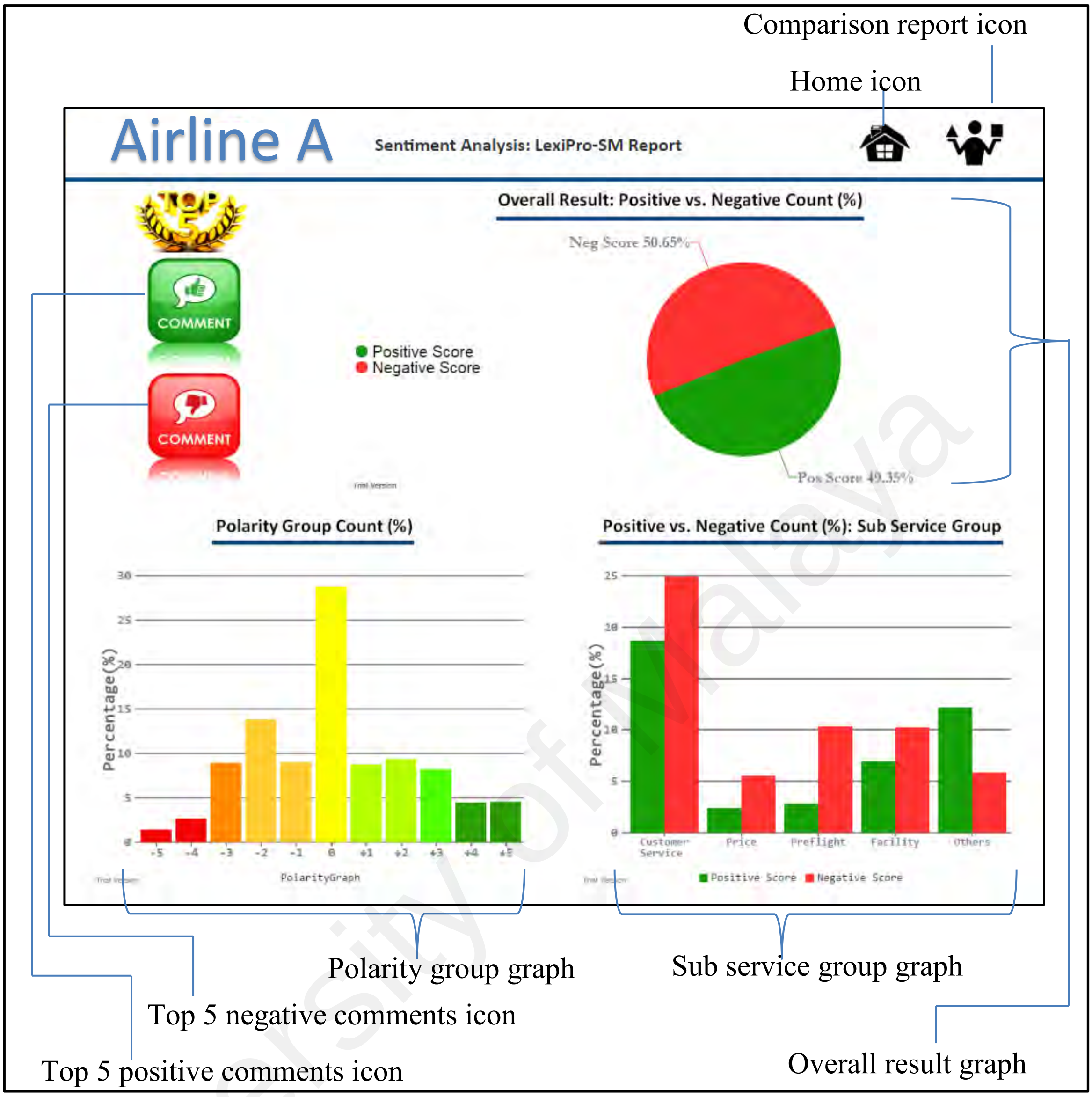
Furthermore, the airline page was linked to another page which is called a comparison report page. This page shows the comparison made between the services of the two airlines which are Airline A and Airline B. Then, comparison comments were produced for each part of the analysis which were overall service result, polarity group result and sub-services result.

University of Malaya

This comparison report would be helpful for the airline management to make comparisons with the competitor airlines and improve the services provided in order to achieve more support from customers.

Following are the screen captures of the web-based portal that presents the LexiPro-SM results. This web-portal has a homepage that linked with the Airline A analysis report page (Figure 4.17), Airline B report page (Figure 4.18) and Comparison report page (Figure 4.19).

University of Malaya



Comparison report icon

Home icon

Polarity group graph

Sub service group graph

Top 5 negative comments icon

Overall result graph

Top 5 positive comments icon

Figure 4.17: Airline A analysis report page

University of Malaya

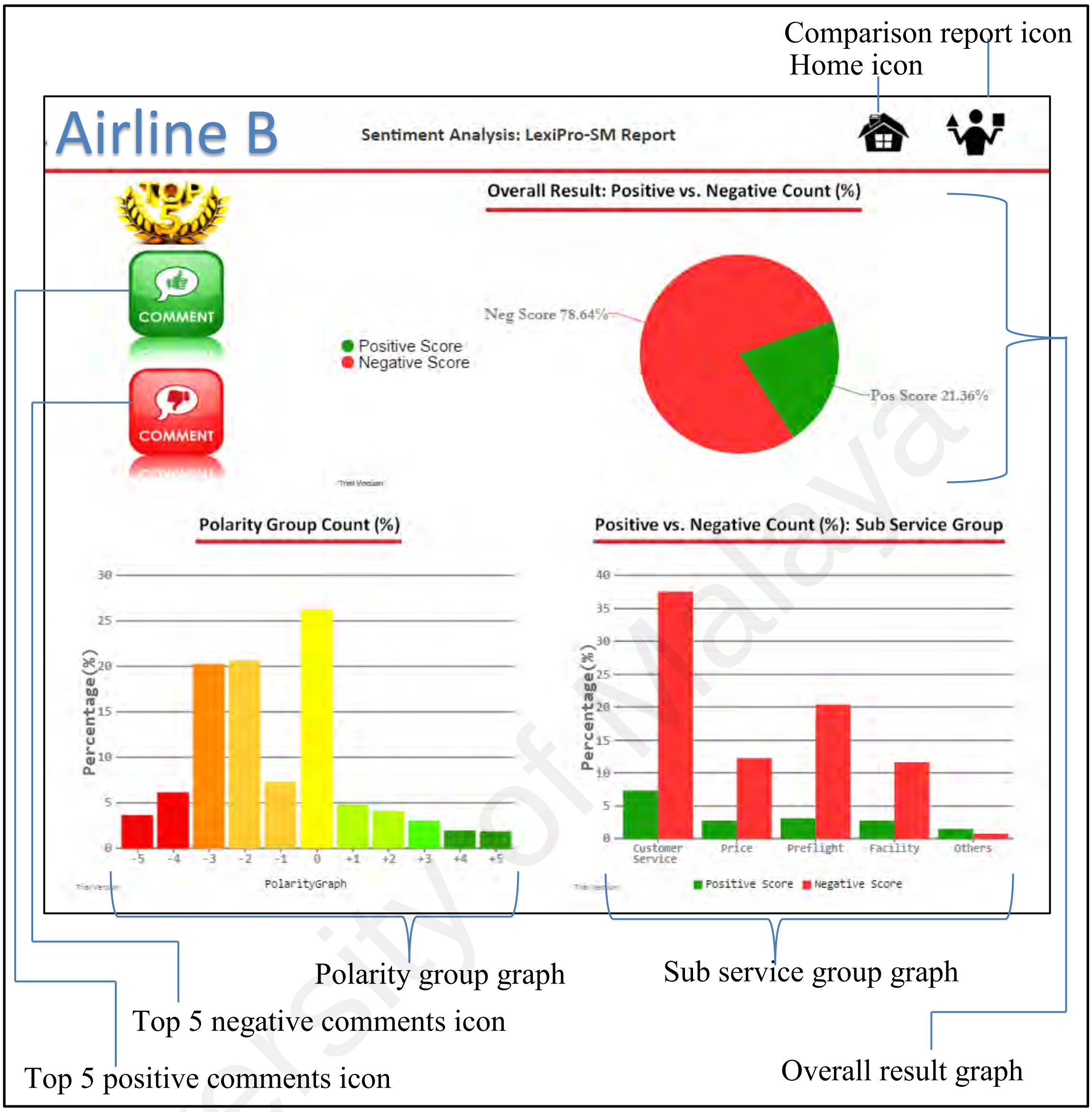
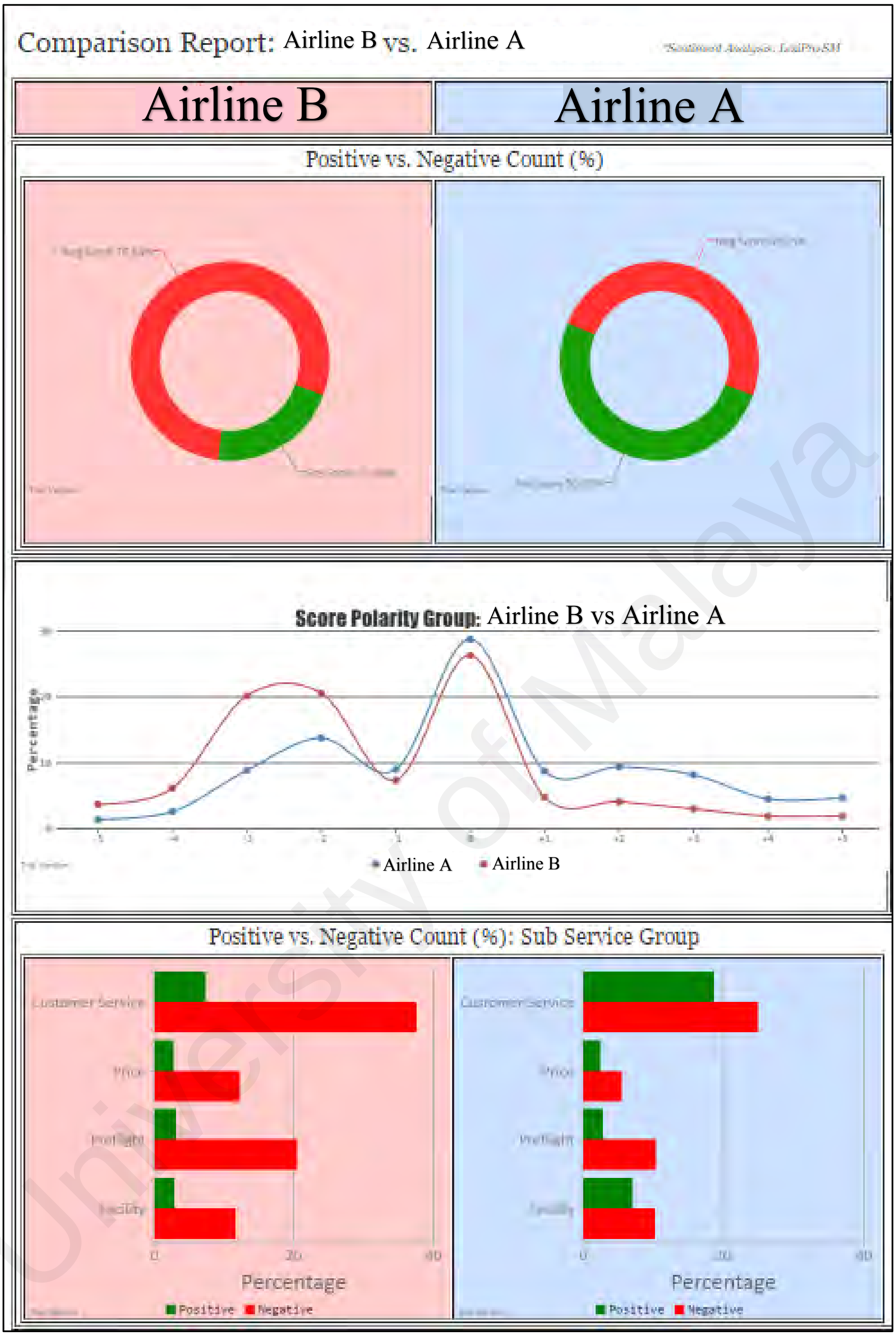


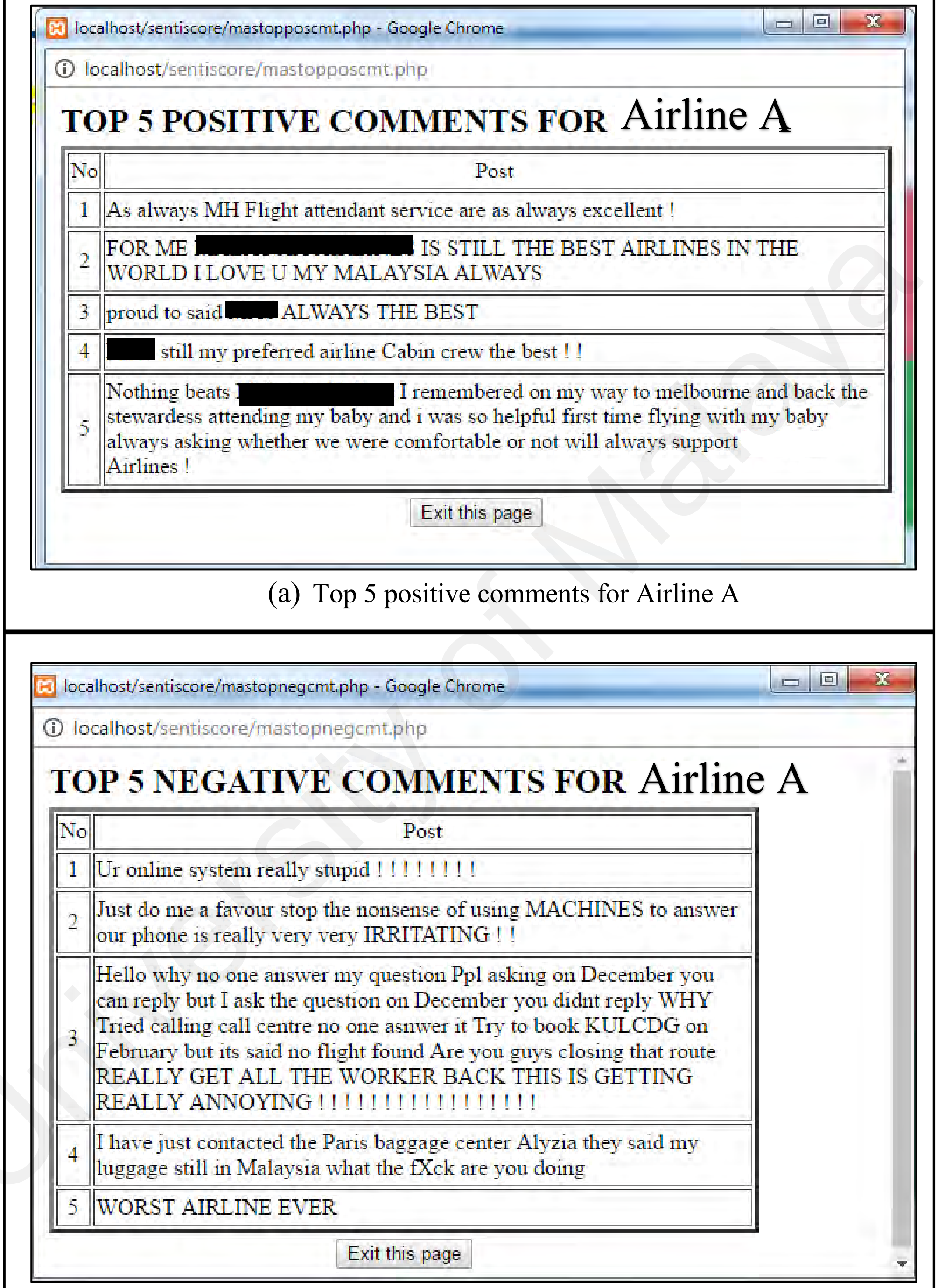
Figure 4.18: Airline B analysis report page

University of Malaya



University of Malaya

Figure 4.19: Comparison report page


localhost/sentiscore/mastopposcmt.php - Google Chrome
localhost/sentiscore/mastopposcmt.php

TOP 5 POSITIVE COMMENTS FOR Airline A

No	Post
1	As always MH Flight attendant service are as always excellent !
2	FOR ME [redacted] IS STILL THE BEST AIRLINES IN THE WORLD I LOVE U MY MALAYSIA ALWAYS
3	proud to said [redacted] ALWAYS THE BEST
4	[redacted] still my preferred airline Cabin crew the best ! !
5	Nothing beats [redacted] I remembered on my way to melbourne and back the stewardess attending my baby and i was so helpful first time flying with my baby always asking whether we were comfortable or not will always support Airlines !

(a) Top 5 positive comments for Airline A

localhost/sentiscore/mastopnegcmt.php - Google Chrome
localhost/sentiscore/mastopnegcmt.php

TOP 5 NEGATIVE COMMENTS FOR Airline A

No	Post
1	Ur online system really stupid ! ! ! ! ! ! ! !
2	Just do me a favour stop the nonsense of using MACHINES to answer our phone is really very very IRRITATING ! !
3	Hello why no one answer my question Ppl asking on December you can reply but I ask the question on December you didnt reply WHY Tried calling call centre no one asnwer it Try to book KULCDG on February but its said no flight found Are you guys closing that route REALLY GET ALL THE WORKER BACK THIS IS GETTING REALLY ANNOYING !
4	I have just contacted the Paris baggage center Alyzia they said my luggage still in Malaysia what the fXck are you doing
5	WORST AIRLINE EVER

(b) Top 5 negative comments for Airline A

Figure 4.20: Example of top five comments

University of Malaya

4.6 Summary

In summary, this chapter presented three types of system design describing the database relationship with an entity, data analyst activities in LexiPRO-SM sentiment analysis, and the data flow between processes and database. The three system designs are Use case diagram, Entity relation diagram (ERD) and Data flow diagram. Detailed explanation given on system implementation that involves data collection, data cleaning and data analysis processes.

In the implementation of data collection, an explanation given on the AP references that was used to access Facebook data, show the type of information that will be extracted from Facebook pages and stored in databases. The extracted data will be stored in a raw database where the tables are named “rawdataA” for Airline A and “rawdataB” for Airline B.

In the implementation of data cleaning, an explanation was given on how scripting work with the functions to remove all noises in raw data and showed examples of data cleaning with data stored in a clean database. The clean data are stored in clean database where the tables named “cleandataA” for Airline A and “cleandataB” for Airline B respectively. Following that, the importance of LexiPro-SM prototype and its features was discussed. Furthermore, a complete testing was performed between LexiPro-SM and SO-CAL on the improved features such has negation, repeated letters, exclamation marks and auto correction word. Finally, the functions of web-based portal for LexiPro-SM analysis were discussed.

University of Malaya

CHAPTER 5: RESULTS AND DISCUSSION

5.1 Introduction

Chapter 5 discusses a case study from the airline industry in Malaysia and the evaluation results of the proposed enhanced scoring mechanism, LexiPro-SM. A comparison is made between LexiPro-SM and SO-CAL to identify the effectiveness of LexiPro-SM.

5.2 LexiPro-SM analysis on Malaysian airline industry

Based on the sentiment analysis results, a case study was conducted between the two airlines: Airline A and Airline B. This case study identified customers' opinions (from each airline) and performed a comparison study to determine the airline that provide better service for passengers and travelers. Also, a web-based portal was created to present the analysis results in a graphical form as illustrated in chapter 4 (the physical design of the portal). The following sections present LexiPro-SM analysis results from the two airlines which involve overall service result, polarity group result, and sub-services result.

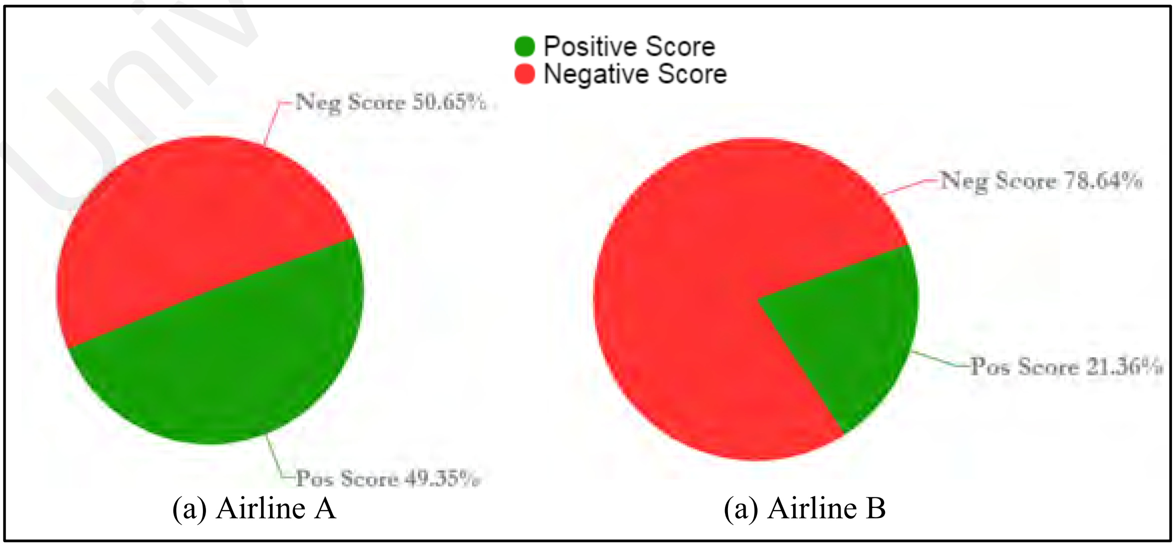
The overall service represents the negative and positive counts that belong to each airline. Polarity group is the categorization of the sentiment scores into eleven categories within the range of -5 to +5, in which -5 (red in chart) represents most negative, +5 (green in chart) represents most positive, and 0 represents neutral (yellow in chart). The counts of the data for each group are presented in a percentage form representing the polarity results. Lastly, the airline service is divided into four categories namely customer service, price, pre-flight and facility. Data that do not match any of the four categories are categorized as "others".

University of Malaya

5.2.1 Overall service: positive vs. negative count (%)

Figure 5.1 illustrates the overall results based on the positive and the negative score counts (in percentage). Figure 5.1(a) shows the results of Airline A, where the score for positive is 49.35% and negative is 50.65%. Figure 5.1(b) shows the results of Airline B, where the positive score is 21.36% and the negative score is 78.64%. As for Airline A, the result shows that the negative score is slightly higher than the positive score in which the difference between the two scores is only 1.3%. Besides, the results of Airline B also show that the negative score is higher than the positive score, but the two scores are substantially different as much as 57.28%. This simply means that Airline B has a higher percentage of negative scores compared to positive scores. By comparing these two charts, results show that Airline B has achieved higher negative score than Airline A, which in turn indicate that many customers were unhappy with the service provided by Airline B. Although Airline A has high negative scores, the positive scores of Airline A are relatively higher than the positive scores of Airline B, indicating that Airline A provides better service than Airline B. The comparison is continued in the following section (with polarity results) to identify the percentage value obtained by both airlines in the highest polarity categories (-4,-5, +4 and +5).

University of Malaya



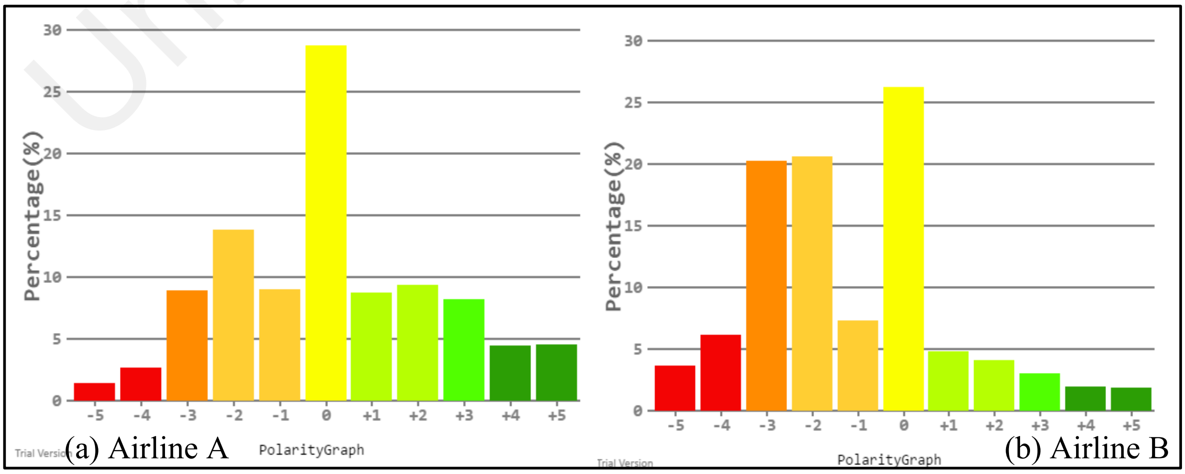
(a) Airline A (a) Airline B

Figure 5.1: Overall service result

5.2.2 Polarity group result

Figure 5.2 illustrates the polarity results from the two bar charts whereby 5.2(a) represents Airline A and 5.2(b) represents Airline B. Both charts have highest percentage of neutral (yellow) score; that means, out of 4000 data sets from each airline, more than 25% data did not contribute to the sentiment result. Besides, from the Airline A results, the highest negative bar is -2 (13.83%), while Airline B is also -2 (20.6%). Airline A's highest positive bar is +2 (9.35%), while Airline B's highest positive bar is +1 (4.08%). These highest polarity results show that most comments from both airlines belong to the moderate sentiment range (-2,+1,+2) indicating that customer satisfaction for both airlines is in the satisfactory level. However, when comparing the percentage of the most positive (+4 and +5) bars between the two airlines, Airline A has higher green bars (9.06%) than Airline B (3.82%), which indicate that customer satisfaction is greater in Airline A compared to Airline B. Similarly, when comparing the most negative (-4 and +5) bars between the two airlines, it shows that Airline B has higher red bars (8.8%) than Airline A (4.08%), which indicate that the number of people who are very unsatisfied with the service is higher in Airline B than Airline A. Overall, the representation of negative bar charts is higher in Airline B and lower in Airline A. This trend concludes that Airline A provides better service than Airline B.

University of Malaya



(a) Airline A

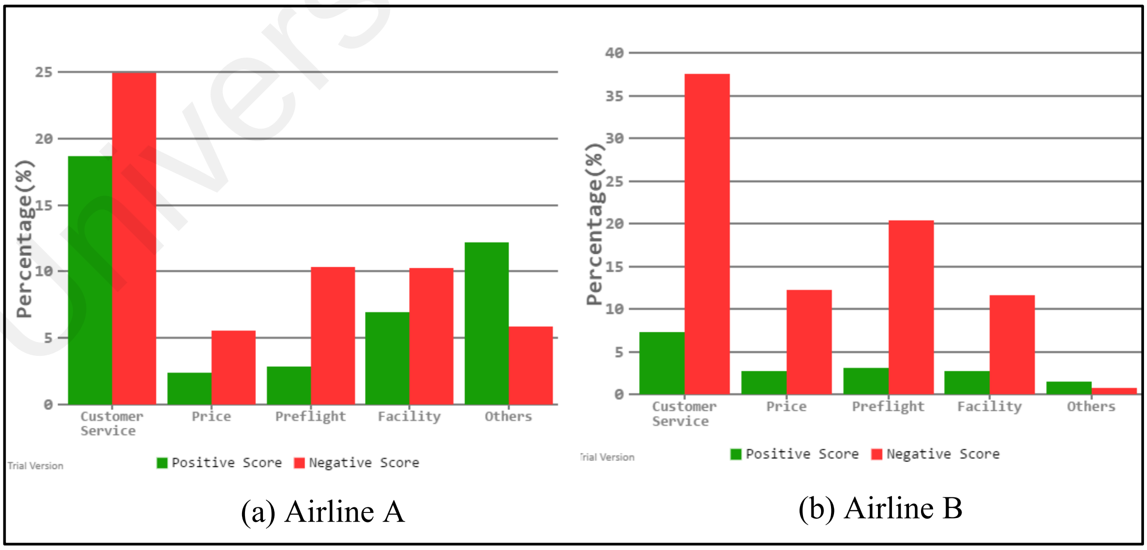
(b) Airline B

Figure 5.2: Polarity results

5.2.3 Airline sub-services score

Figure 5.3 illustrates the scores in graphical forms as shown in the following bar charts. Figure 5.3(a) depicts the scores for Airline A and 5.3(b) for Airline B. These bar charts have five categories namely customer service, price, preflight, facility, and others. The green bars represent positive scores whereas red bars represent negative scores. The function of this bar chart is to show customers' opinions towards the five categories mentioned. By so doing, the score results help airline management to identify the rating score for sub-services and plan for pre-emptive measures to make improvements. From the results, every sub-service group has higher negative scores than positive scores except for the "others" group which is treated as irrelevant information for this sub-service analysis. Therefore, in this case, all categories require improvement. However, the most significant group can be identified by calculating the difference between negative and positive sentiment scores in order to identify the group with the highest negative opinions.

University of Malaya



(a) Airline A

(b) Airline B

Figure 5.3: Sub-services scores for Airline A and Airline B

Table 5.1 shows the difference between negative and positive scores for both airline categories. The highest scores represent highest negative value. As for Airline A, pre-flight has the highest negative difference, whereas for Airline B, the highest negative difference is present in its customer service.

Table 5.1: Difference between negative and positive scores

Negative score	**Customer Service (%)**	**Price (%)**	**Pre-Flight (%)**	**Facility (%)**
Airline A	6.29	3.17	7.53	3.26
Airline B	30.21	9.51	17.3	8.87

University of Malaya

5.3 Results based on evaluation metrics

As discussed in Chapter 3, LexiPro-SM was evaluated using evaluation metrics such as accuracy, recall, precision, and F1-score measure. This evaluation measure was performed using 300 data sets that were randomly selected from the total processed data of both airlines (Airline A and Airline B) and classified by human experts. According to Narr et al. (2012), human annotation for a large data would be costing high. Besides, trained data that based on human expert classification (hand-labelled) should consists only a small amount of data (Narr et al.,2012). Thus, 300 trained data have been finalized by considering the cost and time constraints involved in this research. The human expert classification was performed by three linguistic experts. Their main task was to classify each comment as positive, negative or neutral. With the help of three human experts, the possibility of a tie in identifying sentiment of each post was eliminated as the post was identified as positive, negative or neutral based on majority (Muhammad et al., 2015). Thus, the finalized human expert results were used to perform evaluation measure calculation. SO-CAL was also evaluated with these metrics for the purpose of making a comparison between LexiPro-SM and SO-CAL.

Table 5.2 shows the finalized results of the human expert classification depicting the count of positive, negative, and neutral out of the 300 data sets (chapter 3).

Table 5.2: Human expert classification results

Human expert classification	Positive	Negative	Neutral	Total
	91	106	103	300

5.3.1 Accuracy result

Figure 5.4 shows the accuracy results comparing LexiPro-SM and SO-CAL as illustrated in graphical form. Based on the observation of the accuracy results, it shows that LexiPro-SM has achieved a high accuracy of 90.7% whereas SO-CAL has only achieved moderate accuracy of 58.33%. This indicates that LexiPro-SM achieved a higher accuracy compared to SO-CAL.

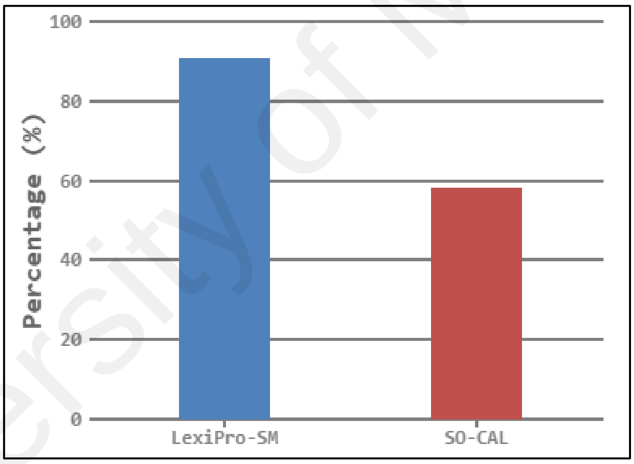


Figure 5.4: Accuracy results

5.3.2 Performance results

Table 5.3 shows the performance results of LexiPro-SM comparing SO-CAL that included the results of recall, precision, and F1 score. For LexiPro-SM, there is not much difference between recall and precision results. However, SO-CAL has a significant difference between recall and precision achieved, especially for positive and negative sentiments where positive recall is 60.44% while the positive precision is 49.6%. Meanwhile, the negative recall is 59.43% and the negative precision is 75%.

University of Malaya

These differences lead to calculating F1score for all the results in order to identify a harmonic average between recall and precision. So, the comparison was made by looking into the F1 score achieved by both tools.

Table 5.3: Performance results

Tool	**LexiPro-SM**			**SO-CAL**		
Metrics	**Positive (%)**	**Negative (%)**	**Neutral (%)**	**Positive (%)**	**Negative (%)**	**Neutral (%)**
Recall	98.9	85.9	88.4	60.44	59.43	55.34
Precision	90	91	91	49.6	75	54.3
F1 Score	94.24	88.4	89.68	54.5	66.31	54.82

For LexiPro-SM, the best F1 score was achieved in positive sentiments (94.24%) followed by neutral sentiment (89.68%) and finally negative sentiment (88.4%). It shows that LexiPro-SM is more sensitive in positive sentiment detection than other sentiments, however the scores achieved by these three sentiments still belong to a high accuracy level (>80%), so the difference among the sentiments do not give much effect.

Meanwhile, for SO-CAL, the F1 score for negative sentiment (66.31%) is higher than positive (54.5%) and neutral (54.82%). This indicates that SO-CAL is more sensitive in negative sentiment detection than other sentiments. However, SO-CAL performance results only achieved a moderate level of accuracy which shows that the sentiment detection of SO-CAL is not as effective as LexiPro-SM.

Based on the overall results of the performance, it shows that LexiPro-SM achieved a higher accuracy in all sentiments (positive, negative and neutral) than SO-CAL. Therefore, it can be concluded that the sentiment detection in LexiPro-SM is more effective than SO-CAL.

University of Malaya

All the differences (i.e. recall, precision and F1 Score) were found to be significant between LexiPro-SM and SO-CAL. To be precise, the difference for recall was noted at $p < 0.001$ ($t = 2.785$), precision at $p < 0.001$ ($t = 1.76$) and F-score at $p = 0.026$ ($t = 1.17$) for positive sentiments. For negative sentiments, the difference for recall was noted at $p < 0.001$ ($t = 1.531$), precision at $p = 0.033$ ($t = 0.889$) and F-score at $p = 0.018$ ($t = 1.07$).

5.4 Discussion

The two research questions which were addressed in this research are:

RQ1: How to enhance the scoring mechanism?

RQ2: How to access the effectiveness of the enhanced scoring mechanism?

Research question one was answered in Chapter 3 and 4 which explained the development process of LexiPro-SM with the incorporation of features to enhance the scoring mechanism. These features are negation, repeated letter, exclamation mark, and auto word corrector. Negation was improved by using three types of negations such as switch negation, intensifier negation, and sentence negation. The variation in negation helps to detect a correct sentiment and produce a better result for negation text analysis. Repeated letter is the main contribution for this research, in which the improvement was made by calculating the sentiment scores based on the length of the repeated letter word instead of assigning a general score. Similarly, the calculation for exclamation mark has improved by assigning the score based on the number of exclamation marks repeated instead of assigning a general score. Then, the final improvement was made by incorporating auto word corrector which can help to detect sentiment from the misspelled word by replacing it with a correct word.

University of Malaya

On the other hand, the answer for research question two was provided from the evaluation measured and comparison made with the existing system. From the observation of performance results, it shows that LexiPro-SM achieved a greater percentage than SO-CAL. This is because, LexiPro-SM was not only developed to calculate semantic orientation in a sentence, but also the scores calculation in LexiPro-SM considered the text style used to express opinions in both airline pages. For example, from the observation of the data collected, many data contained sentence negations and many sentences were expressed with the repetition of an exclamation mark. Furthermore, the addition of acronym dictionary, repeated letter feature, and auto word correction feature has led LexiPro-SM to achieve a better accuracy.

However, the main reason for SO-CAL achieving a moderate level of accuracy is that the inability to process misspelled words (such as wrong spelling/typo, acronym, and repeated letter). In this research data, much text appeared in an informal form which is a challenge for SO-CAL in identifying the sentiment. For example, the presence of acronym in text such as “gd for good” or “hpy for happy” was not defined in SO-CAL dictionary (Taboada et al., 2011), so the acronym text was not detected for the sentiment calculation.

SO-CAL also produced erroneous results in negation detection in which the data that belonged to negative sentiment was given a score of zero (neutral) due to the SO-CAL calculation method that involved general value for negation as “-4”. Moreover, although the detection of negation in sentence was incorporated in SO-CAL (Taboada et al., 2011), it did not work effectively when processing the sentence negation text in this research data.

Therefore, from the comparison made between LexiPro-SM and SO-CAL, the incorporation features such as processing misspelled, enhanced negation technique, and auto word correction on LexiPro-SM has improved the accuracy in the sentiment

University of Malaya

detection. In fact, these accuracy results have proven that LexiPro-SM is more effective than SO-CAL in processing social media data (Taboada et al., 2011).

5.5 Summary

In summary, a case study was conducted by comparing the two airlines in Malaysia (Airline A and Airline B) where the comparison involved overall result, polarity result, and sub-service result. In general, it shows that Airline A has more positive scores than Airline B which can be concluded that MAS provides a better service to customers than AirAsia. However, both services have highest negative scores in every result that require attention by the airline management.

Furthermore, evaluation measures were performed on LexiPro-SM and SO-CAL's analysis results with the reference to human expert results. The performance results showed that LexiPro-SM has a high accuracy for all the sentiments, however SO-CAL only have moderate accuracy. Therefore, it indicates that the enhancement of scoring mechanism (LexiPro-SM) has improved the sentiment detection accuracy and produced more effective results than SO-CAL. The main reason identified for the poor performance of SO-CAL was that, its inability to process misspelled word and producing some fault results in the negation calculation.

University of Malaya

CHAPTER 6: CONCLUSION, LIMITATION AND FUTURE WORK

6.1 Research overview

Sentiment analysis is an activity to analyze opinions towards an entity (Liu, 2012). This analysis can be divided into three levels which are document level, sentence level and entity/aspect level (Pang et al., 2002; Liu, 2012; Hu & Liu, 2004). Sentiment analysis techniques can be divided into two categories; the machine learning technique and the lexicon-based technique (Muhammad et al., 2015). The machine learning technique uses machine language to analyze data. This technique can be further divided into two methods which are supervised learning and unsupervised learning (Shelke et al., 2016;Vohra & Teraiya, 2013;Soni and Patel, 2014). The lexicon-based technique uses a collection of sentiment words and calculates the scores based on semantic orientation of words or phrases (Turney, 2002; Liu, 2016; Serrano-Guerrero et al., 2015). This technique can be divided into two methods which are dictionary based and corpus based (Serrano-Guerrero et al., 2015).

The scope of this research lies within the lexicon-based sentiment analysis domain. Thus, literature studies were conducted on the research that was closely related to the lexicon-based approach and identified the main limitations from each study such as not processing misspelled words, limitations in the length of text and lack of sentiment features incorporated such as lexical valence. Through the literature review process, the problem statement emerged as the sentiment detection in repeated letters was not fully accurate because a general score was assigned for the detection of repeated letter in a sentence and the strength of the sentiment was not defined based on the length of the repetition of letters. According to Brody and Diakopoulos (2011), the length of repeated letters can increase the strength of a sentiment.

University of Malaya

Based on the problem statement, two research objectives were identified which are:

Objective 1: To enhance the scoring mechanism for text-based sentiment analysis.

Objective 2: To evaluate the effectiveness of the enhanced scoring mechanism.

The project scope of this research focuses on applying the lexicon-based sentiment analysis to the airline industry in Malaysia which involved two well established airlines: Airline A and Airline B, where the data were obtained from airline's official Facebook page. The research contribution for this research is mainly about improving the non-lexicon modifier which is repeated letters or characters. This will be discussed in the following section.

Literature studies were performed by identifying problems faced in natural language processing (NLP), discussed the approaches in lexicon creation (manual and automated) and the necessity of lexicon modifiers in sentiment analysis. The literature studies continued with a review of existing lexicon-based sentiment analysis systems such as Semantic Orientation Calculator (SO-CAL), SentiStrength, Linguistic Inquiry and Word Count (LIWC), SentiHealth-Cancer (SHC-pt) and SmartSA. From the literature review, it was found that all the existing studies have similar limitations on the implementation of the non-lexical modifier which can pose a threat when defining strength of sentiment. Thus the focus of this research is to improve the non-lexical modifiers (especially repetition of letters).

The methodology of this research has discussed the planning and steps involved in building the proposed scoring mechanism LexiPro-SM. The system methodology deployed in this research is Rapid Application Development (RAD). The programing language used to develop LexiPro-SM is PHP and used phyMyadmin database for data storage. The most important processes involved in methodology are data collection, data

University of Malaya

cleaning and data analysis. Data collection was performed by collecting airline data (MAS and AirAsia) from their Facebook pages with the aid of Graph API (PHP SDKs). Data cleaning involved removing all the noise that was not relevant for this research where the clean data contained only alphabets and exclamation marks. Data analysis involved the process to determine the sentiment score and categorized the score based on the polarity range. In addition, explanations given for the features were incorporated in LexiPro-SM such as dictionaries, negation, intensifier, exclamation mark, capitalization, repeated letter, auto corrected words and airline service grouping. Then evaluation planning was discussed, using evaluation metrics to determine the effectiveness of LexiPro-SM.

Furthermore, a complete testing was performed between LexiPro-SM and SO-CAL of the improved features such has negation, repeated letter, exclamation mark, and auto corrected words. This improvement has increased the sensitivity of sentiment detection. Thus, the first objective of this research has been achieved.

Evaluation measures (accuracy, recall, precession and F1-score) were performed on LexiPro-SM and SO-CAL's analysis results with the reference of human expert results. The results obtained from the evaluation measure were used to compare LexiPro-SM and SO-CAL to identify the effectiveness of LexiPro-SM. LexiPro-SM achieved a higher accuracy (90.7%) than SO-CAL (58.33%). Furthermore, LexiPro-SM also achieved a high F1-score for all the sentiments which are 94.24% (positive), 88.4% (negative), and 89.68% (neutral). However, the F1-score achieved for all the sentiments in SO-CAL were lower than LexiPro-SM, where the scores are 54.5% (positive), 66.31% (negative) and 54.82% (neutral). Therefore in this research the enhancement of the scoring mechanism (LexiPro-SM) improved the accuracy of sentiment detection and produced more effective results than SO-CAL. Thus, the second objective of this research has been achieved.

University of Malaya

A web-based portal was developed for the airline management to visualize the results of LexiPro-SM sentiment analysis in graphical structure. Each airline has its own page which contains three types of charts which are overall service, polarity group and sub-services. The overall service chart shows the total count of positive and negative sentiments. The polarity group chart shows the count of sentiments present for each group (-5 to +5). These two charts may indicate the performance of an airline. In the sub-services chart, the results were divided into five categories which are customer service, price, pre-flight, facilities and others. This can help the management to identify a service that gives a poor performance and take a specific action to improve that particular service. In addition, having sub-services results may help airline management to reduce the time and cost of doing manual data collection such as questionnaires, interviews etc. Furthermore, each airline portal has added a feature to view the top five comments that belong to negative and positives sentiments. This will help airline management to view the real comments of customers and identify the most important issues that need to be solved immediately. Furthermore, a comparison page was linked to the airline page, where the management of each airline can compare their results with their competitor. This may help to identify the weaknesses of their organization that can bring further improvement to the service provided.

Then, a case study was conducted of two airlines in Malaysia (Airline A and Airline B) by comparing their results for overall service, polarity group and sub-services. In overall service, the positive score of Airline A is 49.35% whereas for Airline B is 21.36%. In polarity group, the most positive (+4 and +5) polarity percentage for Airline A is 9.06% and Airline B is 3.82%. Then, the most negative (-4 and +5) polarity percentage for Airline A is 4.08% and Airline B is 8.8%. These results show that Airline A has more positive scores than Airline B which indicate Airline A is providing better service to customers than Airline B. Having these results means airline management can monitor their performance and improve the service provided to achieve customer

University of Malaya

satisfaction. For example, Airline A has a higher positive score than Airline B but the overall service results of Airline A are at an average level where the positive score is 49.35% and negative is 50.65%. This indicates that Airline A management should improve the service in order to get better reviews which in turn can retain the reputation of the organization.

Moreover, the sentiment analysis results were divided into five sub-services. At both airlines the service that has the highest positive and negative score is customer service. Additionally, the negative scores for all the services of both airlines are higher than positive scores (except the "other services" group which is irrelevant). Thus, the difference between negative score and positive score was calculated for each service to determine a service that has most negative score. For Airline A the service with the highest negative score is pre-flight facilities (7.53%) and in Airline B it is customer service (30.21%). These results indicate that Airline A provided poor service for pre-flight services such as ticketing, baggage, boarding etc. Then, Airline B provided poor service on customer service such as communication, staff issues etc. So, both airlines' management can give priority to the sub-service that has the highest negative scores. Overall the web-based portal could help airline management to determine the marketing strategy of their organization.

6.2 Research contribution

The main contribution of this research was the improvement of one of the non-lexical modifiers which is repetition of letters, where a lexicon will be identified in a typo word that has repetition of letters and gives scores based on the length (number) of repeated letters contained in the typo word. In the data analysis phase, two methods were implemented to analyze the repeated letter which were looping and stemming method. The looping method is a method to analyze the repetition of letters that belong to one type of character, whereas, the stemming method analyzes repetition of letters

University of Malaya

that belong to two or more types of characters. A standard score for repetition of letters was determined as 0.5 (each repeated letter) which was finalized during internal testing. Overall the testing conducted shows that the contribution this research has achieved is the enablement for LexiPro-SM to improve the sentiment calculation for the repeated letter.

6.3 Limitation and future work

There were two important limitations found in this research. First, the division of sentiment in text according to its entity may improve the accuracy of sentiment detection. This is because, the scoring mechanism producing lowest or neutral score for the text that contained both negative and positive sentiments. For example, the sentence “I love the food served on board but hate the seat which is small in size” produced a neutral score in LexiPro-SM. However, this sentence has two different sentiments that described two different entities, which means in a real situation, both sentences are contributing sentiments but for different entities.

Second, the improvement on auto word correction algorithm may increase the accuracy of sentiment detection. This is because misspelled words which do not contribute to sentiment caused wrong detection in auto word corrector. For example, “lyak” (non-sentiment misspelled word) can be detected as “leak”. This wrong word detection can cause a wrong sentiment score calculation and it might affect the strength of the sentiment.

POS tag and sarcasm also could improve the accuracy of sentiment detection. POS tag is the collection of sentence that has a combination of adjectives and adverbs, where most words have more than one POS tag. This feature will be useful to identify sentiment for the word that can contribute two different meanings (Das and Balabantaray, 2014). Sarcasm is a sentiment expressed in a text by using positive or intensified positive words to express negative sentiment (Bharti et al, 2016). For

University of Malaya

example the sentence “I love being ignored”, is contributing negative sentiment that belongs to disappointment but it was expressed with positive sentiment. Several studies were conducted on sarcasm such as Das and Balabantaray (2014) who detected sarcasm in real time tweets, Riloff et al. (2013) identified sarcasm as contrast between positive and negative sentiment and Bharti et al., (2015) recognized parsing-based sentiment in twitter data. Therefore, future studies could look into the possibility of including features such as POS tag, sarcasm etc. to improve sentiment analysis.

6.4 Conclusion

In conclusion, the objectives of this research (as discussed above) have been achieved, where the scoring mechanism was enhanced with the incorporation of improved features such as repetition of letters, negation, exclamation marks and auto word corrector. Then the effectiveness of LexiPro-SM was determined by comparing LexiPro-SM and SO-CAL, where the performance tests showed LexiPro-SM was more effective than SO-CAL by producing higher accuracy results than SO-CAL. Furthermore, a case study was performed using LexiPro-SM sentiment analysis between two airlines in Malaysia (Airline A and Airline B), where the results show Airline A having more positive scores than Airline B which can allow us to conclude that the service provided by Airline A is better than Airline B.

University of Malaya

REFERENCES

Ahire, S. (2014). A survey of sentiment lexicons. Retrieved from http://www.cfilt.iitb.ac.in/resources/surveys/Sentiment-Lexicons-Sagar- Ahire.pdf

Aisopos, F., Tzannetos, D., Violos, J. & Varvarigou,T. (2016).Using N-Gram Graphs for Sentiment Analysis: An Extended Study on Twitter. Pr*oceeding of Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference*, 44-51. doi:10.1109/BigDataService.2016.13

Al-Ayyoub, M., Essa, S.B., & Alsmadi, I. (2015). Lexicon-based sentiment analysis of Arabic tweets. *International Journal of Social Network Mining*, 2(2), 101-114. doi: http://dx.doi.org/10.1504/IJSNM.2015.072280

Al-Kabi, M.N. , Amal H. Gigieh, Izzat M. Alsmadi, Heider A. Wahsheh, & Mohamad M. Haidar. (2014). Opinion Mining and Analysis for Arabic Language. *International Journal of Advanced Computer Science and Applications*, 5, 5

Amiri, F., Scerri S., & Khodashahi M. (2015), Lexicon-based Sentiment Analysis for Persian Text, *Proceedings of Recent Advances in Natural Language Processing*, Hissar, 9–16

Andreevskaia , A. , Bergler, S., & Urseanu, M. (2007). All blogs are not made equal: Exploring genre differences in sentiment tagging of blogs.*ICWSM*

Arora, A., Patil, C., & Correia,S. (2017). Opinion Mining: An Overview. International *Journal of Advanced Research in Computer and Communication Engineering (ISSN: 2278-1021)*, 4(11), 94-98

Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: a case study. *Proceedings of Recent Advances in Natural Language Processing (RANLP-2005),Washington,USA*.Retrieved from https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/new_domain_sentiment.pdf

Avanco, L.V., & Nunes M.G.V. (2014). Lexicon-based Sentiment Analysis for Reviews of Products in Brazilian Portuguese. *Proceeding of Brazilian Conference on Intelligent Systems* 2014, 277-281.doi: 10.1109/BRACIS.2014.57

University of Malaya

Awachate, P.B. & Kshirsagar, V.P. (2016).Improved Twitter Sentiment Analysis Using N Gram Feature Selection and Combinations. *International Journal of Advanced Research in Computer and Communication Engineering*, 5(9), 154-157. Retrieved from:http://www.ijarcce.com/upload/2016/september-16/IJARCCE %2035.pdf

Bharti,S.K., Vachha, B., Pradhan, R.K., Babu,K.S. & Jena S.K. (2016).Sarcastic sentiment detection in tweets streamed in real time: a big data approach. *Digital Communications and Networks*, 2(3), 108-121. doi:http://dx.doi.org/10.1016/ j.dcan.2016.06.002

Beigi, G., Xia, H., Maciejewski, R., & Huan, L. (2016). An overview of sentiment analysis in social media and its applications in disaster relief. *An Environment of Computational Intelligence*, 639, 313-340. doi: 10.1007/978-3-319-30319-2_13

Benamara, F., Cesarano, C., Picariello, A., Reforgiato, D., & Subrahmanian, V.S. (2007). Sentiment analysis: Adjectives and adverbs are better than adjectives alone. *Proceedings of International Conference on Weblogs and Social Medi*a, ICWSM, Boulder, CO

Brody, S., & Diakopoulos, N. (2011). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2011), Stroudsburg, PA, USA*. Retrieved from https://pdfs.semanticscholar. org/5557/fb31fd3e852e8fe6ff902d017c5ac78b65a7.pdf

Bongirwar, V.K. (2015). A Survey on Sentence Level Sentiment Analysis. *International Journal of Computer Science Trends and Technology (IJCST)*, 3(3), 110-113. Retrieved from http://www.ijcstjournal.org/volume-3/issue-3/IJCST-V3I3P21 .pdf

Banea, Carmen, Mihalcea, R., & Wiebe, J. (2008). A bootstrapping method for building subjectivity lexicons for languages with scarce resources. *LREC'08*, 2, 764–767

Bhaskar, J., Sruthi, K., & Nedungadi, P. (2014). Enhanced sentiment analysis of informal textual communication in social media by considering objective words and intensifiers. *Recent Advances and Innovations in Engineering (ICRAIE)*, 2014. doi: DOI: 10.1109/ICRAIE.2014.6909220

Crossley, S.A., Kyle, K., & McNamara, D.S. (2016). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, 1-19. doi: 10.3758/s13428-016-0743-z

University of Malaya

Dadvar, M., Hauff, C. & Franciska, J.D. (2011) . Scope of negation detection in sentiment analysis. *Proceedings of the Dutch-Belgian Information Retrieval Workshop, DIR 2011, Amsterdam, the Netherlands*. 16-20

Das, O. & Balabantaray, R.C. (2014).Sentiment Analysis of Movie Reviews using POS tags and Term Frequencies. *International Journal of Computer Applications (0975 – 8887)*, 96(25), 36-41. Retrieved from : http://research.ijcaonline. org/volume96/ number25/pxc3897048.pdf

Dave, K., Lawrence, S., & David, M. P. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of International Conference on World Wide Web (WWW-2003), Budapest, Hungary*. Retrieved from http://www.kushaldave.com/p451-dave.pdf

Definition of 'not very'.COBUILD Advanced English Dictionary,HarperCollins Publishers. Retrieved 5,February,2017, from https://www.collinsdictionary.com/ dictionary/english/not-very

Duwairi, R.M. (2014). Arabic Sentiment Analysis using Supervised Classification. *Paper presented at The 1st International Workshop on Social Networks Analysis, Management and Security (SNAMS - 2014), Barcelona, Spain*. Retrieved from http://www.just.edu.jo/~rehab/c1.pdf

Elgamal, M. (2016).Sentiment Analysis Methodology of Twitter Data with an application on Hajj season. *International Journal of Engineering Research & Science (IJOER) ISSN - [2395-6992]*, 2(1), 82-87. Retrieved from http://www .ijoer.com/Paper-January-2016/IJOER-JAN-2016-22.pdf

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heil-man, M., Yogatama, D., Flanigan, J. & Smith, N.A. (2011). Part-of speech tagging for twit- ter: annotation, features, and experiments. *Proceedings of the 49th Annual Meeting of the ACL, in: HLT '11, ACL, Stroudsburg, PA, USA*, 2011, 42–47

Gitari, N.D., Zuping, Z., Damien, H., & Long, J. (2015). A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4), 215-230. doi: http://dx.doi.org/10.14257/ijmue.2015.10.4.21

Govindaraj, S., & Gopalakrishnan, K. (2016). Customer Product Reviews Using Acoustic and Textual Features. *ETRI Journal*, 38(3), 494-501. Retrieved from http://dx.doi.org/10.4218/etrij.16.0115.0684

University of Malaya

Hatzivassiloglou, V., & McKeown, K. (1997). Predicting the semantic orientation of adjectives. *Proceedings of 35th Meeting of the Association for Computational Linguistics, 174–181*

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04*, 04(2), 168

Hutto, C.J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the eighth international AAAI conference on weblogs and social media*, 216–225

Hogenboom, A., Bal, D., Frasincar, F., & Bal, M. (2013).Exploiting Emoticons In Polarity Classification Of Text . *Journal Of Web Engineering*. Retrieved from http://people.few. eur.nl/frasincar/papers/JWE2015/jwe2015.pdf

Jindal, N., & Bing, L. (2006a). Identifying comparative sentences in text documents. *Proceedings the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2006),Washington, USA*. Retrieved from https:// www.cs.uic.edu/~liub/publications/sigir06-comp.pdf

Jindal, N., & Bing, L. (2006b). Mining comparative sentences and relations. *Proceedings of National Conference on Artificial Intelligence (AAAI-2006), California, USA*. Retrieved from https://www.cs.uic.edu/~liub/publications/ aaai06-comp-relation.pdf

Kee, M.W, & Ghazali, M. (2011). Branding satisfaction in the airline industry: A comparative study of Malaysia Airlines and Air Asia. *African Journal of Business Management* , 5(8), 3410-3423.

Kennedy, A., & Diana, I. (2006). Sentiment classification of movie and product reviews using contextual valence shifters. *Computational Intelligence*, 22(2), 110-125.

Khan. A., & Baharudin, B. (2011). Sentence Level Semantic Orientation of Online Reviews and Blogs using SentiWordNet for Effective Sentiment Classification. *International Journal of New Computer Architectures and their Applications (IJNCAA)*, 1(2), 627-643.

Kim, S.M., & Hovy, E. (2004). Determining the sentiment of opinions. *Proceedings of COLING 2004*, 1367–1373

University of Malaya

Korkontzelosa, I., Nikfarjamb, A., Shardlowa, M., Sarkerb, A., Ananiadoua, S., & Gonzalezb, G.H. (2016). Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of Biomedical Informatics*,62, 148-158 doi:http://dx.doi.org/10.1016/j.jbi.2016 .06.007

Kotelnikov, E.V., Bushmeleva, N.A., Razova, E.V., Peskisheva T.A. & Pletneva M.V. (2016). Manually Created Sentiment Lexicons: Research and Development. *Proceedings of the International Conference "Dialogue 2016"*. Retrieved from http://www.dialog-21.ru/media/3402/kotelnikovevetal.pdf

Koto, F. & Adriani, M. (2015).HBE: Hashtag-Based Emotion Lexicons for Twitter Sentiment Analysis. *Proceedings of the 7th Forum for Information Retrieval Evaluation, At Gandhinagar, India*. doi: 10.1145/2838706.2838718

Kucuktunc, O., Cambazoglu, B. B., Weber, I., & Ferhatosmanoglu, H. (2012). A large-scale sentiment analysis for yahoo! answers. *Paper presented at the Web Search and DataMining (WSDM2012), Seattle, Washington*. 633-642

Liau, B.Y., & Tan,P.P. (2014). Gaining customer knowledge in low cost airlines through text mining. *Industrial Management & Data Systems*, 114(9), 1344-1359. doi: 10.1108/IMDS-07-2014-0225

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Chicago, Illinois: Morgan & Claypool Publishers.

Martín-Valdivia, M., Martínez-Cámara, E., Perea-Ortega, J., & Ureña-López , L.A. (2012). Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40, 3934–3942.

Mary, A.J.J, & Arockiam, L. (2016).A Framework for Aspect level Sentiment Analysis of Academic Results Data. International Journal of Recent Trends in Engineering & Research (IJRTER), 2(7), 14-19. Retrieved from: http://www .ijcstjournal.org/volume-3/issue-3/IJCST-V3I3P21.pdf

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093-1113. doi: http://dx.doi.org/10.1016/j.asej.2014.04.011

Minqing , H. ,& Bing, L. (2004). Mining and summarizing customer reviews. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Seattle, Washington, USA*. Retrieved from https://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf

University of Malaya

Muhammad, A., Wiratunga, N. & Lothian, R. (2015), Contextual sentiment analysis for social media genres. Knowledge-Based Systems . doi: http://dx.doi.org/10.1016 /j.knosys.2016.05.032

Narendra, B., Uday Sai, K., Rajesh, G., Hemanth, K., Teja, V.C., & Kumar, K.D. (2016). Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies. *I.J. Intelligent Systems and Applications*, 2016, 8, 66-70. doi: 10.5815/ijisa.2016.08.08

Narr, S., Hulfenhaus, M., & Albayrak, S. (2016). Language-independent twitter sentiment analysis. Knowledge Discovery and Machine Learning (KDML), LWA, 12-14.Retrieved from http://www.dailabor.de/fileadmin/Files/ Publikationen/ Buchdatei/narr-twittersentiment-KDML-LWA-2012.pdf

Nasukawa, T., & Jeonghee,Y. (2003). Sentiment analysis: Capturing favorability using natural language processing. *Proceedings of Second International Conference on Knowledge Capture in Proceedings, Florida, USA*. Retrieved from http://www.csce.uark.edu/~sgauch/5013NLP/S13/hw/swetha.pdf

Naz, R. & Khan, M.N.A. (2015). Rapid Applications Development Techniques: A Critical Review. *International Journal of Software Engineering and Its Applications*, 9(11), 163-176. doi:http://dx.doi.org/10.14257/ijseia.2015.9.11.15

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2007). Textual Affect Sensing for Sociable and Expressive Online Communication. *Proceeding of Second International Conference, ACII 2007 Lisbon, Portugal, 4738, 218–229*. doi: 10.1007/978-3-540-74889-2_20

Norman, G. J., Norris, C., Gollan, J., Ito, T., Hawkley, L., Larsen, J. & Berntson, G. G. (2011). Current emotion research in psychophysiology: *The neurobiology of evaluative bivalence. Emotion Review*, 3, 3349-359. doi: 10.1177/1754 073911402403

Oliveira, N., Cortez, P., & Areal, N. (2014). Automatic Creation of Stock Market Lexicons for Sentiment Analysis Using StockTwits Data. *Proceedings of the 18th International Database Engineering & Applications Symposium on - IDEAS Oporto, Portugal: ACM Press*, 14, 115–123 doi:10.1145/2628194.2628 235

Pang, B., Lillian, L., & Shivakumar Vaithyanathan. (2002). Thumbs up?: Sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), Stroudsburg, PA, USA*. Retrieved from http://www.cs.cornell.edu/home/ llee/papers/sentiment.pdf

University of Malaya

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin.

Paltogloua, G., Gobronb, S., Skowronc, M., Thelwalla, M., & Thalmannb, D. (2010). Sentiment analysis of informal textual communication in cyberspace. *Austrian Research Institute for Artifcial Intelligence, 1010 Vienna*

Phillips, W. (1999).Introduction to Natural Language Processing. *The Mind Project*. Retrieved 5,February,2017, from http://www.mind.ilstu.edu/curriculum/ protothinker/natural_language_processing.php

PHP SpellCheck. Retrieved 5,February,2017, from http://www.phpspellcheck.com

Polanyi, L., & Zaenen, A. (2006). Contextual Valence Shifters.*Computing Attitude and Affect in Text: Theory and Applications*,20,1-10. doi: 10.1007/1-4020-4102-0_1

Qadir, A. & Riloff, E. (2013). Bootstrapped Learning of Emotion Hashtags #hashtags4you. *Proceeding of 4th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA2013).* Retrieved from https://www.cs.utah.edu/~riloff/pdfs/wassa13-hashtags.pdf

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). A Comprehensive Grammar of the English Language. Longman, London.

Rahm, E., & Do,H.H. (2000). Data Cleaning: Problems and Current Approaches (2000). *IEEE Data Engineering Bulletin*. Retrieved from http://betterevaluation .org/sites/default /files/data_cleaning.pdf

Rajput, Q., Haider, S., & Ghani, S. (2016). Lexicon-Based Sentiment Analysis of Teachers' Evaluation. *Applied Computational Intelligence and Soft Computing*, 2016. Retrieved from http://dx.doi.org/10.1155/2016/2385429

Rastogi, S.S.K., Singhal, R. & Kumar, R.(2014).A Sentiment Analysis based Approach to Facebook User Recommendation. *International Journal of Computer Applications (0975 – 8887), 90(16)*, 21-25

Riloff, E., Patwardhan, S., & Wiebe, J. (2006). Feature subsumption for opinion analysis. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2006), Sydney, Australia*. Retrieved from http://anthology.aclweb.org/W/W06/W06-1652.pdf

University of Malaya

Riloff, E., Qadir, A., Surve, P., & Silva, L.D. (2013). Sarcasm as Contrast between a Positive Sentiment and Negative Situation. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 704–714. Retrieved from http://www.anthology.aclweb.org/D/D13/D13-1066.pdf

Rodrigues, R.G., Dores, R.M.D., Camilo-Junior, C.G., & Rosa, T.C. (2016). SentiHealth-Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks. *International Journal of Medical Informatics*,85,80-95.

Salas-Zárate, M.D.P, López-López, E., Valencia-García, R., Aussenac-Gilles, N., Almela, Á., & Alor-Hernández, G. (2014). A study on LIWC categories for opinion mining in Spanish reviews. *Journal of Information Science ISSN 0165-5515*, 40(6), 1-13.

Sauri, R. (2008). *A Factuality Profiler for Eventualities* (Doctoral dissertation, Brandeis University,Waltham). Retrieved from http://www.cs.brandeis.edu/~roser/ pubs/sauriDiss_1.5.pdf

Schrauwen, S. (2010) . Machine learning approaches to sentiment analysis using the Dutch Netlog Corpus. *Computational Linguistics and Psycholinguistics Research Center (CTRS-001) , Antwerp, Belgium*. Retrieved from http://www. clips.ua.ac.be/sites/default/files/ctrs-001-small.pdf

Seerat, B., & Azam, F. (2012). Opinion Mining: Issues and Challenges (A survey). *International Journal of Computer Applications (0975 – 8887),* 49(9), 42-51. Retrieved from https://pdfs.semanticscholar.org/fce6/cd80be4abf10650 80a3d2a 6eeffb70 ba4837.pdf

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & HerreraViedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311, 18-38. doi: http://dx.doi.org/10.1016/j.ins.2015.03. 040

Shelke, N.M, Deshpande, S., & Thakre, V. (2012). Survey of Techniques for Opinion Mining. *International Journal of Computer Applications (0975 – 8887),* 57(13), 30-35. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi= 10.1.1.252.9648&rep=rep1&type=pdf

Soni, V., & Patel, M.R. (2014). Unsupervised Opinion Mining from Text Reviews Using SentiWordNet. *International Journal of Computer Trends and Technology (IJCTT) ISSN: 2231-2803*, 11(5), 234-238

Subrahmanian, V. S. & Reforgiato, D. (2008). Ava: Adjective-verb-adverb combinations for sentiment analysis. *Intelligent Systems*, 23(4):43–50

University of Malaya

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2), 267–307.

Taboada, M., Anthony, C., & Voll, K. (2006). Creating semantic orientation dictionaries. *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC),* 427–432

Tarlekar, A., & Kodmelwar M.K. (2015). Sentiment Analysis of Twitter Data from Political Domain Using Machine Learning Techniques. *International Journal of Innovative Research in Computer and Communication Engineering,* 3(6), 5590-5597. Retrieved from: https://www.ijircce.com/upload/2015/june/84_29_Sentiment.pdf

Thelwall, M., Buckley, K., Paltoglou, G., Cai,D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology archive*, 61 (12), 2544-2558.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, 63(1), 163-173. Retrieved from: http://sentistrength.wlv.ac.uk/ documentation/ SentiStrengthChapter.pdf

Turney, P.D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002), Philadelphia, Pennsylvania, USA*. Retrieved from http://www.aclweb.org/anthology/P02-1053.pdf

Turney,P. & Littman, M. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information System*, 21(4):315–346

Troussas, C., Virvou, M., Espinosa, K.J., Llaguno, K., & Caro, J. (2013).Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning. Proceedings of Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International. Retrieved from https://pdfs.semanticscholar.org/8a1f/a9ceee8332f35c118504f237c536cd500112.pdf

Vohta, S.M., & Teraiya, J.B. (2013). A Comparative Study of Sentiment Analysis Techniques. *Journal of Information, Knowledge And Research In Computer Engineering*, 2(02), 313-317. Retrieved from http://www.ejournal.aessangli.in/ ASEEJournals/CE63.pd

University of Malaya

Waldo, J. (2006). On System Design. *Sun Labs*. Retrieved from http://scholar.harvard.e du /files/waldo/files/ps-2006-6.pdf

Wang, H., Castanon, J.A & Jose,S. (2015). Sentiment Expression via Emoticons on social Media, 2015, *IEEE*

Weber, I., Ukkonen, A., & Gionis, A. (2012). Answers, not links: Extracting tips from yahoo! answers to address how-to web queries. *Proceedings of the fifth ACM international conference on web search and data mining (WSDM '12)*, 613-622

Weaver, J., & Tarjan, P. (2012). Facebook Linked Data via the Graph API. *IOS Press*, 1, 1-6. Retrieved from http://www.cs.rpi.edu/~weavej3/papers/swj2012-fbld.pdf

Wiebe, J.M. (2000). Learning subjective adjectives from corpora. *Proceedings of Seventeenth National Conference on Artificial Intelligence (AAAI-2000), California, USA*. Retrieved from https://people.cs.pitt.edu/~wiebe/pubs/papers /aaai2000.pdf

Wunnava, G. (2015, June). Applying Machine Learning to Text Mining with Amazon S3 and RapidMiner. *AWS Big Data Blog*. Retrieved from: https://aws.amazon. com/blogs/big-data/applying-machine-learning-to-text-mining-with-amazon-s3-and-rapidminer/

Wu, X., Hai-tao, L., & Shao-jian, Z. (2015). Sentiment analysis for Chinese text based on emotion degree lexicon and cognitive theories. *Journal of Shanghai Jiaotong University (Science).*20(1), 1-6

Xiong, H.,Steinbach, M., Pandey, G., & Kumar, V. (2006).Enhancing Data Analysis with Noise Removal. *IEEE Transactions on Knowledge and Data Engineering*, 18(3), 304-318

Zafarani, R., Abbasi, M.A, & Liu. H. (2014). Social Media Mining: AnIntroduction. *Cambridge University Press*, 2014. Retrieved from http://dmml.asu.edu/ smm/SMM.pdf

Ziora, L. (2016). The Sentiment Analysis as a tool of Business analytics in ontemporary organizations. *Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach (ISSN 2083-8611), 281, 234-241*. Retrieved from http://www.ue.katowice.pl/fileadmin/user_upload/ wydawnictwo /SE_Artyku%C5%82y_271_290/SE_281/19.pdf

University of Malaya