SARCASM DETECTION AND CLASSIFICATION TO SUPPORT SENTIMENT ANALYSIS: A STUDY IN MALAY SOCIAL MEDIA

MOHD SUHAIRI BIN MD SUHAIMIN

FACULTY OF COMPUTING AND INFORMATICS

UNIVERSITI MALAYSIA SABAH

2017

SARCASM DETECTION AND CLASSIFICATION TO SUPPORT SENTIMENT ANALYSIS: A STUDY IN MALAY SOCIAL MEDIA

MOHD SUHAIRI BIN MD SUHAIMIN

THESIS SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF MASTER OF SCIENCE

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH

2017

CERTIFICATION

NAME : **MOHD SUHAIRI BIN MD SUHAIMIN**

MATRIC NO. : **MI1511003T**

TITLE : **SARCASM DETECTION AND CLASSIFICATION TO SUPPORT SENTIMENT ANALYSIS: A STUDY IN MALAY SOCIAL MEDIA**

DEGREE : **MASTER OF SCIENCE (COMPUTER SCIENCE)**

VIVA DATE :

CERTIFIED BY

1. **SUPERVISOR**

   Dr. Mohd Hanafi bin Ahmad Hijazi Signature

   .....

2. **CO-SUPERVISOR**

   Assoc. Prof. Dr. Rayner Alfred Signature

   .....

DECLARATION

I hereby declare that this thesis and the works presented in it are my own and have been performed by me as the result of my original research. The recorded result and finding have not been submitted previously for a higher degree in any universities. The material in this thesis is my own except for quotations, equations, summaries and references, which have been duly acknowledged.

30 July 2017 .....

Mohd Suhairi bin Md Suhaimin

MI1511003T

ACKNOWLEDGMENT

First and foremost, *"Alhamdulillah"*, praise to Allah for giving me the good health and easing my journey to completion. I would like to express my gratitude to my wife, Dr. Nor Asyikin Yahya, for moral and financial support. This would not have been possible without you. Also, thanks go to my children, Furqan, Insyirah and Fatihah, my parents and mother-in-law for being patient and understanding through the tough times.

My deepest gratitude goes to my helpful supervisor, Dr. Mohd Hanafi Ahmad Hijazi for constantly provided me consultation and guidance to pursue this research. The experience sharing, critical thinking and research expertise have helped me in finding solution, conducting experiment and finally accomplished this research. May Allah grant you the best, *"Jazakallahu khairan kathira"*.

I also extend my gratitude to Assoc. Prof. Dr. Rayner Alfred for giving me advice in pursuing the experiment, and Prof. Frans Coenen for giving insight to improve the quality of this research and co-authoring the publication article.

Finally, my gratitude goes to University Malaysia Sabah for funding this research through UMSGreat, Grant GUG0061-TK-2/2016 and Ministry of Higher Education Malaysia that has given me the opportunity to pursue this research.

MOHD SUHAIRI BIN MD SUHAIMIN

30 July 2017

ABSTRACT

The classification of users' sentiment from social media data can be used to determine public opinion on certain issues. The presence of sarcasm in text may hamper the performance of sentiment analysis. This thesis presents research work conducted on sarcasm detection and classification to support sentiment analysis. A Malay social media dataset, specifically focused on economic and political domain, was acquired from public comments posted on Facebook. The proposed work consists of two phases: (i) sarcasm detection and (ii) sentiment analysis with sarcasm detection and classification. In the first phase, the development of a mechanism for detecting sarcasm on bilingual data was explored. To achieve this, a feature extraction process was proposed to identify sarcasm features. Five feature categories of that can be extracted using natural language processing were considered: lexical, pragmatic, prosodic, syntactic and idiosyncratic. A non-linear Support Vector Machines classifier was employed to measure the performance of the features using the adopted evaluation metric, average F-measure. The best-performing features were then used as input for the second phase. In the second phase, a framework for sentiment analysis that considers sarcasm detection and classification was proposed. The framework consists of six modules, namely preprocessing, feature extraction, feature selection, sentiment classification, sarcasm detection and classification, and actual sentiment classification. Results obtained from the evaluation conducted demonstrate that the proposed features and framework are able to improve the performance of sentiment analysis. The best performance for sarcasm detection was found using a combination of syntactic, pragmatic, and prosodic features with an average F-measure score of 0.852. The best result of sentiment classification using the proposed framework, considering both sarcasm detection and classification, recorded an average F-measure score of 0.905, outperforming the baseline sentiment classification score of 0.839.

ABSTRAK

PENGESANAN DAN KLASIFIKASI SARKASME UNTUK MENYOKONG ANALISIS SENTIMEN: SATU KAJIAN DALAM MEDIA SOSIAL MELAYU

Klasifikasi sentimen oleh pengguna-pengguna daripada data media sosial boleh digunakan untuk mendalami pendapat awam mengenai isu-isu tertentu. Kehadiran sarkasme dalam teks mungkin menjejaskan prestasi analisis sentimen. Tesis ini membentangkan kerja penyelidikan yang dijalankan ke atas pengesanan dan klasifikasi sarkasme untuk menyokong analisis sentimen. Satu set data media sosial Melayu, khususnya tertumpu pada domain ekonomi dan politik, telah diperolehi dari komen awam yang diposkan di "Facebook". Kerja yang dicadangkan terdiri daripada dua fasa: (i) pengesanan sarkasme dan (ii) analisis sentiment dengan pengesanan dan klasifikasi sarcasme. Dalam fasa pertama, pembangunan satu mekanisme untuk mengesan sarkasme dari data dwibahasa telah diterokai. Untuk mencapai matlamat ini, satu proses pengekstrakan fitur telah dicadangkan bagi mengenalpasti fitur-fitur sarkasme. Lima kategori fitur yang boleh diekstrak menggunakan pemprosesan bahasa tabii telah dipertimbangkan iaitu leksikal, pragmatik, prosodi, sintaksis dan idiosinkratik. Satu pengelas bukan linear "Support Vector Machines" telah diguna pakai untuk mengukur prestasi fitur-fitur tersebut dengan mengguna pakai penilaian metrik, purata "F-measure". Fitur-fitur yang berprestasi terbaik telah dijadikan sebagai input untuk fasa kedua. Dalam fasa kedua, satu rangka kerja untuk analisis sentimen yang mengambilkira pengesanan dan klasifikasi sarkasme telah dicadangkan. Rangka kerja ini terdiri daripada enam modul iaitu pra pemprosesan, pengekstrakan fitur, pemilihan fitur, klasifikasi sentimen, pengesanan dan klasifikasi sarkasme, dan klasifikasi sentimen sebenar. Keputusan yang diperolehi dari penilaian yang dijalankan menunjukkan bahawa fitur-fitur dan rangka kerja yang dicadangkan dapat meningkatkan prestasi analisis sentimen. Fitur-fitur yang berprestasi terbaik untuk pengesanan sarkasme adalah dari gabungan fitur sintaksis, pragmatik dan prosodi dengan skor purata "F-measure" berukuran 0.852. Keputusan terbaik bagi klasifikasi sentimen menggunakan rangka kerja yang dicadangkan, dengan mempertimbangkan pengesanan dan klasifikasi sarkasme, merekodkan skor purata "F-measure" berukuran 0.905, mengatasi skor garis asas klasifikasi sentimen berukuran 0.839.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIS	Artificial Immune System
ADJ	Adjective
ADV	Adverb
ANN	Artificial Neural Network
BOW	Bag-Of-Words
CHI	Chi-Square
CNB	Complement Naïve Bayes
DF	Document Frequency
DT	Decision Tree
FS-INS	Feature Selection Immune Network System
GI	Gini Index
GR	Gain Ratio
IG	Information Gain
IMDB	Internet Movie Database
k-CV	k-fold Cross Validation
k-NN	k-Nearest Neighbors
LIWC	Linguistic Inquiry and Word Count
LogR	Logistic Regression
ME	Maximum Entropy
MI	Mutual Information
MM	Markov Models
MPQA	Multi-Perspective Question Answering
NB	Naïve Bayes
NBM	Naïve Bayes Multinomial
NLP	Natural Language Processing
NLTK	Natural Language Toolkit

OneR	One Rule
PCA	Principal Component Analysis
POS	Part-Of-Speech
RBF	Radial Basis Function
RF	Relief-F
SA	Sentiment Analysis
SMO	Sequential Minimal Optimization
SVM	Support Vector Machines
TF-IDF	Term Frequency - Inverse Document Frequency

LIST OF SYMBOLS

γ	Gamma
C	Cost error
d	degree
F	F-measure
F_{avg}	Average F-measure
P	Precision
R	Recall
r	reasonable number

LIST OF APPENDIX

CHAPTER 1

INTRODUCTION

1.1 Overview

Sentiment Analysis (SA) classifies user generated content such as opinion, believe, views and emotion in written text towards their entities and attributes (B. Liu, 2015). Generally, SA focuses on opinions in generated content whether it expresses positive or negative sentiments. Focus can be categorized into sentiment orientation (positive, negative or neutral), sentiment intensity (different level of strength), and sentiment rating (expression degree such as 1-5). Different levels of analysis at the document, aspect and sentence levels have been investigated as found in the literature. The latter is the focus of the work presented in this thesis. Sentence level sentiment analysis can be defined determining whether an opinionated sentence expresses a positive or negative opinion (Indurkhya & Damerau, 2010; B. Liu, 2015).

The rise of the social networking platform and web technologies has encouraged users to create and share content in the form of opinion, believe, views and emotion. This user-generated content is increasing vast on social networking media such as Facebook, Twitter, Google+ and forum discussion. Research in sentiment analysis has been extended to learn and gain knowledge and benefits from user generated content. It has been used extensively in review summarization, decision making, ranking and recommender systems, and been applied in industry, organization, government and business (Farzindar & Inkpen, 2015). Social media sentiment analysis applications include predicting voting intention for political benefit, security defense application to identify national threats, and media monitoring for business intelligence.

Social media SA's primary issues include classification accuracy, cross language SA, informal medium type and ambiguity. The classification accuracy issue concerns a high percentage of sentiments incorrectly classified as neutral (Madhoushi, Hamdan, & Zainudin, 2015). Cross language SA issue includes lack of resource for applying SA in multiple language or target language in non-English that resulting in weak prediction performance (Dashtipour et al., 2016; Korayem, Aljadda, & Crandall, 2016). Informal medium issues include incorrect words and limitation of length for providing opinion (Giachanou & Crestani, 2016). Ambiguity concerns figurative language such as sarcasm to convey the actual meaning sentiment in delivering opinion (Balahur & Jacquet, 2015; Ravi & Ravi, 2015). The last factor has been identified as the most significance challenge in social media SA (Farzindar & Inkpen, 2015; Joshi, Bhattacharyya, & Carman, 2016; B. Liu, 2015; Serrano-Guerrero, Olivas, Romero, & Herrera-Viedma, 2015; Weitzel, Prati, & Aguiar, 2016).

In communication, sarcasm is used to express opinion that is different from the initially apparent meaning (Ghosh, Guo, & Muresan, 2015). Therefore, sarcasm existence in sentences tend to confuse the SA system and misclassify the sentiment. In an automatic system, detecting sarcasm from genuine subjectivity opinion, an opinion that contains personal orientation or sentiment towards an entity, is tough. Sarcasm is difficult to resolve as words used in a comment are usually associated to the opposite polarities. Failure to detect sarcasm in the sentences will affect the actual sentiment prediction and misclassification (Farzindar & Inkpen, 2015). Example of sarcasm is "hmmm…, soon the college fee will rise, good job", which could be classified as positive since the words used are usually presenting positive sentiment. However, it is obvious that the comment carries negative sentiment.

This thesis addresses a number of issues raised due to sarcasm in SA and proposes several solutions to overcome those issues (see Section 1.2 and 1.3 for detail). In the literature, most work has focused on the detection of sarcasm, including identification of features to recognize sarcasm, techniques to improve detection and classification, and a background study related to linguistic and computational sarcasm. The work presented in this thesis address the sarcasm detection issue in bilingual social media texts, and subsequently employs sarcasm detection to support sentiment analysis.

This introductory chapter has been organized as follows. Section 1.2 describes the research motivation and Section 1.3 elaborates the research objectives. Section 1.4 briefs the scope of the research and Section 1.5 describes the research methodology. Section 1.6 details the evaluation criteria for the research. Section 1.7 presents research contributions and Section 1.8 provides details of the published work as a result of this research. Section 1.9 describes the organization of the thesis.

1.2 Research Motivation

Detecting sarcasm (and also SA) is made more complex when social media texts are written in more than one language (bilingual). Misspelled words, shortened word forms and stylistic text coupled with the use of dual language are commonplace, and it is not unusual to mix different languages. The crucial part is to extract the features that could better identify the sarcasm content.

To the best knowledge of the author, no work has been done to adopt sarcasm detection and classification to support SA. The challenge is thus to identify mechanism of how this could be done. Therefore, the motivation of this research is to produce an approach for social media SA on bilingual text that considers sarcasm detection and classification to make sentiment prediction. It is conjectured that by considering sarcasm content, better SA performance could be produced.

1.3 Research Objective

Given the research motivation described in Section 1.2, the main research question for the work presented in this thesis is: *"What is the appropriate approach to classify sentiment using sarcasm detection and classification for bilingual social media data?"* Two subsidiary questions raised from this research question are:

1. *"What are the features that can be extracted from social media containing bilingual data that can better identify sarcasm features?"*
2. *"How the sarcasm detection and classification can be employed into SA system?"*

Based on the identified research questions, three research objectives were derived:

1. To investigate and identify features for sarcasm detection on bilingual social media data.
2. To investigate and implement a framework for SA with sarcasm detection and sarcasm classification to produce better sentiment classification's performance.
3. To evaluate the results of the proposed approaches in (1) and (2).

1.4 Research Scope

The preliminary focus of this research is sarcasm on bilingual social media data. Malay social media data was chosen rather than English for showing high levels of bilingual comments in sentence form (Samsudin, Puteh, & Hamdan, 2011; Samsudin, Puteh, Hamdan, & Nazri, 2013a). According to Dress, Kreuz, Link, and Caucci (2008), usage and factors for sarcasm also vary according to geographical area; thus the study and result might be slightly different from the English or other language. However, the approach, techniques and methodology could be useful for adoption and implementation. Sentence levels of sentiments are concentrated on for this foundation investigation, and only comments from discussions are considered. In depth topics of discussion such as topic-based SA or contextual features such as commentator profiles are beyond the scope of this research.

With respect to classification process, a supervised machine learning approach is used in this work. Supervised machine learning has been shown to be more effective in sentiment classification than a lexicon-based approach (Blinov, Klekovkina, Kotelnikov, & Pestov, 2013; Hailong, Wenyan, & Bo, 2014; Yusof, Mohamed, & Abdul-Rahman, 2015). The classification algorithm examined is Support Vector Machines (SVM) due to its superiority over other classification algorithms in SA and sarcasm detection tasks (Bouazizi & Ohtsuki, 2015; Chandrakala & Sindhu, 2012; Ghosh et al., 2015; Hailong et al., 2014; Medhat, Hassan, & Korashy, 2014; Muresan, Gonzalez-Ibanez, Ghosh, & Wacholder, 2015; Yusof et al., 2015).

1.5 Research Methodology

To achieve the research objectives of the work in this thesis, a methodology of two phases is set up, with which an additional preliminary phase is added. The overall methodology is illustrated in Figure 1.1.

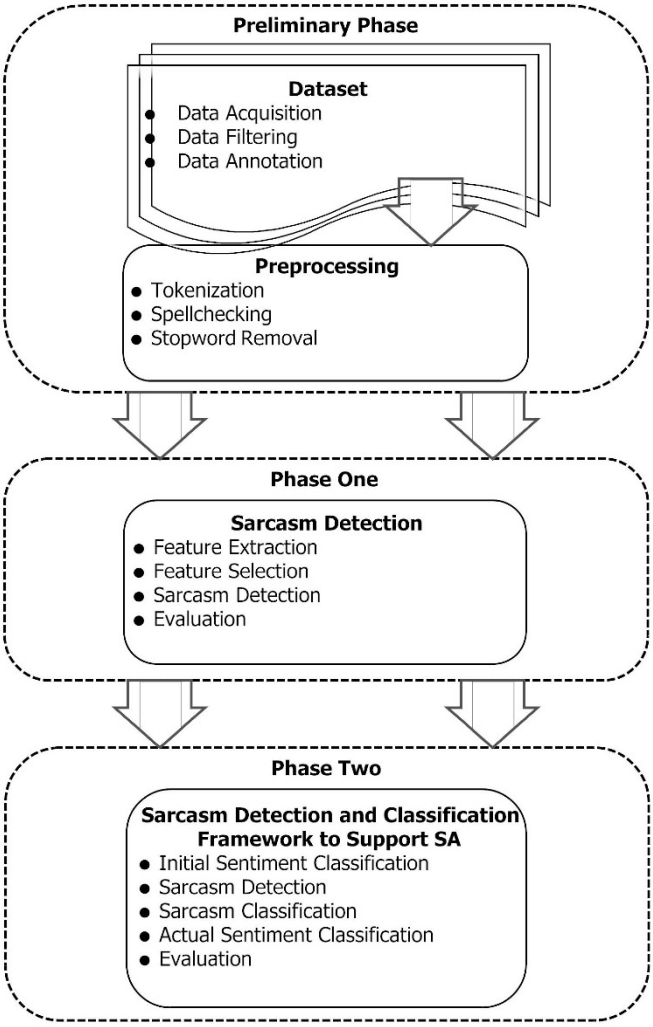


Figure 1.1: Methodology phases for research

The preliminary phase is data acquisition, filtering and annotation followed by data preprocessing. Tokenization, spellchecking and stopwords removal is conducted in the preprocessing stage. Details of the preliminary phase is presented in Chapter 3.

Phase one: Sarcasm detection. Investigation to identify features that best represent sarcasm content in bilingual social media data is investigated in this phase. These features are extracted from the preprocessed comment. Feature selection is then performed and sarcasm detection is then conducted using supervised classification. Details of this phase are described in Chapter 4.

Phase two: Sarcasm detection and classification framework to support SA. A framework is proposed to embed the detection and classification of sarcasm into SA system that could improve SA performance. The proposed framework consists of sentiment classification, sarcasm detection, sarcasm classification and actual sentiment classification modules. The best features identified in phase one is used to perform the sarcasm detection and sentiment classification. Details of this phase are described in Chapter 5.

1.6 Evaluation Criteria

The focus of the work proposed in this thesis is identifying the best results for sentiment classification for SA system by using the best result of sarcasm detection. To evaluate the proposed approaches, an evaluation metric of average F-measure (F_{avg}) is considered. F_{avg} is an average of F-measure evaluated binary class of positive and negative or sarcastic and non-sarcastic. The formula is given as follows:

$$F_{avg} = \frac{F_i \times c_i + F_j \times c_j}{c_i + c_j} \tag{1.1}$$

where F_i is the F-measure for class i and c_i is the number of documents in class i, while F_j is the F-measure for class j and c_j is the number of documents in class j.

F-measure (F), the harmonic mean of precision (P) and recall (R) (Manning, Raghavan, & Schütze, 2008) for each class i and j, is given as follows:

$$F = 2 \times \frac{P \times R}{P + R} \quad (1.2)$$

where P is the precision and R is the recall for each class when equally weighted.

P in detail is the measurement of the correctly classified document (for positive or negative) over the total number of document being classified (for positive or negative). The formula is:

$$P = \frac{\Sigma true\ positive}{\Sigma true\ positive + \Sigma\ false\ positive} \quad (1.3)$$

where $\Sigma\ true\ positive$ represents the sum of all document that is correctly classified and $\Sigma\ false\ positive$ represents the sum of all document which is incorrectly classified.

R is the measure of the number of correctly classified documents over the total documents marked as having respective polarity. The formula is:

$$R = \frac{\Sigma true\ positive}{\Sigma true\ positive + \Sigma\ false\ negative} \quad (1.4)$$

where $\Sigma\ true\ positive$ represents the sum of all correctly classified document and $\Sigma\ false\ negative$ represents the sum of all the documents that the system has misclassified.

1.7 Research Contribution

The main contributions of the research work presented in this thesis can be summarized as follows:

1. A process to identify sarcasm features for sarcasm detection on bilingual social media data.
2. A framework for SA that considers sarcasm detection and classification.

1.8 Published Work

Some of the work described in this thesis has been published in a number of refereed publication as itemized below.

1. **Conference Papers**

   (a) Md Suhaimin, M. S., Ahmad Hijazi, M. H., Alfred, R., & Coenen, F. (2016). *Mechanism for Sarcasm Detection and Classification in Malay Social Media.* Paper presented at the 3rd International Conference on Computational Science and Technology (ICCST), Kota Kinabalu. This paper presents some initial work to identify the best mechanism for sarcasm detection and classification, specifically for Malay social media bilingual dataset.

   (b) Md Suhaimin, M. S., Ahmad Hijazi, M. H., Alfred, R., & Coenen, F. (2017). *Natural Language Processing Based Features for Sarcasm Detection: An Investigation Using Bilingual Social Media Texts.* Paper presented at the Proceeding of the The 8th International Conference on Information Technology (ICIT17), Amman, Jordan. This paper presents the performance of the proposed feature extraction process for sarcasm detection. Natural Language Processing (NLP) based features were used in this paper.

2. **Research Competition**

   (c) Ahmad Hijazi, M. H., Md Suhaimin, M. S., Alfred, R., & Coenen, F. (2016). *Mechanism for Sarcasm Detection and Classification in Malay Social Media.* Part of paper in (a) has been presented at research competition of "Pertandingan Penyelidikan dan Rekacipta UMS (PEREKA 2016), Kota Kinabalu". This paper has been awarded silver medal.

1.9 Thesis Organization

This thesis consists of six chapters. The rest of this thesis is organized as follows:

Chapter 2 presents the literature review of the related research on SA and sarcasm detection. Previous works on general framework for SA and sarcasm detection is presented in this chapter. Further, this chapter describes the method with the feature used, and algorithms for SA and sarcasm detection using supervised learning approach, both English and non-English language study. Selected classification algorithms considered in this thesis are also presented in this chapter.

Chapter 3 describes the dataset used and preprocessing stage for the preliminary phase of the methodology. Data acquisition, filtration and annotation performed are also described.

Chapter 4 presents the proposed approach to identify best feature for sarcasm detection on bilingual social media data. A series of experiment is conducted to find the best-performing features to detect sarcasm. The evaluation is also presented in this chapter.

Chapter 5 describes the proposed SA framework using sarcasm detection and classification. The evaluation and comparison of the result produced by the framework are presented in this chapter.

Chapter 6 summarizes research work and identifies possible future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The SA consists of a number of tasks including data preprocessing, extraction of features and sentiment analysis or prediction. The output of the SA, in most cases, is in the form of prediction of the text's sentiment; positive or negative. However, sarcastic contents are common in social media data, which could affect the SA performance. Therefore, some previous work has extended the prediction to three classes: positive, negative and sarcasm. Nevertheless, sarcasm itself could contain either positive or negative sentiment. Hence, the work presented in this thesis proposes an approach to perform sarcasm detection and classification, which subsequently will be used to predict the text's sentiment. This chapter presents a review of relevant previous work on framework for SA and sarcasm detection using supervised learning. Section 2.2 elaborates a framework of SA with details including a supervised learning approach applied to English and non-English data (including Malay). Section 2.3 describes the framework of sarcasm detection and classification using supervised learning approach. Section 2.4 reviews the selected supervised classification algorithm, which is SVM and Section 2.5 summarizes this chapter.

2.2 Framework of Sentiment Analysis

This section presents the literature review of general framework for SA using supervised machine learning approach. An overview of the general framework for the SA is provided

in Sub-section 2.2.1. Details of the SA using supervised learning approach are reviewed in Sub-section 2.2.2. SA using supervised learning approach for Non-English language is reviewed in Sub-section 2.2.3, and specifically Malay language is reviewed in Sub-section 2.2.4.

2.2.1 Overview of Framework

A machine learning approach, either supervised, semi-supervised, or unsupervised, differs from a lexicon-based approach in terms of two major factors. First, the machine learning approach requires training data for classification. Second, machine learning algorithms is used to perform classification (Madhoushi et al., 2015; Yusof et al., 2015). For learning approach, supervised requires data to be labelled, while unsupervised does not need data to be labelled but finds a hidden structure using learning algorithms (Chew, 2013; Yusof et al., 2015). Semi-supervised learning is a combination of those two, where joint distribution is used to perform classification (Madhoushi et al., 2015). Among three discussed approaches, supervised learning approach has been showing the effectiveness in sentiment classification task (Blinov et al., 2013; Hailong et al., 2014; Yusof et al., 2015). Therefore, the focus of this research has employed this approach.

Generally, SA focuses on opinions in the generated content, whether positive or negative. Focus can be categorize into sentiment orientation (positive, negative or neutral), sentiment intensity (different level of strength) and sentiment rating (expression degree such as 1-5). Different levels of analysis at the document, sentence and aspect levels have been investigated through data mining, natural language processing, web mining and information retrieval as well as pragmatics and semantics. At the sentence level, classification involves two steps: subjectivity classification and sentiment classification. Subjective classification concerns personal orientation, from objective opinion containing factual information. The sentiment classification later determines subjective opinion, whether a positive or negative sentiment (Medhat, Hassan, et al., 2014). The formal definition of sentiment classification in sentence level of SA thus can be formulated as follows: given an opinionated sentence, determine

whether it expresses a positive or negative opinion (Indurkhya & Damerau, 2010; B. Liu, 2015).

Sentiment classification for SA is generally a text classification problem and closely related to Natural Language Processing (NLP) problem which any existing supervised learning approach can be applied directly (Agarwal & Mittal, 2016; B. Liu, 2015). The general approach for sentiment classification using machine learning is comprised of preprocessing of text, extraction and transformation of features from text, and finally classification using robust machine learning method. Figure 2.1 shows the framework for general sentiment classification approach for SA.

Figure 2.1: General Framework for SA

First, the labelled data for training is preprocessed. Then, feature extraction is performed to extract the feature. Finally, the classification will classify a feature into its corresponding class. Details of the approach are provided below:

a. Preprocessing. The preprocessing may consist of data cleaning or filtering, tokenization, standardization, normalization, stopwords removal, stemming and lemmatization.
    i. Data cleaning is needed to remove the unnecessary characters like symbols and filtering data such as pictures, videos or links (Newell, Potharaju, Xiang, & Nita-Rotaru, 2014).
    ii. Tokenization is the breaking or splitting of each sentence into a list of words. Each word that represents the sentence is called a token (Chandrakala & Sindhu, 2012).
    iii. Standardization is performed to produce same type of structure such as lowercase of text (Pustejovsky & Stubbs, 2012; Weiss, Zhang, & Indurkhya, 2015) and normalization creates equivalence canonical form

of class for the selected token (Aggarwal & Zhai, 2012; Manning et al., 2008). Normalization can be applied during preprocessing or feature transformation such as vector normalization or document length normalization.

iv. Stopword removal involves filtering only meaningful words and avoiding most common words such "is", "a" and "the".

v. Stemming usually involves removing affixes and leaving the root word or stem, whereas lemmatization remove inflection in words by leaving the lemma of the words. Stemming for example will return "sing" as a stem from the word "sings" and "singing". Lemmatization will return lemma word "sing" from "sings", "sang" and "sung" (Jurafsky, 2000; Manning et al., 2008).

b. Feature extraction. After preprocessing, features are extracted from the text through some functional mapping and transform it to the feature vector for weighting.

i. Vectorization. The most common extracted feature is in the form of bag-of-words (BOW) model, in which all tokens are considered and represented using a frequency of the same feature, and later weighted the using term frequency - inverse document frequency (TF-IDF) (Jurafsky, 2000; H. Liu & Motoda, 1998).

ii. Feature selection. Feature selection process is applied to select only the relevant and informative features to improve classification performance, reduce processing time or to gain knowledge about processing data (Guyon, Gunn, Nikravesh, & Zadeh, 2006). Various strategies have been adopted for feature selection, such as feature ranking or selecting only subset of features for classification. Feature selection, however, is optional.

c. Classification. Finally, the selected feature is trained using classification algorithm or classifier for classification. SVM, Naïve Bayes (NB) and Maximum Entropy (ME) are the most common supervised classifiers (Ravi & Ravi, 2015; Serrano-Guerrero et al., 2015; Yusof et al., 2015).

Medhat, Yousef, and Korashy (2014) utilized a framework to prepare corpora or dataset from social network sites for SA. The framework was used for sentiment classification on movie review data. Data from Twitter, Facebook, and Internet Movie Database (IMDB) websites were acquired and preprocessed. The preprocessing involved cleaning of the data by means of replacing negation and stopwords removal. The feature in the form of tagged part-of-speech (POS) was extracted from the dataset, where only the adjective and verb were selected. Features in the form of unigram and bigram were extracted. Supervised classifiers used for testing are NB and Decision Trees (DT). The dataset was split to training and testing for classification; the best result recorded using NB with bigram feature.

A concept of framework for SA has been proposed using NLP task to extract opinionated information from text (Cambria, Poria, Bisio, Bajpai, & Chaturvedi, 2015). The framework consists of eight modules: micro-text analysis, semantic parsing, subjectivity detection, anaphora resolution, sarcasm detection, topic spotting, aspect extraction and polarity detection. Each module aimed to normalize irregular text, deconstructing natural language text into concept, filtering non-opinionated text, resolving reference in the discourse, detecting sarcastic opinion and flip the polarity, contextualizing opinion to a specific topic, deconstructing text into different opinion target, and detecting polarity value for each opinion target respectively. Documents will need to undergo each module in order to extract the sentiment information.

The general framework above and the two proposed frameworks provide insight to the view of the framework for SA. Medhat, Yousef, et al. (2014) have modified the framework for sentiment classification by with dataset preparation and (Cambria et al., 2015) proposed a concept for framework using NLP task for SA. Using this guidance, any modification general framework can be generated for sentiment classification tasks.

2.2.2 Sentiment Analysis Using Supervised Learning Approach for English Language

This sub-section reviewed the SA approach using supervised machine learning for English language from the literature. First, the dataset domain with acquisition and

annotation are presented in Sub-subsection 2.2.2.1. Later, preprocessing is presented in Sub-subsection 2.2.2.2. Feature extraction and feature selection are presented in Sub-subsection 2.2.2.3 and Sub-subsection 2.2.2.4, respectively. Finally, classifier generation and classification are presented in Sub-subsection 2.2.2.5.

2.2.2.1 Dataset Acquisition and Annotation

Most of the early work on SA was applied to the business domain. Sentiment classification of movie reviews is most frequent focus in the domain. A few studies were also conducted on customers feedback on product review (Dang, Zhang, & Chen, 2010; Dave, Lawrence, & Pennock, 2003), travel domain review (Ye, Zhang, & Law, 2009) and news article review (Read, 2005). The objective was to classify positive or negative sentiment of the reviews towards a target attribute or entity. Datasets were acquired from reviewer comments on certain threads, then labelled or annotated. The process of labelling dataset can be done by using either manual annotation or rule-based, or crowdsourcing. Manual annotation is usually done by human annotator, rule-based by applying specific rule or algorithm that automatically labels the dataset, and crowdsourcing by using a service to assign task to group of people via the internet. Frequently for experimenting, the balance distribution of positive and negative label is chosen. However Lane, Clarke, and Hender (2012) experimented with using an imbalanced dataset manually annotated for sentiment classification of customer feedback. Pang, Lee, and Vaithyanathan (2002) first used balanced 700 positive and 700 negative movie reviews using manual annotation by three annotators. The third annotator was used to produce majority vote in case ties of the two annotators. Later the study (Pang & Lee, 2004) expanded to 1000 positive and 1000 negative review datasets available publicly that have been used for many later works. The focus on English reviews is longstanding as only English comments are selected during filtering, whereas non-English comments are filtered out (Read, 2005).

2.2.2.2 Preprocessing

As presented in Sub-section 2.2.1, preprocessing of text involving the removal of stopwords, stemming or lemmatization is a favorite choice of researchers. Only uncommon terms are left for feature extraction, thus reducing variation of presented text (Bakliwal et al., 2012; Dave et al., 2003; Kiritchenko, Zhu, & Mohammad, 2014; Mohammad, Kiritchenko, & Zhu, 2013; Moraes, Valiati, & Neto, 2013; Riloff, Patwardhan, & Wiebe, 2006). However, some researchers have not performed stopwords removal, stemming or lemmatization in their study for the reason that such original words might indicate sentiment orientation. Thus, variation is preserved (Pang et al., 2002; Pham, 2014; Ye et al., 2009).

2.2.2.3 Feature Extraction

Various features have been proposed in previous work which can be categorized as either BOW or NLP based feature. The most commonly used is BOW due to its simplicity yet efficiency in extracting features that represent the sentiment of the texts. BOW chose words that are considered equally important; it ignores word order and semantics of the sentence (Moraes et al., 2013; Newell et al., 2014; Pham, 2014; Xia & Zong, 2010). NLP based features, in contrast, take into account the word order and semantic representation of the sentences. The characteristic of chosen features to be extracted from language processing perspective could bring better meaning to represent the sentences for classifier generation (Bakliwal et al., 2012; Kiritchenko et al., 2014).

Five categories of NLP based feature have been used to extract the feature for classifier generation from the literature. The categories are lexical, syntactic, pragmatic, prosodic, and semantic.

a) Lexical features. This feature is the most common form of feature extracted using NLP. These features are used to represent the sentence's content using tokenized words in the form of n-grams, that can uncover information for processing sentence (Indurkhya & Damerau, 2010). Three types of n-gram are frequently used in the literature: 1-gram (unigram), 2-gram (bigram) and 3-gram (trigram). Consider the

three words from the sentence, "thank you government". Unigram represent the tokenized word for the feature as single token: "thank", "you", "government", while bigram represent it as coupled: "thank you" and "you government". So the trigram represents it as a triplet: "thank you government".

b) Syntactic feature. Syntactic features provide important information with regards to syntactic structure of documents. A common syntactic feature is tagged POS that may be associated with words in a document. Syntactic features such as adjectives, nouns and verbs have been shown to produced significant improvement, with respect to SA, when paired with word (word-tag pairing) as it identifies discriminant features compared to using a word or tag alone (Salvetti, Lewis, & Reichenbach, 2004; Xia & Zong, 2010). Examples of using word only includes: "thank", "government", while using tag only is: "VERB", "NOUN". Used of word-tag pairing is then: "thank-VERB", "government-NOUN".

c) Pragmatic features. This feature is intended to emphasis the meaning of the content of sentences (Daniel & James, 2009). Emoticons, punctuation, hashtag (#) and repeated words are examples of pragmatic features. Previous work found pragmatic such as emoticon, punctuation marks, hashtag and target word have contributed to better classification. However, it only works well when a large dataset is available. Analysis conducted on misclassified data has found that the existence of pragmatic features could probably signal sarcasm (Bakliwal et al., 2012; Read, 2005).

d) Prosodic features. This feature represents different pitches, loudness, timing and tempos in writing (Bikel & Zitouni, 2012). Interjections are an example of prosodic features. Some examples of interjection are: "argh", "aww" and "woops".

e) Semantic features. The semantic features require deep understanding of the relation words in the sentences or sentiment towards the entity and aspect. Consider a word "crash" that have multiple meaning related to the used toward the entity or aspect. "Crash" can mean accident, or also can mean attend a party without invited, or significant drop in stock market depend on the aspect that the sentence focused on. Thus, semantic features in sentences are deemed to represent the original sentiment expression into some kind of semantic meta language

(Indurkhya & Damerau, 2010; B. Liu, 2012). A number of dictionaries of semantic features are available such as Wordnet[1] and SentiWordnet[2] .

The extracted features are usually vectorized, then converted into a matrix with a numerical format. TF-IDF is widely used to convert the features into vectors (Moraes et al., 2013; O'Keefe & Koprinska, 2009; Pham, 2014; Sharma & Dey, 2012). In another work, Ng, Dasgupta, and Arifin (2006) performed vector normalization on document length that produced better performance result of SA. This normalization is mostly helpful when the features are varied across different documents (Manning et al., 2008; Weiss et al., 2015). Document to length normalization is defined as:

$$V_{norm} = \frac{V_{act} \times L_{avg}}{L_d} \qquad (2.1)$$

where V_{norm} represents normalized value of vector, V_{act} represent actual value of vector sparse represent the feature, L_{avg} represent average document length and L_d represent document length of the feature.

2.2.2.4 Feature Selection

Feature selection is often used to eliminate meaningless feature using statistical approach. A number of different feature selection techniques were employed in the literature such as Information Gain (IG) (Dang et al., 2010; Moraes et al., 2013; Riloff et al., 2006; Ye et al., 2009), Mutual Information (MI) (Xia & Zong, 2010) and Chi-Square (CHI) (Lane et al., 2012). (Selvi, Ahuja, & Sivasankar, 2015) have conducted various feature selection study for sentiment classification using IG, MI, CHI, Gain Ratio (GR), Relief-F (RF), One Rule (OneR) and Document Frequency (DF). Most the feature selection approach is directed at identifying the most indicative features by means of feature ranking or subset selection. The classification performance when consider

[1] https://wordnet.princeton.edu/
[2] http://sentiwordnet.isti.cnr.it/

feature selection usually better compared to without feature selection. However, some feature selection approaches do not perform well when employed on particular domains. Some studies have found that ineffective feature selection harms performance (Lane et al., 2012; Moraes et al., 2013; Selvi et al., 2015).

2.2.2.5 Classifier Generation and Classification

Various supervised classification algorithm or classifier had been chosen or compared to perform sentiment classification. The aim was to determine the best classification algorithm that can improve classification performance. Some of favorable algorithms are SVM, NB, ME, DT, Markov Models (MM), k-Nearest Neighbors (k-NN) and Artificial Neural Networks (ANN). The evaluation method is used to perform the classification either using splitting the features into training and testing proportion or by using k-fold Cross Validation (k-CV). k-CV divides the dataset into k partitions. The evaluation will be repeated k times. In each repetition, one partition of the data will be considered as test data and the remaining as training data. The classification performance of each evaluation will then be recorded.

Table 2.1 shows the summary of selected previous work of SA on English language along with non-English dataset using supervised classification algorithm. The table presents the features extracted, feature selection used, advantage and disadvantage of each work. From Table 2.1, the most common feature extracted was lexical. Lexical feature type of unigram is largely extracted for training due to simplicity in representing tokenized words. The syntactic was the second common feature chosen to be extracted as showing discriminative feature. The common feature selection chosen were CHI and IG and most chosen classifier to perform classification was SVM. The approaches used have indicated that better performance can be achieved using a combination of features with feature selection and classifier generation. A number of issues also arose, such as the need for linguistic knowledge to detect sentiment (such as affected of sarcasm), size and imbalanced dataset, and ineffective feature selection. A review of previous work on SA for non-English language using the supervised learning approach is presented in the following subsection.

Table 2.1: Summary of Previous Work Using Supervised Learning Approach

Author	Extracted Feature (Feature Selection)	Classifier	Strength	Remaining issues	Language
Pang et al. (2002)	BOW, Lexical - unigram, Syntactic - POS	NB, SVM, ME	Simple. Easy to implement.	Difficult to detect sentiment for thwarted expression (across, contrary).	English
Dave et al. (2003)	Lexical - n-gram, Syntactic - POS, Semantic - Wordnet (IG),	NB, ME, SVM	Various feature to represent.	Rating inconsistency in labelled data.	English
Gamon (2004)	Lexical - n-gram, Syntactic - POS, Semantic - NLPWin	SVM	NLP feature improved performance.	Need large size dataset.	English
Salvetti, Lewis, & Reichenbach (2004)	Lexical - n-gram, Syntactic - POS word/tag, Semantic - WordNet	NB, HMM	POS (word/tag) adjective improved performance.	WordNet feature generalization cause performance drop.	English
Read (2005)	Lexical - n-gram, Pragmatic - emoticon	SVM, NB	Construct large emoticon corpus.	Noisy data. Mixed sentiment (probably sarcasm).	English
Kennedy & Inkpen (2006)	Pragmatic - contextual valence shifter	SVM	Valence shifter contribute to performance.	Difficult to implement. Need linguistic knowledge.	English
Ng et al. (2006)	Lexical - n-gram, tag polarity, objective info, Syntactic - POS	SVM	High accuracy using n-gram for subjectivity detection.	Some feature ineffective for polarity/sentiment classification.	English
Riloff et al. (2006)	Lexical - n-gram, Syntactic - POS, Hierarchy (IG)	SVM	Detail pattern feature, support subjectivity classification.	Complex hierarchy based.	English
Ye et al. (2009)	Lexical - unigram, char n-gram (IG)	SVM, NB	Simple. Easy to implement.	Need large size dataset.	English
Xia & Zong (2010)	Syntactic - POS (MI, IG)	NB, SVM, Ensemble	POS (word-tag) produced discriminative feature. Feature selection improved performance.	Need ensemble & joint model.	English
Bakliwal et al. (2012)	Lexical - n-gram, Pragmatic - emoticon, hashtag, target	SVM, NB	Various feature for training.	Need contextual study. Misclassified probably sarcasm.	English
Lane et al. (2012)	Lexical - n-gram, Semantic - dependency (CHI)	SVM, k-NN, NB, DT	Works for imbalanced data.	Feature selection is ineffective.	English

Table 2.1 Continued: Summary of Previous Work Using Supervised Learning Approach

Author	Extracted Feature (Feature Selection)	Classifier	Strength	Remaining issues	Language
Hamdan, Béchet, & Bellot (2013)	Lexical - unigram, Syntactic - POS, Prosodic - Slang word, Semantic - WordNet, SentiWordNet	SVM, NB	Various feature for training (use SamEval dataset).	Need large size dataset.	English
Moraes et al. (2013)	BOW (IG)	NB, SVM, ANN	Simple. Easy to implement. Works for balanced & imbalanced data.	Feature selection (IG) is ineffective with ANN.	English
Pham (2014)	Lexical - n-gram, rating based	SVM	Rating based contribute to performance.	Only work with rating based data.	English
Selvi et al. (2015)	BOW - unigram (IG, MI, CHI, GR, RF, OneR, DF)	ME, SVM, k-NN, NB, DT	Various feature selection & classifier.	Less feature. Depend on feature selection & classifier.	English
Tan & Zhang (2008)	BOW, Lexical - unigram (IG, MI, DF, CHI)	NB, SVM, k-NN, Winnow Classifier	Various feature selection techniques.	Need large size dataset.	Chinese
Ghorbel & Jacot (2011)	Lexical - n-gram, Syntactic - POS, Semantic - SentiWordNet	SVM	Can be applied to bilingual. Use translation.	Need thorough preprocessing. Misspelled affect translation, lemmatization.	French
Shi & Li (2011)	Lexical - unigram, frequency	SVM	Simple. Easy to implement.	Frequency of feature is ineffective.	Chinese
Shoukry & Rafea (2012)	Lexical - unigram, bigram (IG)	NB, SVM	Simple. Easy to implement. Works well for Arabic.	Filter only sentiment data. Discard other language, sarcasm.	Arabic
Balahur & Turchi (2013)	Lexical - unigram, bigram	Hybrid, SVM, SMO	Can be used to other language. Use translation.	Limited resources available for multilingual SA.	English, French, Italian, German, Spanish
Habernal et al. (2013)	Lexical - n-gram, char n-gram, Syntactic - POS, Pragmatic - Emoticon	SVM, MaxEnt	Various feature. Can be used for other study.	Dataset only applicable to Czech SA.	Czech
Kim et al. (2015)	Lexical - n-gram, char n-gram, Syntactic - POS	ANN, SVM, NB, DT	Various classifier used.	English data removed. Use specific Korean NLP.	Korean
Wang et al. (2015)	Lexical (MI)	NB	Simple. 2-step SA strategy can be implement.	Use specific Chinese word segmentation tool.	Chinese

2.2.3 Sentiment Analysis for Non-English Language

The non-English SA approach is similar to English in general. The distinctions can be found in the preprocessing and feature extraction phases. The issue for non-English is less resource such as spellchecker, POS tagger or linguistic resource (semantic dictionary like SentiWordnet) to perform the preprocessing and feature extraction. This is due to availability of resources that have been done were largely in English (Dashtipour et al., 2016; Korayem et al., 2016). A few work on Korean and Chinese had come out with their own language resource such as POS tagger and spellchecker to cope with the limited resource's issue (Kim, Do Young Kwon, & Jeong, 2015; X. Wang, Zhang, & Wu, 2015). Other non-English studies used available resources in English and adapted them in languages such as French, Arabic, Czech, Urdu and Malay. Ghorbel and Jacot (2011) study SA for French movie review. They translate the word from French to English, then extract the syntactic and semantic features. The translation method was also used by Balahur and Turchi (2013) to extract features using various languages. They emphasized the importance of spellchecking for improvement of feature extraction, as misspelled words will affect translation and lemmatization, thus misclassifying the results.

Like English, lexical features are most common for extraction of features and some feature selection were also applied. X. Wang et al. (2015) investigated 2-step strategy for sentiment classification of Chinese dataset, including subjectivity classification and sentiment classification on subjectivity. Better classification performance was produced when considering the 2-step strategy.

The classification algorithm used for learning is also not much different from English SA, which includes supervised classification algorithms such as NB, k-NN, SVM, DT and ANN. Among these, SVM is the most popular and produces the best sentiment classification performance.

From Table 2.1 the most common feature used for non-English language dataset were lexical and syntactic, same as described for non-English language. SVM was the most chosen classifier to perform classification. The most remaining issue rose in SA for non-English language were limited resource to perform preprocessing and feature selection.

2.2.4 Sentiment Analysis for Malay Language

A number of works on SA for Malay datasets can be found in the literature, such as (Samsudin et al., 2011; Samsudin et al., 2013a; Samsudin, Puteh, Hamdan, & Nazri, 2013b). The framework employed is similar to those presented in the foregoing sub-section, where the sentiment of the data was classified as either positive, negative, or neutral (Isa, Puteh, and Kamarudin (2013). Works of (Al-Moslmi, Gaber, Al-Shabi, Albared, & Omar, 2015; Alsaffar & Omar, 2014; Samsudin et al., 2013a) highlighted the importance of feature selection with respect to Malay SA.

Most of the work on Malay SA has employed supervised learning approach. With respect to the size of the datasets, most work used a total of 2000 opinions which consist of 1000 positive and 1000 negative opinions (Al-Moslmi et al., 2015; Alsaffar & Omar, 2014; Isa et al., 2013; Samsudin et al., 2011, 2013a).

For preprocessing, normalization, spelling correction, tokenization and stopwords removal were applied to clean the data from noisiness. Some works also consider more than one language which includes both Malay and English to remove stopword. Translation of Malay language to English was performed to permit the utilization of English stemmer and lemmatization dictionaries (Samsudin et al., 2011), since none is available for Malay. A challenge for social media with public comments and reviews is that users tend to use dual languages (bilingual), in addition to noisy and unstructured text. Samsudin et al. (2013b) proposed an approach for reducing noisy text by means normalization in the preprocessing stage. The proposed normalization consists of spellchecking, vocabulary comparison to the corpus and dictionary checking. Informal word that did not passed the normalization was removed.

A number of different feature selection techniques have been employed such as IG, CHI, Gini Index (GI), Principal Component Analysis (PCA), Feature selection Immune Network System (FS-INS) and Features Artificial Immune System (AIS) (Al-Moslmi et al., 2015; Alsaffar & Omar, 2014; Samsudin et al., 2013a). The reduction of features using these feature selection techniques managed to improve the classification performance of the SA system. With respect to classification algorithms, SVM, Sequential

Minimal Optimization (SMO), k-NN, and NB have been used in the literature (Al-Moslmi et al., 2015; Alsaffar & Omar, 2014; Samsudin et al., 2013a). Table 2.2 shows a summary of previous SA work for Malay data using a supervised classification algorithm with brief details of each work.

Table 2.2: Summary of Previous Work on Malay Sentiment Analysis Using Supervised Learning Approach

Author	Approach	Extracted Feature (Feature Selection)	Classifier	Strength	Remaining issue
Samsudin et al. (2011)	Annotation dataset, tokenization, normalization micro text, spellchecking, translation to English, stemming, lemmatization, vectorization.	Syntactic - POS	SVM, NB, k-NN.	Preprocessing can be improve & applied to bilingual.	Less NLP feature extracted. Translate all data to English.
Samsudin et al. (2013a)	Acquisition dataset, normalization, tokenization, spellchecking, translation manually, stopwords removal English & Malay.	Lexical - n-gram, negation, (Feature selection Immune Network System (FS-INS))	SVM, NB, k-NN	Preprocessing can be improve & applied to bilingual.	Ignore low frequency feature using FS-INS.
Samsudin et al. (2013b)	Standardization, normalization Malay Mixed Text Normalization (MyTNA), translation, stopwords removal English & Malay, tokenization.	Lexical - n-gram, negation, (Immune Network System (FS-INS))	SMO	Preprocessing with standardization & normalization (MyTNA) improves performance.	Feature selection alone was inefficient
Isa et al. (2013)	Acquisition, standardization, normalization, tokenization, stopwords removal (modified), stemming (Reverse Porter Algorithm based & Backward-Forward Algorithm).	Lexical - unigram, frequency, polarity, (Features Artificial Immune System (AIS))	k-NN	Stemming of Malay text using Reverse Porter Algorithm.	Difficult to implement. Need linguistic knowledge.
Alsaffar & Omar (2014)	Tokenization, stopwords removal, spell correction,	BOW, Lexical - ngram (IG, CHI, GI, PCA, FS-INS & AIS)	SVM, NB, k-NN	Easy to implement. Various feature selection.	Different feature size works for different feature selection & algorithm
Al-Moslmi et al. (2015)	Standardization, tokenization, stop words removal.	BOW, Lexical - ngram (IG, CHI, GI)	SVM, NB, k-NN	Easy to implement. Various feature selection.	Different feature size works for different feature selection & algorithm

2.3 Framework of Sarcasm Detection and Classification

This section presents a literature review of framework for sarcasm detection and classification in SA using supervised learning approach. The definition of sarcasm is presented in Sub-section 2.3.1. An overview of the frameworks been proposed for the SA is presented in Sub-section 2.3.2. Details of the sarcasm detection and classification using supervised learning approach are reviewed in Sub-section 2.2.3. Sarcasm detection and classification using supervised learning approach for Non-English language is then reviewed in Sub-section 2.2.4.

2.3.1 Sarcasm Definition

Sarcasm or verbal irony can be simplified as the use of linguistic and stylistic process in expressing meaning differently from literal, which usually opposite and slightly different from the actual meaning depending on the viewpoint of the listener (Carvalho, Sarmento, Silva, & de Oliveira, 2009; R. Gibbs & Colston, 2007). It is a form of insincere communication that contradict to lying. In lying, speakers tend to hide the insincerity from the listener. On the other hand, sarcasm speakers intend to highlight the insincerity to convey criticism or covering up embarrassment. Usually it was done in dramatic or humorous manner but in polite and less aggressive than direct confrontation (Shany-Ur et al., 2012). Raymond W. Gibbs (2000) found sarcasm as a frequently used type of irony along with hyperbole, jocularity, rhetorical questions, and understatements to convey varied kinds of obvious and subtle interpersonal meanings. Using sarcasm typically expresses a negative attitude is directed towards an individual or a group. The exclamation "You're really brilliant!" about someone who has committed a thoughtless act is an example of sarcasm (Roger J. Kreuz & Roberts, 1993). In addition, sarcasm considered an explicitly aggressive type of irony, with clearer target and markers/cues (Attardo, 2000).

The concept of sarcasm and irony is segregated by a thin line (de Freitas, Vanin, Hogetop, Bochernitsan, & Vieira, 2014; López & Ruiz, 2016) and it is difficult to come up with a formal definition among researchers since sarcasm and irony are not static (Filatova, 2012; Weitzel et al., 2016). Irony can be simplified as saying or writing in an

opposite way of what is meant. Carvalho et al. (2009) assumed that the term irony refers to the specific case where a word or expression with prior positive polarity is figuratively used for expressing a negative opinion. In the work of sarcasm detection some SA researchers still depend on users in social media in defining the terms of sarcasm or irony, since these differences are limited to professionals and people involved in the field. Social media users do not clearly distinguish between the terms sarcasm or irony, or other figurative language (Davidov, Tsur, & Rappoport, 2010; Tsonkov & Koychev, 2015).

2.3.2 Overview of Sarcasm Framework

A concept of SA using sarcasm detection was proposed by Cambria et al. (2015). In their approach, sarcasm detection has to be performed first on the texts before polarity flipping takes place to identify the actual sentiment. To detect sarcastic opinions and polarity flipping, the authors suggest work by Polanyi and Zaenen (2006) that applied contextual valence shifter using positive and negative valence. Valence calculation is performed to the sentence using positive and negative valence corpus to flip the polarity. In addition, the authors also suggest the work of Davidov et al. (2010) using pattern-based with punctuation-based features, and work by González-Ibáñez, Muresan, and Wacholder (2011) using lexical and pragmatic features to characterize sarcastic word. However, no experiment has been conducted to evaluate the proposed framework. Another work proposed a decision making framework, based on users' opinion, to recognize sarcasm sentiment using a parsing based algorithm in unsupervised approach (Bharti, Babu, & Jena, 2015). The framework employed two approaches to identify sarcastic tweets; the parsing-based lexicon generation and interjection word enrichment. The framework consists of social media, social media entities, polarity detector and sentiment classification. Social media refers to website or networking platform while social media entities consist of entities for user to review, express opinion, and retrieve comment. A polarity detector is an automated system to identify the actual sentiment or sarcasm from the text. From the tweets in social media entity, the detector classified the sentiment into negative, positive or neutral. Further checks were applied

to positive or negative sentiments to see whether they had actual sentiment or sarcastic sentiment. The propose algorithm were tested on two categories of tweets, with sarcasm hashtag and without sarcasm hashtag. The final sentiments were either actual positive, actual negative or sarcastic.

Recently a case study of sarcasm detection and classification in social media by Muresan et al. (2015) proposed a computational framework that includes comments acquisition, post-processing, corpus creation, features extraction and selection, and classification. The comments were acquired from Twitter, and annotation and indexing were based on the hashtag used by users. The post-processing filters and eliminates the meaningless tweets. Only the verified tweets were considered for corpus creation. A final corpus consists of 900 tweets in three categories, sarcastic, positive and negative was created. Lexical and pragmatic features were extracted. Lexical features used were n-gram in the form of unigram and bigram. Semantic features included Linguistic Inquiry and Word Count[3] (LIWC) (Pennebaker, Boyd, Jordan, & Blackburn, 2015) and Wordnet-Affect[4] (Strapparava & Valitutti, 2004). Emoticon, punctuation and common ground (used of user name to reply or "retweet", and emoticon characteristic) were used as pragmatic features. CHI was used to identify useful features. NB, Logistic Regression (LogR) and SVM were employed as classifiers. Binary and three-ways classification experiments were conducted using 5-CV. The best results were produced using combination of unigram, semantic and pragmatic features. The output of the sentiment classification was either positive, negative, or sarcasm.

The following two sub-sections present some discussion of previous work on sarcasm in SA using supervised learning approach (Sub-section 2.3.1) and for non-English language (Sub-section 2.3.2).

[3] https://liwc.wpengine.com/
[4] http://wndomains.fbk.eu/wnaffect.html

2.3.3 Sarcasm Detection and Classification Using Supervised Learning Approach for English Language

This sub-section reviewed the sarcasm detection and classification approach using supervised machine learning for English language from the literature. First the dataset acquisition and annotation are presented in Sub-subsection 2.3.3.1. Later, data preprocessing in is presented in Sub-subsection 2.3.3.2. Feature extraction and feature selection are presented in Sub-subsection 2.3.3. Finally, classifier generation and classification are presented in Sub-subsection 2.3.3.5.

2.3.3.1 Dataset Acquisition and Annotation

Sarcasm in SA warrants serious interest in domains where sarcastic contents are high such as business, economic and politic (Farzindar & Inkpen, 2015; B. Liu, 2015). Some work have been conducted in education (Altrabsheh, Cocea, & Fallahkhair, 2015) and social (Wallace, Choe, & Charniak, 2015; Wallace, Choe, Kertz, & Charniak, 2014) domains. Most of these works used data acquired from Twitter and Facebook. There has been a few studies directed at product review dataset acquired from web such as Amazon (Buschmeier, Cimiano, & Klinger, 2014; Ramteke, Malu, Bhattacharyya, & Nath, 2013; Reyes & Rosso, 2012).

Level of analysis have at the document level, sentence and aspect level for solving specific objective. For sarcasm detection on previous work, focus has been directed at detecting, identifying or recognizing of sarcasm from dataset such as producing final class of sarcastic or non-sarcastic (sarcasm vs. non-sarcasm). While sarcasm classification aims to categorized, segregated or classified the sarcasm into specific class such as sarcastic and positive, sarcastic and negative, as well as three class of sarcastic, positive and negative (González-Ibáñez et al., 2011; Muresan et al., 2015; Nagwanshi & Veni Madhavan, 2014). Although both focuses are related, there has been little work done on sarcasm classification. In this thesis, previous works on sarcasm detection or sarcasm classification can be categorized into two tasks. The first task is the majority type of work that focuses on automatic detection of sarcasm or sarcasm classification. The second task is relational study related to sarcasm in SA such

as corpus generation for sarcasm (Maria Alcaide, Justo, & Ines Torres, 2015; Muresan et al., 2015; Sulis, Irazú Hernández Farías, Rosso, Patti, & Ruffo, 2016), recognition of sarcasm features (Tepperman, Traum, & Narayanan, 2006; Tungthamthiti, Shirai, & Mohd, 2014), and influence of sarcasm (Weitzel et al., 2016). The contribution of both tasks includes sarcasm detection or classification in SA. Some work has involved two or more task, such as recognizing of sarcasm features from corpus and later sarcasm detection work by (Bouazizi & Ohtsuki, 2015; González-Ibáñez et al., 2011).

Dataset annotation into sarcasm or non-sarcasm has been conducted, in most cases, using rule-based. Acquiring data using hashtags such as #sarcasm or #sarcastic will save time compared to manual or crowdsourcing annotation. Nagwanshi and Veni Madhavan (2014) and Z. Wang, Wu, Wang, and Ren (2015) used hashtags to identify positive and negative polarity along with sarcasm. Hashtag for positive polarity includes #happy, #joy, and #lucky while negative polarity includes #sadness, #angry and #frustrated. The data usually acquired using Twitter API[5]. However the use of hashtag for annotation is still debatable as the hashtag itself is not used formally in Twitter and depends on the person who tags the sentence (Davidov et al., 2010; Justo, Corcoran, Lukin, Walker, & Torres, 2014). To address this issue, manually reviewing the sentence or the comment to validate the labelling is required (González-Ibáñez et al., 2011; Liebrecht, Kunneman, & van den Bosch, 2013).

Manual annotation assigned human annotators to label the acquired dataset. One, two or more annotators were assigned to perform the task. Two or more annotators were selected to perform majority annotation, thus avoiding bias and misconceptions. Often the odd number of annotator is chosen when there is two class of annotation, such as sarcastic and non-sarcastic. Kappa statistic metric can be used to measure the inter annotator agreement. The annotation process in disambiguated tasks such as in linguistic often requires more than an annotator, thus the agreement of among them are measured for the reliability of the produced annotation. This is to ensure the reliability of the outcomes of the annotation process among the annotator using the scale of agreement (Fleiss, 1971; Landis & Koch, 1977). Cohen's kappa is usually used if there are only two annotators (Kunneman, Liebrecht, van Mulken, & van

[5] https://dev.twitter.com/streaming/overview

den Bosch, 2015; Liebrecht et al., 2013; Riloff et al., 2013), while Fleiss's kappa is used if there are three or more annotators (González-Ibáñez et al., 2011; Muresan et al., 2015; Ptácek, Habernal, & Hong, 2014) .

Crowdsourcing is another annotation method to speed up the process. Crowdsourcing services such as Mechanical Turk[6] and CrowdFlower[7] are popular choices as shown in the previous work (Barbieri, Ronzano, & Saggion, 2015; Justo et al., 2014; Maria Alcaide et al., 2015; Xu, Santus, Laszlo, & Huang, 2015)..

2.3.3.2 Preprocessing

Preprocessing of text for sarcasm detection or classification is similar to text preprocessing in SA described in Sub-section 2.2.1 and Sub-subsection 2.2.2.2. Some differences can be found during filtration or noise removal and standardization or normalization. Filtration usually filters out irrelevant hashtag symbol (#); some hashtags are deemed significant with respect to sarcasm detection such as #sarcasm (Kunneman et al., 2015; Ptácek et al., 2014; Weitzel et al., 2016). With respect to texts or comments normalization, Xu et al. (2015) performed thorough normalization such as replacing heavy punctuation (punctuation used repeatedly) and word segmentation from social media's comment with simpler form. For example, heavy punctuations like "?!?!!" are replaced with "?!" and segmented "yeahright" with "yeah right". Normalization produces meaningful features and avoids 'dispersion'. The dispersion is where features that should be considered to be the same feature are treated as different features, result in poor performance when creating training data with which to build a classifier (Forman, 2007).

2.3.3.3 Feature Extraction and Selection

Most work in the literature extracts lexical features from their dataset. Others have investigated various syntactic, pragmatic and prosodic features. Syntactic features includes tagged POS of noun, verb, adjective and adverb using various taggers.

[6] https://www.mturk.com/mturk/
[7] https://www.crowdflower.com/

Pragmatic includes emoticon, emotion label, capital words, and punctuation mark while prosodic considers interjection, explicit and implicit incongruity. Due to the adverse effect of sarcasm on SA, identification of features that best represent sarcasm is pertinent. NLP based features have been shown to be able to produce meaningful feature to detect sarcasm. Many previous work that used supervised learning for SA employed lexicon-based approach to extract semantic features. The lexicon-based approach uses dictionary such as WordNet, Wordnet-Affect SentiWordNet[8], SenticNet[9], SentiStrength[10] and LIWC to identify these features. Some researchers have used available corpus coupled with the dictionaries to extract relevant features such as Multi-Perspective Question Answering (MPQA)[11], AFINN[12] or own corpus (González-Ibáñez et al., 2011; Justo et al., 2014; Muresan et al., 2015; Ptácek et al., 2014; Xu et al., 2015).

In addition to these five types of feature, another feature that worth investigation is idiosyncratic feature. Idiosyncrasy is a mode of behavior or way of thought peculiar to an individual[13]. In the linguistic study of metaphorical language, an idiosyncrasy is an isolated metaphor rarely used in common conversation and yet intended to bring meaning to the overall message. Idiosyncratic features have been used as non-systematic metaphorical expressions as part of 'message delivery' during conversation. Examples include "head of cabbage", "foot of the mountain", and "leg of a table" (Lakoff & Johnsen, 2003). The point of metaphor is that the actual meaning should not be interpreted as the literal meaning, as in the case of sarcasm. Idiosyncratic features may be considered to be an indicator of the presence of sarcasm.

Feature selection also conducted in some of the previous work of sarcasm detection. CHI is most often chosen for feature selection in sarcasm detection (Joshi, Sharma, & Bhattacharyya, 2015; Maria Alcaide et al., 2015) and sarcasm classification (Muresan et al., 2015). In addition, example of work Xu et al. (2015) attained good performance in Semantic Evaluation Workshop (SemEval) of sarcasm detection using

[8] http://sentiwordnet.isti.cnr.it/
[9] http://sentic.net/
[10] http://sentistrength.wlv.ac.uk/
[11] http://mpqa.cs.pitt.edu/
[12] http://neuro.imm.dtu.dk/wiki/AFINN
[13] https://en.oxforddictionaries.com/definition/idiosyncrasy

feature ranking by Pearson's correlation coefficient and ruling out less important features.

2.3.3.4 Classifier Generation and Classification

Various classification algorithms or classifiers have been used to detect and classify sarcasm. From the literature, the most common used are SVM, SMO, ME, NB and DT. In addition, Altrabsheh et al. (2015) maximized the variation of NB with experimented Complement Naïve Bayes (CNB) and Naïve Bayes Multinomial (NBM) as sarcasm detection was better in CNB than other classifiers.

Table 2.3 shows the summary of previous sarcasm detection and classification work on English and non-English data using supervised classification algorithm. The detail of the features extracted, detection or classification approach, strength and issues are also given. From Table 2.3, various feature in was extracted in type of lexical, syntactic, pragmatic, prosodic and semantic. The CHI was chosen as common feature selection and SVM was most common classifier. The advantage highlights the variation of proposed feature had improved the sarcasm detection and classification. The issue centered on the need for features to represent sarcasm, to better distinguish it from non-sarcastic. A review of previous sarcasm detection and classification work on non-English data is presented in the following subsection.

Table 2.3: Sarcasm Detection and Classification Using Supervised Learning Approach

Authors	Feature Details	Detection/ Classification	Classifier	Strength	Remaining issues	Language
González-Ibáñez et al. (2011)	1. Lexical factors: unigrams, dictionary-based. 2. Prosodic: interjections. 2. Pragmatic factors: punctuations, positive emoticons, negative emoticons, ToUser. 3. Feature ranking: presence & frequency.	Sarc vs. Pos vs. Neg, Sarc vs. Non-Sarc, Sarc vs. Neg, Sarc vs. Pos, Pos vs. Neg.	SMO, LogR.	Various feature for training.	Lexical & pragmatic ineffective in pos vs. neg classification	English
Lunando and Purwarianti (2013)	1. Lexical: Unigram (SentiWordNet score), Negation, Word context, Affix. 2. Negativity: negative sentiment based topic. 3. Number of interjection words. 4. Question word - Boolean value.	Pos vs. Neg vs. Neu, Sarc vs. Non-Sarc.	NB, ME, SVM.	Negativity & interjection number improved accuracy.	Sarcasm texts have no global topic affected accuracy.	Indonesian
Riloff et al. (2013)	1. Heuristic syntactic: POS tag bootstrapping. 2. Lexical: Unigram, unigram & bigram.	Sarc vs. Non-Sarc.	SVM, Lexicon-based.	Hybrid approach improves recall. Approach can be use as guidance for polarity flip.	Complex hybrid. Difficult to implement.	English
Liebrecht et al. (2013)	1. Features: unigram, bigram, trigram, frequency >3 times. 2. Weighted feature: CHI metric.	Sarc vs. Non-Sarc.	Balanced Winnow (Littlestone, 1988).	Simple. Easy to implement.	Hard to distinguish sarcastic tweets from literal tweets (no hashtag)	Dutch
Justo et al. (2014)	1. Mechanical Turk Cues: Lexical - n-gram. 2. Statistical Cues: unigram, bigram & trigram (feature selection CHI) 3. Linguistic information: Syntactic - POS. 4. Semantic information (LIWC dictionary) 5. Length information: number of words, characters, sentences, average words per sentence, average character per sentence) 6. Concept & Polarity Information (SenticNet-3.0)	Sarc vs. Non-Sarc, Nasty vs Non-Nasty.	NBM	Various feature for training.	Concept & polarity information do not improve result.	English

Table 2.3 Continued: Sarcasm Detection and Classification Using Supervised Learning Approach

Authors	Feature Details	Detection/ Classification	Classifier	Strength	Remaining issues	Language
Tungthamthiti et al. (2014)	1. Semantic: Concept level & common-sense - ConceptNet, Contradiction - SentiStrength & SenticNet, Coherence - POS. 2. Lexical: unigram, bigram, trigram 3. Sentiment feature: (low, medium & high) of pos/neg polarity, 4. Pragmatic: punctuation & special symbols feature - emoticon, slang, exclamation mark, capitalize word, repetitive.	Sarc vs. Non-Sarc.	Baseline: SVM trained with N-gram features.	Various feature for proposed. N-gram feature most effective.	Dependable to sarcasm hashtag.	English
Ptácek et al. (2014)	1. Lexical: Character n-gram, N-gram, Skip-bigram. 2. Pattern: Pattern high frequency words (HFWs), content words (CWs), Word-shape pattern 3. Syntactic: POS characteristics - number of nouns, verbs, & adjectives, ratio of nouns to adjectives & verbs to adverbs, number of negative verbs. 4. Pragmatic & Prosodic: punctuation, emoticon, pointedness, word-case.	Sarc vs. Non-Sarc.	ME, SVM	Various feature for training & experiment. Good for implement in other language.	Feature better in English only. Challenge in non-English language.	Czech
Nagwanshi and Veni Madhavan (2014)	1. Lexical: n-grams (up to bigram) occur > twice. 2. Syntactic: POS tag – Stanford POS tagger 3. Semantic: words which contradict or are nearly opposite of each other – WordNet. 4. Pragmatic: sentiment of the sentence differs from the emoticons or similes. 5. Politeness rating: words like "extremely" , "too" used before positive sentiment carrying word. 6. Flipping of sentiment: sentiment dissimilarity in the sentiment progression within a sentence - SentiWordNet.	Sarc vs. Pos, Sarc vs. Neg.	NB, SVM	Various feature combination experimented. Better result if more combination used.	Proposed algorithms differ to human evaluation dataset.	English
Maria Alcaide et al. (2015)	1. Statistical cues - Lexical unigram, bigram, trigram (feature selection CHI). 2. Semantic information - LIWC semantic. 3. Alternative combine: Weighted combination of features.	Sarc vs. Non-Sarc.	NBM, SVM.	Statistical n-gram features with selection show effectiveness.	Combination feature do not contribute to the best result.	English

Table 2.3 Continued: Sarcasm Detection and Classification Using Supervised Learning Approach

Authors	Feature Details	Detection/ Classification	Classifier	Strength	Remaining issues	Language
Altrabsheh et al. (2015)	1. Lexical: unigram. 2. Pragmatic: emotion label, polarity label, no. of punctuation characters, no. of question & exclamation marks, no. of emoticons, no. of hashtags, time of tweet.	Sarc vs. Non-Sarc.	NB, NBM, CNB, ME, SVM, SMO, RF.	Various classifier experimented. Simple feature used for training.	Feature without preprocess lowered the performance.	English
Joshi et al. (2015)	1. Lexical: unigram by feature selection CHI 2. Pragmatic: emoticons, laughter expressions, punctuation marks, capital words. 3. Prosodic: Explicit incongruity - numeric, qualitative features; overtly expressed through sentiment words of both polarities. 4. Prosodic: Implicit Incongruity - implicit phrases features; covertly expressed through phrases of implied sentiment, opposed opposing polar words.	Sarc vs. Non-Sarc.	SVM	Various feature for training.	Incongruity difficult to implement. Need linguistic knowledge.	English
Bouazizi and Ohtsuki (2015)	1. Sentiment-related Features: 14 features qualified as "sentiment-related" including number of positive & negative words, highly emotional positive & negative. 2. Punctuation-Related Features: number of exclamation marks, question marks, dots, capital words & quotes, words contain repeated vowels. 3. Syntactic Features: uncommon words, common sarcastic expressions, number of interjections & laughter. 4. Pattern-Related Features: POS tag based classification	Pos vs. neg, Sarc vs. Non-Sarc	A: NB, SVM, NE B: RF	Recall performance improves after consider sarcasm, concentrate on negative tweets only (sarcasm does not always mean that what is said is the opposite of what is meant). Approach can be used as guidance for polarity flip.	Concentrate to correct positive tweets only.	English
Kunneman et al. (2015)	1. Features: unigram, bigram, trigram, frequency >3 times. 2. Weighted feature: CHI metric.	Sarc vs. Non-Sarc.	Balanced Winnow (Littlestone, 1988).	Simple. Easy to implement. Extend to French.	More clues/feature needed to signify sarcasm.	Dutch, French
Muresan et al. (2015)	Bag-of-Features representation: 1. Lexical: n-gram, lexicon-based - LIWC & WordNet-Affect 2. Pragmatic: emoticons, common ground. 3. Feature Ranking 4. Combination Feature Feature selection: CHI	Sarc vs. Pos vs. Neg, Sarc vs. Non-Sarc, Sarc vs. Neg, Sarc vs. Pos.	NB, SVM, LogR	Adding more data with feature selection & ranking increase performance.	Lexical & pragmatic feature do not provide sufficient information to differentiate sarcasm from positive & negative.	English

Table 2.3 Continued: Sarcasm Detection and Classification Using Supervised Learning Approach

Authors	Feature Details	Detection/ Classification	Classifier	Strength	Remaining issues	Language
Xu et al. (2015)	1. Lexical: a. UniToken- unigrams put in a bag with tags describing emphasis types (duplicate vowel, capitalized, heavy punctuation, emoticon) b. BiToken- bigrams of the normalized tokens 2. Syntactic: DepTokenPair- "parent-child" pairs based on dependency structures 3. Semantic: a. PolarityWin- four sentiment dictionaries: Opinion Lexicon, AFINN, MPQA & SentiWordnet. b. PolarityDep- similar to PolarityWin, differs in negation checked in the dependency structure c. PolarShiftWin- designed for Irony, based on a 5-window, shift of polarity is present check. d. PolarShiftDep- similar to PolarShiftWin, differs in shift is checked in the dependency structure. Normaization & feature selection Pearson's r.	Sarc vs. Non-Sarc.	DT, SVM	Various feature for training. Feature selection Pearson's r contribute to better performance. Good result in SemEval 2015 task.	Bigram feature harms performance.	English
Z. Wang et al. (2015)	1.BOW 2. Word Cluster Type: 1. History-based Context 2. Conversation-based Context 3. Topic-based Context	Sarc vs. Pos vs. Neg	SVM multiclass, SVM sequential - Markov Model (MM)	History-based context feature improves performance.	Word cluster better that BOW. SVM sequential better than SVM multiclass.	English
Weitzel et al. (2016)	1. Machine Learning: Word2Vec (W2V), TFIDF vector representation, combination. 2. Lexicon Based: SentiWordNet, Happiness Index, SenticNet, SentiStrength, Patern.en, Vader.	Sarc vs. Non-Sarc.	1.Supervised Learning SVM, LogR. 2.Lexicon-Based	Simple supervised learning approach. Result better than lexicon-based.	Lexicon-based approaches easily misled by sarcastic tweets	English

2.3.3 Sarcasm Detection and Classification for Non-English Language

The amount of work conducted in sarcasm detection for non-English language is limited compared to that in the English language. Some work just started a few years back in Indonesian, Dutch, and Czech, as shown in Table 2.3. Early work on sarcasm detection in the context of the Indonesian language, using lexical features combined with syntactic and prosodic features on Twitter data, can be found in Lunando and Purwarianti (2013). The English SentiWordNet was translated into Indonesian using Google Translate[14]; only words exist in the translated SentiWordNet were considered. Lexical features in the form of unigrams were then extracted from the texts. The features extracted include negation, word context, affix, number of interjections and question words. The best-performing features for sarcasm detection (sarcasm vs. non-sarcasm) were combination of negation and interjection coupled with SVM classifier. In other work on sarcasm detection using a Dutch Twitter dataset, only lexical features were used (unigram, bigram and trigram) with a frequency count of more than three (Liebrecht et al., 2013). The collected dataset was based on the Twitter hashtag #sarcasme (sarcasm in Dutch), thus signaling the presence of sarcasm. The hashtag #sarcasme was used as sarcastic marker that alerted the reader to the fact that the text was sarcastic. The features were then weighted using the CHI metric to select meaningful features. The weighted rankings demonstrated that exclamation mark occurred most frequent in Tweets with the #sarcasme label. To perform the classification, Balanced Winnow classifier was employed in a balanced and imbalanced distribution. The threshold was set for each ranking of features performed the learning for each threshold. Ptácek et al. (2014) used lexical and syntactic features to detect sarcasm in Czech Twitter data. The lexical features used include various n-gram and frequency patterns while the syntactic features used were POS tags, which are nouns, verbs, and adjectives, as well as the ratios of nouns to adjectives and adverbs. Lexical features with frequency patterns were found to produce the best detection result. An F score recorded using a SVM classifier outperformed the ME classifier.

To the best of the author's knowledge, there is no work has been conducted on sarcasm detection in Malay SA, whether using machine learning nor lexicon-based

[14] https://translate.google.com/

approach. Malaysian definition from formal '*Kamus Dewan'* for sarcastic is "*bersifat menyindir (dgn kata yg tajam, pedas) yg boleh menyakitkan hati seseorang*" (Baharom, 2007) which means "satirical (with sharp, penetrating word) that can hurt someone". Interesting point in Malaysian's culture studied by Hei (2009) proved that politeness of Malaysian people used sarcastic conversation to show refusal instead of saying "No". The author observed scenarios in different locations, events and participants. Likewise Mansor, Ahmad, and Yaakub (2010) studied the language politeness among higher education students and found use of sarcastic words as indirect refusal or disagreement. Malay student used metaphor sarcastic regularly compared to non-Malay student. Therefore, detection of sarcasm was deemed necessary in order to achieve better SA performance on a Malay dataset.

2.4 Support Vector Machines

SVM is a classification algorithm that has solid theoretical foundation and performs classification accurately in application involving high dimensional data (Gao & Sun, 2010; B. Liu, 2011; Manning et al., 2008). The work on SA and sarcasm detection and classification presented in the foregoing sections have employed SVM as classifier, most of which produced better results than other classifiers (Bouazizi & Ohtsuki, 2015; Joshi et al., 2015; Kennedy & Inkpen, 2006; Maria Alcaide et al., 2015; Muresan et al., 2015; O'Keefe & Koprinska, 2009; Pang et al., 2002; Ptácek et al., 2014; Read, 2005; Ye et al., 2009).

SVM classifiers perform classification by finding the optimal hyperplane that maximizes the margin between two classes, let say +1 (positive comment) and -1 (negative comment). Note that hyperplane refers to features in higher dimensional space. While features in 3-dimensions are called a plane or surface and in 2-dimensions are called a line (B. Liu, 2011).

For training, D can be represented as $\{(x_1, x_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $x_i = (x_{i1}, x_{i2}, ..., x_{ir})$ is a r-dimensional input vector in a real-valued space $X \subseteq R^r$ with y_i is its class label (output value) and $y_i \in \{1, -1\}$. In linear separable training, separating

a hyperplane in terms of an intercept term b and a normal vector $\vec{w}$ which is perpendicular to the hyperplane is given as:

$$f(x) = sign(\vec{w} \cdot \vec{x} + b) \tag{2.2}$$

In practice, training data (comment) is non-separable linearly and contains noise and errors such as outliers and mislabeling (B. Liu, 2011; Manning et al., 2008). To cope with the case, SVM must consider the noise and error in the training data. Thus, slack variables $\xi_i \ (\geq 0)$ are introduced as follows:

$$\begin{aligned} &\vec{w} \cdot \vec{x}_i + b \geq 1 - \xi_i, \quad \text{for } y_i = 1 \\ &\vec{w} \cdot \vec{x}_i + b \geq -1 + \xi_i, \quad \text{for } y_i = -1 \end{aligned} \tag{2.3}$$

The new constraint is:

$$\begin{aligned} &\text{Subject to: } y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 - \xi_i, \quad i = 1, 2, ..., n, \\ &\xi_i \geq 0, \ i = 1, 2, ..., n. \end{aligned} \tag{2.4}$$

Figure 2.2 shows the geometric interpretation for allowing errors in training data x_a and x_b in the wrong region with a slack variable.

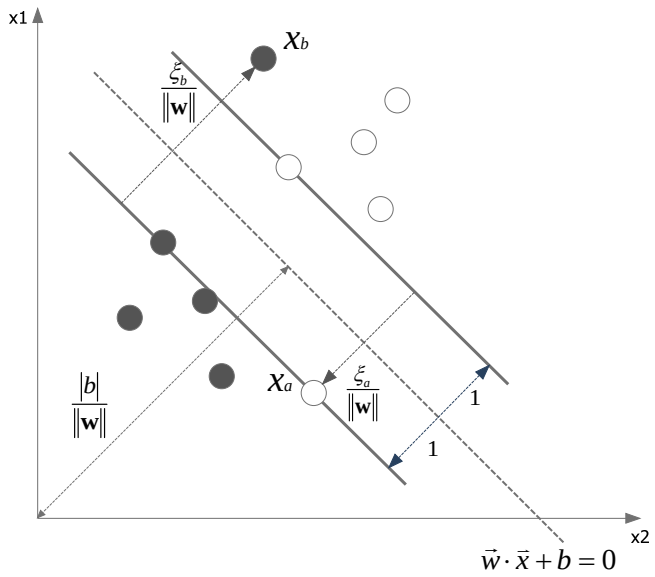


Figure 2.2: The Linear Non-Separable Case Allowing Data Points Error

Source : B. Liu (2011)

An extra cost for errors to change the function to penalize the errors in non-separable case is given by:

$$\frac{\|w\|^2}{2}+C\sum_i \xi_i$$
$$\text{Subject to: } y_i(\vec{w}\cdot\vec{x}_i+b)\geq 1-\xi_i, \quad i=\ 1,\ 2,\ ...,\ n, \tag{2.5}$$
$$\xi_i\geq 0,\ i=\ 1,\ 2,\ ...,\ n.$$

where $C\geq 0$ is a user defined parameter for controlling the extent of misclassified data points in training. Its dual is:

$$\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}y_i y_j\alpha_i\alpha_j\vec{x}_i\vec{x}_j$$
$$\text{Subject to: } \sum_{i=1}^{n}y_i\alpha_i=0, \tag{2.6}$$
$$0\leq\alpha_i\leq C,\ i=\ 1,\ 2,\ ...,\ n.$$

where $\alpha_i \geq 0$ are Lagrange multipliers.

In a real-life scenario, decision boundaries are non-linear. To deal with this SVM will use the same formulation and techniques in the linear case to transform their input data (input space) to higher dimensional space (feature space). Figure 2.3 shows the transformation's illustration from input space into feature space. Non-linear mapping ϕ is introduced as follows:

$$x \mapsto \phi(x) \tag{2.7}$$

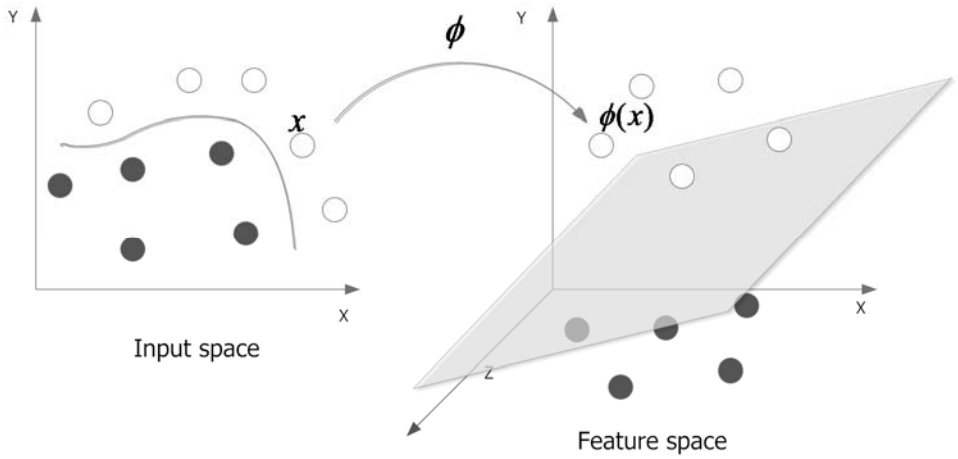


Figure 2.3: Non-Linear SVM Transformation from Input Space into Feature Space

Source : Adapted from B. Liu (2011)

With the transformation, the optimization problem in corresponding for dual becomes:

$$\begin{aligned} &\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \phi(\vec{x}_i)\phi(\vec{x}_j) \\ &\text{Subject to: } \sum_{i=1}^{n} y_i \alpha_i = 0 \\ &0 \leq \alpha_i \leq C,\ i = 1,\ 2,\ ...,\ n. \end{aligned} \tag{2.8}$$

The classification for training data in the final decision rule is then formulated by:

$$\sum_{i=1}^{n} y_i \alpha_i \phi(\vec{x}_i) \cdot \phi(\vec{x}) + b \tag{2.9}$$

To compute the dot product in the feature space let say $\phi(x) \cdot \phi(z)$ the kernel function, K is used without need to know the feature vector $\phi(x)$ or even the mapping function ϕ. This also known as kernel trick. Commonly used kernels are Polynomial (Poly) and Gaussian Radial Basis Function (RBF) (B. Liu, 2011; Manning et al., 2008):

$$\text{Poly: } K(x,z) = (\gamma \vec{x} \cdot \vec{z} + r)^d, \ \gamma > 0. \tag{2.10}$$

$$\text{Gaussian RBF: } K(x,z) = e^{-\gamma \|x-z\|^2}, \ \gamma > 0 \tag{2.11}$$

where $\lambda, r \text{ and } d$ are kernel parameters.

2.5 Summary

This chapter presented the background knowledge of the work focused in this thesis. The chapter started with a general overview of SA framework and approach. Next, various approaches such as preprocessing, feature extraction and selection can be used in ways that sentiment classification techniques may be applied using chosen classification algorithms. The review considers English, non-English and Malay languages. Later, a general overview of sarcasm detection framework and approach was presented. Detail approaches that can be used in ways that sarcasm detection and classification techniques may be applied was described. The review also had considered English and non-English language. Finally, selected classification algorithm to conduct works in this thesis was presented.

Based on the literature review presented in this chapter, it can be concluded that most of the work on SA or sarcasm detection and classification used similar approach

that consists of preprocessing, feature extraction, feature selection and classification. Two notable research opportunities are found: (i) the identification of features that best represent sarcastic content in bilingual data and (ii) the consideration of sarcasm detection and classification for SA. To address the first issue, NLP based feature could be useful to represent sarcasm content. A feature extraction process that can extract NLP based features from bilingual data need to be defined. To address the second identified gap, an SA system that incorporates sarcasm detection and classification is required. The SA framework by Medhat, Yousef, et al. (2014) and conceptual framework proposed by Cambria et al. (2015) could be used as a base for such system.

CHAPTER 3

THE DATASET AND PREPROCESSING

3.1 Introduction

This chapter describes the dataset used for evaluating the proposed approach and preprocessing of the dataset in preliminary phase. Details of the dataset used is described in Section 3.2. Acquisition of the data and the filtering process are explained in Section 3.3, followed by data annotation and distribution of the produced dataset in Section 3.4. Data preprocessing stage later is explained in Section 3.5. The tokenization, spellchecking and stopword removal processes are explained in detail. Finally, the chapter is summarized in Section 3.6.

3.2 The Dataset

The dataset in this research was extracted from Malay Facebook public page. Berita Harian and Astro Awani were two public pages that ranked as top public news pages for year 2015 (Chaffey, 2016). Topics in the economic and political domain were chosen. The consideration is taking due to the domain perspective, where sarcasm appears to be high in political and government topics such as economic (Farzindar & Inkpen, 2015; B. Liu, 2015). In line with this consideration, the dataset was acquired from the budget speech proposal of the Malaysian Prime Minister from 23rd of October 2015 till 12th November 2015.

3.3 Data Acquisition

Firstly, the threads or posts of topics were queried to the Berita Harian and Astro Awani Facebook public pages. Only threads posted in Malay language and related to economics were chosen. Threads posted in English or related to other topics were omitted. Then for each selected thread, comment that made by public towards the thread or post was scrapped and acquired. This acquisition was made using Graph Application Programming Interface (API) Explorer[15] and Facebook Query Language (FQL). Figure 3.1 shows a screenshot of the acquisition process. Topic and comments in Java Server Object Notation (JSON) format later transferred into table for filtration.

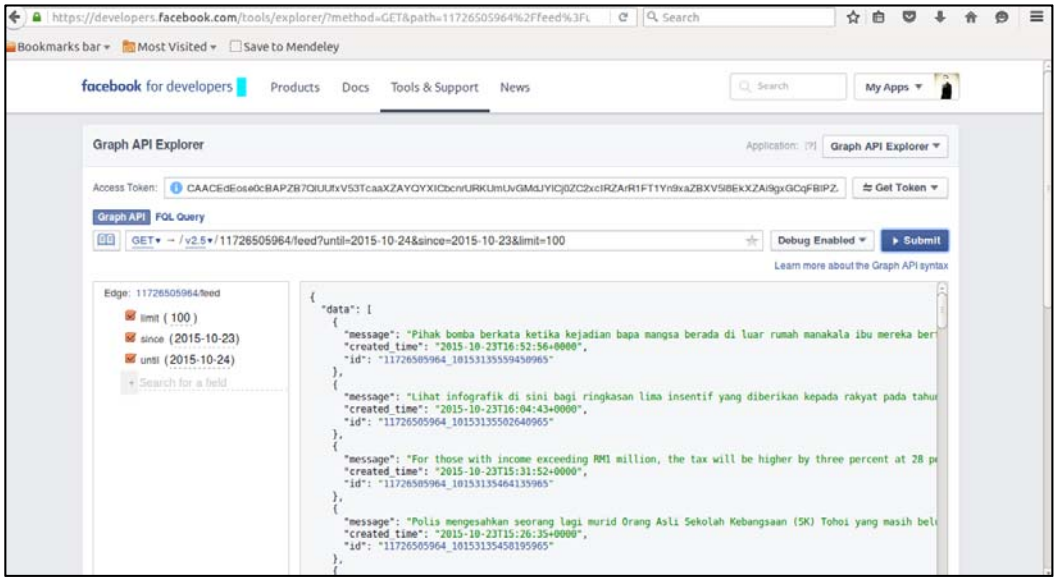


Figure 3.1: Screenshot of Facebook API

From the related comments, only text based comments were selected. Comments made in the form of picture, video, emoticon or Uniform Resource Locator (URL) was removed. Most text comment was made majority in Malay and some mixed with English. It was not uncommon to use dual language (bilingual) while commenting with used of short form and stylistic word. Misspelled and slang words can be found

[15] https://developers.facebook.com/tools/explorer

throughout the dataset. Figure 3.2 shows some comments acquired with said characteristics. At the end of data acquisition process, a total of 3000 comments were selected.

Characteristic	Comment
Malay	*"**Tahniah malaysia**"*
Bilingual	*"Terus maju tak kan korbankan rakyat. **Devils**"*
Short form	*"**Xmmpu** berkate2..**hny** mampu doa..supaya stabil..bru ok **nk** trade..hmm"*
Stylistic	*"**tsk tsk tsk**. x**lme** lg yuran **nek** la tu. good job."*
Slang	*"**dio** ni **gilo doh**"*
Misspelled	*"Dia ni syok sendiri ke...**tk** faham ke apa yg **raayat** inginkan..u got to go go go...that it."*

Figure 3.2: Comment Characteristic of Malay Social Media Dataset

3.4 Data Annotation

Once the data has been collected, manual data annotation is commenced. The annotation is conducted to obtain ground truth of the sentiments of each comment acquired. Three annotators were selected to perform the annotation. The selected annotators are native Malay speakers with English as their second language (the educational background of the annotators are varied from postgraduate to graduate level). The selection of these three annotators allow majority voting to be used to label the data; the third annotator provides judgment in case of tie decisions (Kunneman et al., 2015; Liebrecht et al., 2013; Wallace et al., 2015). Three types of annotations were produced by the annotators: (i) the sentiments of the comments (positive, negative or neutral), (ii) the existence of sarcasm on positive and negative comments only (sarcastic vs. non-sarcastic), and (iii) the positivity and negativity of the sarcasm.

The first annotation was produced to allow sentiment classification. For the work presented in this thesis, only positive and negative labels were considered. As a result, a subset of 1970 comments was derived from the original set of 3000 comments where 802 were positive and 1068 were negative labeled comments. The second annotation was performed to identify sarcastic and non-sarcastic comments. The annotation was conducted only on the selected 1970 comments from the first annotation. Sarcasm annotation of the 1970 comments produced 969 sarcastic comments and 1001 non-sarcastic comments. The final distributions of 'sentiment' and 'sarcasm' labeled annotations are shown in Table 3.1. Majority voting of the annotators was used to identify the label for both sentiment and sarcasm annotations.

Table 3.1: Dataset Distribution of Sentiment and Sarcasm Annotation

Category	i. Sentiment		ii. Sarcasm	
Label	Positive	Negative	Sarcastic	Non-sarcastic
Total	802	1168	969	1001

The third annotation is produced to identify sarcasm positivity and negativity of the selected comments. Comments that have been labeled as positive in the first annotation and sarcastic in the second were labeled as 'positive sarcastic', while positive comments that were labeled as non-sarcastic in the second annotation were labeled as 'true positive'. Similar annotation process was also applied for negative comments. Figure 3.3 shows the process of annotation for sarcasm positivity and negativity while the distribution of 'sarcasm positivity' and 'sarcasm negativity' of the comments is shown in Table 3.2. Comments labeled as positive sarcastic (usually known as false positives) have an actual negative sentiment and vice versa. Per Table 3.1, 802 are positive labeled comments. However, only 172 comments are actually positive while the remaining 630 are negative (as they feature sarcasm) as shown in Table 3.2. Similarly, for negative labeled comments, only 829 of the comments out of 1168 are actual negative.

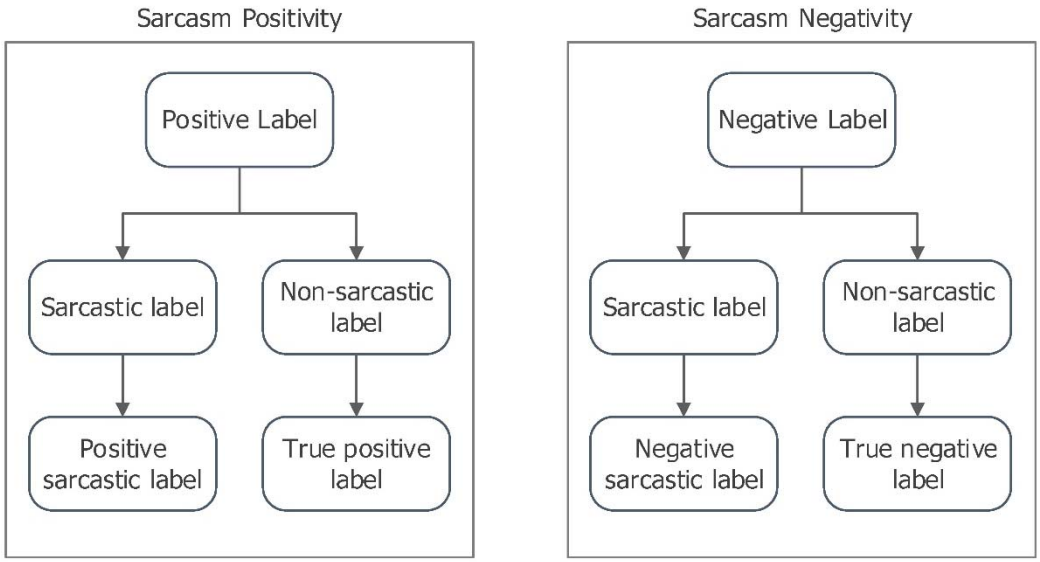


Figure 3.3: Annotation of Sarcasm Positivity and Sarcasm Negativity

The distribution of 'sarcasm positivity' and 'sarcasm negativity' of the comments are shown in Table 3.2. Comments labeled as positive sarcastic (usually known as false positives) have an actual negative sentiment and vice versa. Per Table 3.1, 802 are positive labeled comments. However, only 172 comments are actually positive while the remaining 630 are negative (as they feature sarcasm) as shown in Table 3.2. Similarly, for negative labeled comments, only 829 of the comments out of 1168 are actual negative.

Table 3.2: Dataset Distribution for Sarcasm in Positive and Negative Class (Positivity and Negativity)

iii. Sarcasm	Positivity	Positive sarcastic	630
		True positive	172
	Negativity	Negative sarcastic	339
		True negative	829

To ensure the level of agreement between annotators are acceptable, Fleiss's kappa inter annotator agreement score was used. The Fleiss's kappa (Fleiss, Nee, & Landis, 1979) formula for large sample calculation is given by:

$$s_k^2 = \frac{\sqrt{2}}{\sum_{j=1}^{q} p_j(1-p_j)\sqrt{nr(r-1)}} \sqrt{\left[\sum_{j=1}^{q} p_j(1-p_j)\right]^2 - \sum_{j=1}^{q} p_j(1-p_j)(1-2p_j)} \qquad (3.1)$$

where s_k^2 represents the variance of kappa conditional on the rater sample, p_j represents proportion of responses in category j, q represents total number of categories, n represents number of subjects or items, and r represents number of raters.

The calculated Fleiss's kappa for the annotated dataset was 0.71 for sentiment and 0.47 for sarcasm. Therefore the agreement between three annotator was substantial for sentiment annotation and moderate for sarcasm annotation (Landis & Koch, 1977). The moderate sarcasm score was acceptable (> 0.4) by considering the difficulty of sarcasm content for annotation task (Davidov et al., 2010; González-Ibáñez et al., 2011; Muresan et al., 2015; Ptácek et al., 2014).

3.5 Data Preprocessing

Annotated dataset was preprocessed to prepare for feature extraction and selection. Tokenization, spellchecking and stopword removal was applied to the annotated dataset in this preprocessing stage.

3.5.1 Tokenization

Each comment from the distribution underwent word tokenization. The process had split the sentence in the comment words. List of tokens produced for each comment. In this preliminary stage, all the character such as punctuation marks and hashtag (#) was

kept together with words. The tokenization process was performed using Python Natural Language Toolkit (NLTK) (Bird, Klein, & Loper, 2009) with data manipulation tool Pandas (McKinney, 2012). Figure 3.4 shows an example of word tokenization for the comments.

Comment		Tokenized
"Tahniah malaysia"	→	"Tahniah", "malaysia"
"itu naik ini naik letih oiii nk follow semuanya....#Terpaksa"	→	"itu", "naik", "ini", "naik", "letih", "oiii", "nk", "follow", "semuanya....#Terpaksa"
"Xmmpu berkate2..hny mampu doa..supaya stabil..bru ok nk trade..hmm"	→	"Xmmpu", "berkate2..hny", "mampu", "doa..supaya", "stabil..bru", "ok", "nk", "trade..hmm"
"tsk tsk tsk. xlme lg yuran nek la tu. good job."	→	"tsk", "tsk", "tsk.", "xlme", "lg", "yuran", "nek", "la", "tu.", "good", "job."
"dio ni gilo doh"	→	"dio", "ni", "gilo", "doh"
"Dia ni syok sendiri ke...tk faham ke apa yg raayat inginkan..u got to go go go...that it."	→	"Dia", "ni", "syok", "sendiri", "ke...tk", "faham", "ke", "apa", "yg", "raayat", "inginkan..u", "got", "to", "go", "go", "go...that", "it."

Figure 3.4: Word Tokenization

3.5.2 Spellchecking

Spellchecking was performed to avoid dispersion as discussed in literature review. The dataset was first lowercased for standardization before spellchecked. Malay and English dictionaries were used to spellcheck the tokenized bilingual data. In addition to overcoming misspelled words, short form, stylistic and slang forms not in a formal dictionary, a customized dictionary of regular used words was built to map all the word that cannot be corrected by spellchecker.

Phrases such as *"xlme lg"*, *"x lame lg"*, *"tak lama lg"* or *"tak lama lagi"* are examples of texts containing short form words, stylistic words and spelling errors; the correct standard phrase is *"tidak lama lagi"*. The presence of the above cause dispersion, where the same features are considered as different features. Such cases will results in poor performance when creating training data with which to build a classifier (Forman,

2007).This process returned a set of spellchecked and corrected words. For the work presented in this thesis, PyEnchant[16] spellchecker was employed. Figure 3.5 shows examples of spellchecked and corrected words.

Tokenized		Spellchecked
"Tahniah", "malaysia"	→	"tahniah", "malaysia"
"itu", "naik", "ini", "naik", "letih", "oiii", "nk", "follow", "semuanya....#Terpaksa"	→	"itu", "naik", "ini", "naik", "letih", "ooi", "nak", "follow", "semuanya", "....", "#", "terpaksa"
"Xmmpu", "berkate2..hny", "mampu", "doa..supaya", "stabil..bru", "ok", "nk", "trade..hmm"	→	"tidak", "mampu", "berkata", "..", "hanya", "mampu", "doa", "..", "supaya", "stabil", "..", "baru", "okay", "nak", "trade", "..", "hmmm"
"tsk", "tsk", "tsk.", "xlme", "lg", "yuran", "nek", "la", "tu.", "good", "job."	→	"tsk", "tsk", "tsk", ".", "tidak", "lama", "lagi", "yuran", "naik", "lah", "itu", ".", "good", "job", "."
"dio", "ni", "gilo", "doh"	→	"dia", "ini", "gila", "dah"
"Dia", "ni", "syok", "sendiri", "ke...tk", "faham", "ke", "apa", "yg", "raayat", "inginkan..u", "got", "to", "go", "go", "go...that", "it."	→	"dia", "ini", "syok", "sendiri", "ke", "…", "tidak", "faham", "ke", "apa", "yang", "rakyat", "inginkan", "..", "u", "got", "to", "go", "go", "go", "…", "that", "it", "."

Figure 3.5: Spellchecking of Tokenized Word

3.5.3 Stopword Removal

Malay and English stopword removal were used to remove regular words that are non-discriminative; it cannot differentiate comments of different labels of sentiments nor

[16] http://pythonhosted.org/pyenchant/

sarcasms. Malay regular words such as *"yang"*, *"ini"* and *"itu"*, and English regular words such as "a", "is" and "the" were removed from the dataset. Standard Malay[17] and English[18] stopword lists were used and modified for the work described in this thesis. The Malay stopword list can be found in Appendix A and the English stopword list in Appendix B. The process also removed unnecessary punctuation and left only spellchecked words. Figure 3.6 shows the example spellchecked words after stopword removal.

Spellchecked		Stopword Removal
"tahniah", "malaysia"	→	tahniah malaysia
"itu", "naik", "ini", "naik", "letih", "ooi", "nak", "follow", "semuanya", "....", "#", "terpaksa"	→	naik naik letih ooi nak follow semuanya terpaksa
"tidak", "mampu", "berkata", "..", "hanya", "mampu", "doa", "..", "supaya", "stabil", "..", "baru", "okay", "nak", "trade", "..", "hmmm"	→	tidak mampu berkata mampu doa stabil baru okay nak trade hmmm
"isk", "isk", "isk", ".", "tidak", "lama", "lagi", "yuran", "naik", "lah", "itu", ".", "good", "job", "."	→	isk isk isk tidak lama yuran naik lah good job
"dia", "ini", "gila", "dah"	→	dia gila dah
"dia", "ini", "syok", "sendiri", "ke", "…", "tidak", "faham", "ke", "apa", "yang", "rakyat", "inginkan", "..", "u", "got", "to", "go", "go", "go", "…", "that", "it", "."	→	dia syok sendiri tidak faham rakyat inginkan awak got go go go

Figure 3.6: Returned Word After Stopword Removal

[17] http://nlp.cs.nyu.edu/GMA_files/resources/malay.stoplist
[18] http://www.nltk.org/book/ch02.html

3.6 Summary

This chapter has described the data acquisition and preprocessing phase for the dataset used to evaluate the proposed approach. The data has been annotated manually by three human annotators. The distribution of the dataset has been described. Detail explanations of work conducted in this thesis at the preprocessing stage that consists of only tokenization, spellchecking and stopword removal were then provided. The output of the described processes is cleaned data where the next process can now be performed.

CHAPTER 4

FEATURE FOR SARCASM DETECTION ON BILINGUAL SOCIAL MEDIA DATA

4.1 Introduction

This chapter describes the process proposed to identify the best features for sarcasm detection in the context of bilingual social media data. The feature extraction process consists of two phases; (i) feature extraction from the original bilingual text and (ii) translate the original text to English and extract further features from the translated text. Feature selection is then applied to identify the most discriminative features. In order to evaluate the performance of the identified features to detect sarcasm, supervised classification approach is employed. The details of the proposed feature extraction process and selection are presented in Section 4.2. The experimental setup is described in Section 4.3. Section 4.4 presents the generated results and discussion of results. Section 4.5 concludes this chapter.

4.2 The Proposed Feature Extraction Process

This section describes the process conducted to extract features indicative of the presence of sarcasm based on NLP with respect to bilingual social media texts, in this work the focus is on Malay and English languages. NLP has shown the ability to identify the presence of sarcasm and resolve the challenge (Indurkhya & Damerau, 2010) by means of recognizing and extracting a specific feature to simplify the complex meaning of texts (Reyes, Rosso, & Buscaldi, 2012). The proposed process is comprised of two

phases. In the first phase, lexical features are extracted, followed by pragmatic and prosodic features, in the order that they appear in a given bilingual text. In the second phase the bilingual dataset is translated to English and further prosodic features are extracted along with syntactic and idiosyncratic feature. A lexical representation was used as the fundamental document representation because this is the simplest, and therefore the most common encountered. However, lexical tends to ignores grammar, sentence structure and usually punctuation, thus ignoring the linguistic structure (Provost & Fawcett, 2013). On its own, it is clearly not expressive enough to provide information regarding the presence of sarcasm, hence the overlay of various feature categories as proposed in this paper.

The following sub-section presents the features extracted using NLP that have been adopted for sarcasm detection in the context of bilingual data. In the work described in this thesis, both the original bilingual dataset and its translation in English are considered for feature extraction. The process of feature extraction consists of two main steps: (i) extraction of the lexical, pragmatic and Malay prosodic features from the original bilingual dataset (Sub-section 4.2.1); and (ii) translation of the bilingual dataset to English and extraction of English prosodic feature, along with syntactic and idiosyncratic feature (Sub-section 4.2.2). The process returns sets of extracted features as summarized in Table 4.1. Once the features are identified and extracted from the dataset, feature vectorization was performed using TF-IDF vectorization and normalized to document length.

Table 4.1: Types of Feature Extracted

Feature	Types	Dataset
1. Lexical	Unigram	Original bilingual
2. Pragmatic	Punctuation marks, Hashtag	Original bilingual
3. Prosodic	Interjection	Original bilingual and English
4. Syntactic	Part of Speech	English
5. Idiosyncratic	Idiosyncratic	English

Figure 4.1 shows the feature extraction process for lexical, pragmatic, prosodic, syntactic, and idiosyncratic, from bilingual (A) and translated dataset (B).

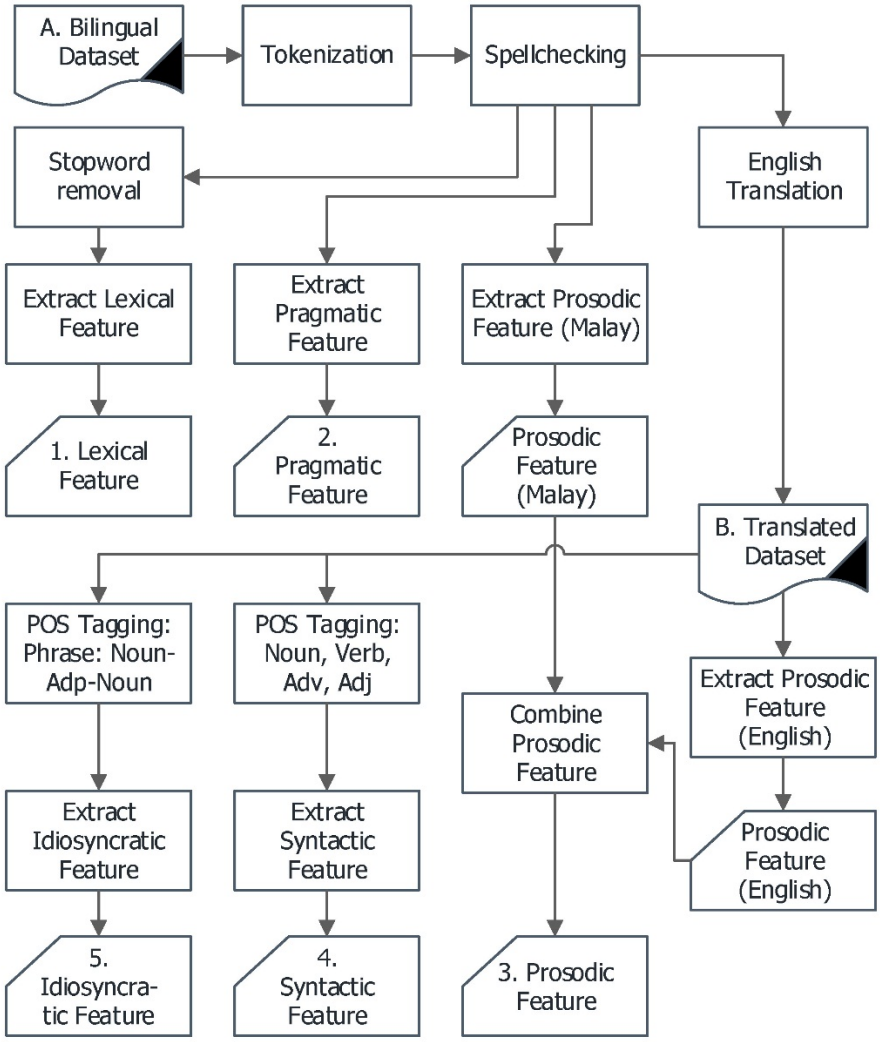


Figure 4.1: Feature Extraction Process

4.2.1 Extraction from Bilingual Dataset

This section describes the process of feature extraction from the original bilingual dataset. Prior to the extraction of feature, the dataset was preprocessed as described in Chapter 3.5. The processes for feature extraction from bilingual dataset involving three

steps as presented in Figure 4.1. Details of the processes are described in Sub-sections 4.2.1.1 to 4.2.1.3 below.

4.2.1.1 Lexical Feature Extraction

Lexical features were extracted from the preprocessed bilingual dataset. The lexical feature was represented in the form of n-grams. 1-gram (n=1) or unigram was selected to represent the lexical feature. When applied on the bilingual text, a total of 2730 number of lexical features was extracted. Table 4.2 shows the examples of the extracted lexical feature.

Table 4.2: Examples of Lexical Feature

Original Comment	Preprocessed	Lexical Feature
"itu naik ini naik letih ooiii nk follow semuanya....#Terpaksa"	itu naik ini naik letih ooi nak follow semuanya .... #terpaksa	naik naik letih ooi nak follow semuanya terpaksa
"langkah berimbang untuk keseksaan...kesengsaraan rakyat...kesejahteraan tang mana tuhhh!!!????"	langkah berimbang untuk keseksaan ...kesengsaraan rakyat … kesejahteraan tang mana tu !!! ????	langkah berimbang keseksaan kesengsaraan rakyat kesejahteraan

4.2.1.2 Pragmatic Feature Extraction

With respect to the work described in this thesis, punctuation marks were considered to be pragmatic features, instead of sentence segmentators, because of their ability to identify sarcasm (Carvalho et al., 2009). Heavy punctuation, for example high occurrences of various punctuation marks, is often an indicator of the presence of sarcasm in text. Examples heavy punctuation found in the dataset are *"…kesejahteraan tang mana tuhhh**!!!????**"* and *"MENYUSAHKAN RAKYAT**!!!** Janji tk ditepati**!!!**"*. The punctuation marks considered in this proposed work were question marks (?), exclamation marks (!) and quotation marks (‘” and ‘’). In addition, hashtag (#) was also

considered as it is another indicator of sarcasm (Bakliwal et al., 2012; Davidov et al., 2010; Kunneman et al., 2015; Ptácek et al., 2014; Read, 2005). The length of sequences of punctuation marks was reduced to a maximum of three characters to avoid dispersion (Forman, 2007; Liebrecht et al., 2013). At the end, four pragmatic features in total were extracted from tokenized and spellchecked dataset. Table 4.3 shows the examples of the extracted pragmatic feature.

Table 4.3: Examples of Pragmatic Feature

Original Comment	Preprocessed	Pragmatic Feature
"itu naik ini naik letih ooiii nk follow semuanya....#Terpaksa"	itu naik ini naik letih ooi nak follow semuanya .... #terpaksa	#
"langkah berimbang untuk keseksaan...kesengsaraan rakyat...kesejahteraan tang mana tuhhh!!!????"	langkah berimbang untuk keseksaan ...kesengsaraan rakyat … kesejahteraan tang mana tu !!! ????	!!! ???

4.2.1.3 Prosodic (Malay) Feature Extraction

Malay prosodic feature extraction was performed using Malay list of interjections[19]. It should be noted that interjections differ according to language; for example, *"ooi"*, *"puii"* and *"weii"* are only found in Malay. A total of 43 Malay interjections were identified and used. The Malay list of interjections can be found in Appendix C. In the end, 40 prosodic features were extracted from original bilingual dataset (tokenized and spellchecked). Table 4.4 shows the examples of the extracted Malay prosodic feature.

[19] https://ms.wikipedia.org/wiki/Kata_seru

Table 4.4: Examples of Prosodic Feature (Malay)

Original Comment	Preprocessed	Prosodic Feature (Malay)
"itu naik ini naik letih ooiii nk follow semuanya....#Terpaksa"	itu naik ini naik letih ooi nak follow semuanya .... #terpaksa	ooi
"woww igt byk ke 500 tu"	wow ingat banyak ke 500 tu	wow

4.2.2 Extraction of Features from English Translated Dataset

Once the aforementioned features were extracted from the original bilingual dataset, the remaining features were then extracted from the English translated dataset. The steps involved are described in Sub-sections 4.2.2.1 to 4.2.2.4.

4.2.2.1 Dataset Translation to English

The spellchecked and stopword removed bilingual dataset from Sub-section 3.5.3 was translated into English using Google Translate[20]. Although the resulting translations were by no means perfect, they were judged to produce translations that were sufficiently accurate to support further analysis, better than the translations using Moses or Bing (Balahur & Turchi, 2014). The employment of deep learning recently also had boost the performance of Google Translate (Castelvecchi, 2016). The translation preserved some Malay words such as names, locations and abbreviations such as 'idris', 'malaysia' and 'ptptn'.

[20] https://translate.google.com/

4.2.2.2 Prosodic (English) Feature Extraction and Combination

English prosodic feature extraction was performed using English list of interjections, obtained from publicly available source[21]. Interjections that are already listed in the Malay interjection as described in Sub-section 4.2.2.3 were removed to avoid redundant features. A total of 66 English interjections were identified as shown in Appendix D. In the end, a total of 26 prosodic features were extracted from the translated dataset. The extracted English prosodic features were then combined with the Malay prosodic feature identified from Sub-section 4.2.2.3. Table 4.5 shows the examples of the extracted English prosodic feature throughout the extraction process.

Table 4.5: Examples of Prosodic Feature (English)

Original Comment	Preprocessed	Prosodic Feature (English)
"Waa banyak maju ohh.... banyak maju..."	waa very advance oh very advance	waa
"yes.. yes..yes!!!"	yes yes yes	yes

4.2.2.3 Syntactic Feature Extraction

Four groups of POS were chosen for the work described in this thesis: NOUN, VERB, ADJECTIVE and ADVERB based on Japerson's Theory for ranking content in language and from the literature. The groups are common distinctive type in the form of POS in grammar, thus provide discriminative feature (Lakoff & Johnsen, 2003; Ptácek et al., 2014; Zhang, Xu, Su, & Xu, 2015). From the translated dataset, tokenization was applied. Each of the tokenized word was then POS tagged. Only the tokenized words associated with the four selected POS groups, as described above, were retained in the text; all other words were removed. The word-tag pair were used to represent the syntactic feature as it had been shown to produced better sentiment classification performance when used together (Xia & Zong, 2010). Table 4.6 shows the examples of

[21] http://grammar.yourdictionary.com/parts-of-speech/interjections/list-of-interjections.html

the extracted syntactic features. The total number of syntactic features extracted was 3695.

Table 4.6: Examples of Syntactic Feature

Original Comment	Preprocessed	Syntactic Feature
"Cantiknya karangan ini . Nak kata semut besar dari gajah la ni."	this beautiful bouquet to say ants than elephants this is	beautiful_ADJ bouquet_NOUN say_VERB ants_NOUN elephants_NOUN
"Ssah mau ckap mlas tngok ni orang, mcm mka jerung,"	hard to say lazy to see this people face of sharks	hard_ADJ say_VERB lazy_ADJ see_VERB people_NOUN face_NOUN sharks_NOUN

4.2.2.4 Idiosyncratic Features Extraction

To extract idiosyncratic feature from the dataset, a syntax rule was created in the form of NOUN-ADPOSITION-NOUN. The syntax was based on linguistic study. An example of idiosyncratic phrase is "head of cabbage", since "head", "of" and "cabbage" will be tagged as noun, adposition and noun by the POS tagger (which satisfies the syntax rule for idiosyncratic). To extract the features, POS tagging was first performed on the dataset. This is similar to the syntactic feature extraction but only the NOUN and ADPOSITION tags were considered to extract idiosyncratic features. The POS tagged dataset was then scanned for phrases that satisfy the above syntax rule. A total of 177 idiosyncratic features were identified from POS tagged translated dataset. Examples of idiosyncratic features found in the dataset were "puppet of buffalo", "face of sharks", "kinds of beans", "ants than elephants", "people in clown" and "joke between continents". Table 4.7 shows the examples of the extracted syntactic feature throughout the extraction process. The identified idiosyncratic features were then replaced in the text (for both sarcastic and non-sarcastic examples) with a unique identifier, idiosyncratic_x, where $1 \leq x \leq 177$.

Table 4.7: Examples of Idiosyncratic Feature

Original Comment	Preprocessed	Idiosyncratic Feature
"Cantiknya karangan ini . Nak kata semut besar dari gajah la ni."	this beautiful bouquet to say ants than elephants this is	"ants than elephants"
"Ssah mau ckap mlas tngok ni orang, mcm mka jerung,"	hard to say lazy to see this people face of sharks	"face of sharks"

4.2.3 Feature Selection

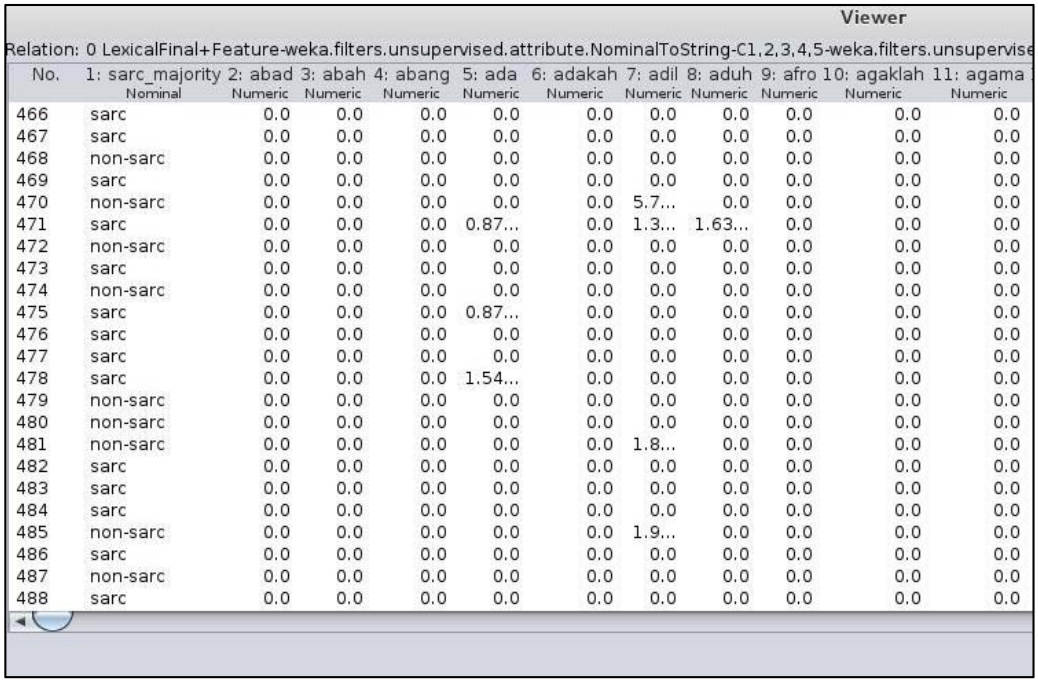
Once the features were extracted, feature selection was conducted. With respect to the work described in this thesis, Pearson's correlation coefficient, as described in Chapter 2, was used to rank features according to their correlated class (sarcasm or non-sarcasm). Then, the top N features were selected for classifier generation.

4.3 Experimental Setup

This section describes the objective of the experiment conducted and parameter setting for the evaluation of the proposed approach.

4.3.1 Experiment Objective

The objective of the evaluation was to identify the effectiveness and the best combination of the different categories of features, for the detection of sarcasm in the context of bilingual data used in this work. Figure 4.2 shows a screenshot of the vectorized and normalized features of some of the data used in this work.



Viewer

Relation: 0 LexicalFinal+Feature-weka.filters.unsupervised.attribute.NominalToString-C1,2,3,4,5-weka.filters.unsupervise

No.	1: sarc_majority Nominal	2: abad Numeric	3: abah Numeric	4: abang Numeric	5: ada Numeric	6: adakah Numeric	7: adil Numeric	8: aduh Numeric	9: afro Numeric	10: agaklah Numeric	11: agama Numeric
466	sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
467	sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
468	non-sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
469	sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
470	non-sarc	0.0	0.0	0.0	0.0	0.0	5.7...	0.0	0.0	0.0	0.0
471	sarc	0.0	0.0	0.0	0.87...	0.0	1.3...	1.63...	0.0	0.0	0.0
472	non-sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
473	sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
474	non-sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
475	sarc	0.0	0.0	0.0	0.87...	0.0	0.0	0.0	0.0	0.0	0.0
476	sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
477	sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
478	sarc	0.0	0.0	0.0	1.54...	0.0	0.0	0.0	0.0	0.0	0.0
479	non-sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
480	non-sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
481	non-sarc	0.0	0.0	0.0	0.0	0.0	1.8...	0.0	0.0	0.0	0.0
482	sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
483	sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
484	sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
485	non-sarc	0.0	0.0	0.0	0.0	0.0	1.9...	0.0	0.0	0.0	0.0
486	sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
487	non-sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
488	sarc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 4.2: Screenshot of TF-IDF Vectorization and Normalization to Document Length

There were five sets of experiment conducted as itemized in Table 4.9. Each set was designed to identify the performance of different individual or combination of feature categories. Set 1 refers to single feature category of experiment with lexical feature was selected as baseline for comparison purpose. Set 2, 3 and 4 combined two, three and four categories of feature respectively. Finally Set 5 refers to combination of all categories of feature. The evaluation metric F_{avg} was used to evaluate the results. All experiments were conducted using 10-CV in Weka Knowledge Flow (Hall et al., 2009). The results are presented in Section 4.4.

Table 4.8: Experimental Combination of Feature

Experiment	
Set I:	Single feature
Set II:	Two combinations
Set III:	Three combinations
Set IV:	Four combinations
Set V:	Five combinations (All)

4.3.2 Parameter Setting

For the experiments, non-linear SVM was used as classification model because it had been shown to perform well with respect to sarcasm detection as shown in the literature (Joshi et al., 2015; Lunando & Purwarianti, 2013; Ptácek et al., 2014; Riloff et al., 2013; Weitzel et al., 2016; Xu et al., 2015). LibSVM (Chang & Lin, 2011), as provided as part of the Weka data mining workbench, was selected to implement the non-linear SVM. The kernel used was RBF kernel where the parameters were set to C = 3.0 and λ = 0.03. These values were selected based on the recommendations made by Ozdemir and Bergler (2015) and Fersini, Pozzi, and Messina (2015). A series of CV grid search test (which differs in parameter value) was also conducted to find the optimal C and λ values for the dataset used in this thesis. The experiment was conducted using the lexical feature only. In the experiment, the top N features (where N equal to 25%, 50%, 75% and 100%) were used for classification. The experiment was conducted using 10-CV. The value of C tested were in a range of 1.0 to 5.0 with 5 steps iteration (1.0, 2.0, 3.0, 4.0 and 5.0) and λ was 0.01 to 0.05 with 5 steps iteration (0.01, 0.02, 0.03, 0.04 and 0.05).

Table 4.9 shows the best classification result of CV grid search test, to find the optimal value of C and λ using lexical feature (baseline), recorded the best for each top N features as suggested by the experiment. The best performance, F_{avg} of 0.840 was recorded when C = 3.0 and λ = 0.03 using top 50% features, thus was set as

default parameter in the experiments of this thesis. The result conformed the recommendation made by Ozdemir and Bergler (2015) and Fersini et al. (2015).

Table 4.9: Result of CV Grid Search Test

% Feature size	25%	50%	75%	Full
Cost (C)	3.0	**3.0**	2.0	1.0
Gamma (λ)	0.02	**0.03**	0.01	0.02
Result (F_{avg})	0.783	**0.840**	0.785	0.711

4.4 Result and Discussion

This section presents the results of the experiments conducted to identify the best combination of feature categories and evaluate the proposed approach. Evaluations and comparisons are presented in Sub-section 4.4.1 and analysis of the result is presented in Sub-section 4.4.2.

4.4.1 Evaluation and Comparison

For feature selection, the N value was set at 25%, 50% and 75%. The performance of sarcasm detection when applied on all extracted features was also measured (without feature selection). The details of the number of features in each category are given in Table 4.10. The results of the experiment using the proposed features to detect sarcasm are shown in Table 4.11.

Table 4.10: The Number of Features Used for Experimentation

Experiment	% Feature size / Feature	25%	50%	75%	Full
		Feature size			
Set I	1. Lexical (Baseline)	683	1365	2048	2730
	2. Pragmatic	1	2	3	4
	3. Prosodic	17	33	50	66
	4. Syntactic	924	1848	2771	3695
	5. Idiosyncratic	44	89	133	177
Set II	6. Lexical + Pragmatic	684	1367	2051	2734
	7. Lexical + Prosodic	699	1398	2097	2796
	8. Lexical + Syntactic	1606	3213	4819	6425
	9. Lexical + Idiosyncratic	727	1454	2180	2907
	10. Syntactic + Pragmatic	925	1850	2774	3699
	11. Syntactic + Prosodic	940	1881	2821	3761
	12. Syntactic + Idiosyncratic	968	1936	2904	3872
	13. Pragmatic + Prosodic	18	35	53	70
	14. Pragmatic + Idiosyncratic	45	91	136	181
	15. Prosodic + Idiosyncratic	61	122	182	243
Set III	16. Lexical + Pragmatic + Prosodic	700	1400	2100	2800
	17. Lexical + Pragmatic + Idiosyncratic	728	1456	2183	2911
	18. Lexical + Prosodic + Idiosyncratic	743	1487	2230	2973
	19. Syntactic + Pragmatic + Prosodic	941	1883	2824	3765
	20. Syntactic + Pragmatic + Idiosyncratic	969	1938	2907	3876
	21. Syntactic + Prosodic + Idiosyncratic	985	1969	2954	3938
	22. Pragmatic + Prosodic + Idiosyncratic	62	124	185	247
Set IV	23. Lexical + Pragmatic + Prosodic + Idiosyncratic	744	1489	2233	2977
	24. Syntactic + Pragmatic + Prosodic + Idiosyncratic	986	1971	2957	3942
Set V	25. Lexical + Syntactic + Pragmatic + Prosodic + Idiosyncratic	1668	3336	5004	6672

Table 4.11: Sarcasm Detection Performance

Experiment	% Feature size / Feature	25%	50%	75%	Full
		F_{avg}			
Set I	1. Lexical (Baseline)	0.783	0.840	0.775	0.708
	2. Pragmatic	0.397	0.407	0.407	0.407
	3. Prosodic	0.556	0.566	0.567	0.560
	4. Syntactic	0.761	**0.847**	0.724	0.656
	5. Idiosyncratic	0.388	0.427	0.461	0.461
Set II	6. Lexical + Pragmatic	0.777	0.836	0.768	0.702
	7. Lexical + Prosodic	0.782	0.835	0.764	0.705
	8. Lexical + Syntactic	0.757	0.822	0.684	0.538
	9. Lexical + Idiosyncratic	0.778	0.834	0.769	0.707
	10. Syntactic + Pragmatic	0.765	0.850	0.731	0.662
	11. Syntactic + Prosodic	0.756	**0.851**	0.738	0.659
	12. Syntactic + Idiosyncratic	0.766	0.845	0.725	0.657
	13. Pragmatic + Prosodic	0.586	0.594	0.598	0.596
	14. Pragmatic + Idiosyncratic	0.429	0.429	0.429	0.428
	15. Prosodic + Idiosyncratic	0.542	0.590	0.625	0.625
Set III	16. Lexical + Pragmatic + Prosodic	0.773	0.831	0.762	0.708
	17. Lexical + Pragmatic + Idiosyncratic	0.777	0.827	0.763	0.705
	18. Lexical + Prosodic + Idiosyncratic	0.778	0.827	0.762	0.716
	19. Syntactic + Pragmatic + Prosodic	0.755	**0.852**	0.737	0.664
	20. Syntactic + Pragmatic + Idiosyncratic	0.767	0.844	0.733	0.659
	21. Syntactic + Prosodic + Idiosyncratic	0.770	0.846	0.732	0.662
	22. Pragmatic + Prosodic + Idiosyncratic	0.571	0.614	0.641	0.641
Set IV	23. Lexical + Pragmatic + Prosodic + Idiosyncratic	0.770	0.827	0.761	0.707
	24. Syntactic + Pragmatic + Prosodic + Idiosyncratic	0.761	**0.848**	0.739	0.664
Set V	25. Lexical + Syntactic + Pragmatic + Prosodic + Idiosyncratic	0.765	**0.825**	0.674	0.533

From table 4.11, the best performance for single feature groups was produced using the syntactic category (F_{avg} = 0.847) and 50% of the features. The best performance for two combinations was produced by syntactic and prosodic categories with an F_{avg} score of 0.851, using 50% of the features. The best overall performance F_{avg} was recorded by the combination of three categories, using also 50% of the features. The score was recorded by combination of syntactic, pragmatic and prosodic categories with an F_{avg} score of 0.852. The best performance for four combinations was produced by syntactic, pragmatic, prosodic and idiosyncratic categories, recorded an F_{avg} score of 0.848. The combination of all features only recorded the best F_{avg} score of 0.825, which is lower than the baseline (lexical feature) score of 0.840. All best scores were recorded when using the top 50% of the features. The results also showed that the reduced features always produced the best performances compared to when all features were used.

4.4.2 Analysis of Result

For set I, syntactic feature has produced the best sarcasm detection performance, followed by lexical, prosodic, idiosyncratic and pragmatic features respectively. Table 4.12 shows the feature categories (sorted in ascending order according to their performance) with example of features used.

Table 4.12: Feature Performance Ranking for Experiment Set I

	Feature	Example
1	Syntactic	beautiful_ADJ bouquet_NOUN say_VERB ants_NOUN elephants_NOUN
2	Lexical	naik naik letih ooi nak follow semuanya terpaksa
3	Prosodic	oiii
4	Idiosyncratic	"face of sharks"
5	Pragmatic	!!! ???

From Table 4.12, the syntactic feature that employs word-tag pairing able to generate more discriminative features compared to if using the word alone (lexical). This results are similar to the findings of Salvetti et al. (2004) and Xia and Zong (2010). The worst performance was generated from pragmatic feature. This could be due to the fact that the same punctuations can be found not only in sarcastic but also non-sarcastic comments, and also the low size of features. This is supported by Davidov et al. (2010) that found using punctuation alone without combination with other feature type produced the worst classification performance.

In set II, the sarcasm detection performances were better than Set I for most categories of feature. The best result was produced by combination of syntactic and prosodic with an F_{avg} score of 0.851. The combination was better when word-tag pairing coupled with interjection, a prosodic type feature. Interjection were found to be good in distinguishing sarcastic word from non-sarcastic, when combining with other feature categories, supported by the result of (Muresan et al., 2015) and (Bharti et al., 2015). The combination of syntactic and pragmatic produced the second best classification performance may be due to the fact that people commonly used punctuation marks to emphasize sarcastic words in their comments. For example *"Ini smua poyo!!!"*, translated as "This all lame!!!". The combination of pragmatic and idiosyncratic features performed the worst in Set II with the highest F_{avg} score of 0.429. Idiosyncratic phrases, with respect to the dataset used in this work, did not followed by punctuations. Hence, the results produced were similar to Set I when both features were employed separately.

For three categories of feature combined as used in Set III, the combination of syntactic, pragmatic and prosodic features produced an F_{avg} score of 0.852 (the best performance). Based on the result, it is conjectured that people always use the syntactic, pragmatic and prosodic to deliver sarcasm. For example *"Ekonomi xmenentu, tp malaysia makin maju?? Kahkah!!"* ("Economy is uncertain, but Malaysia is more advance?? Haha!!"). Syntactic features are: "Economy_NOUN", "uncertain_ADJ", "Malaysia_NOUN", "more_ADV", and "advance_VERB". Pragmatic features are "??" and "!!", and prosodic feature is "Haha". The worst results, with respect to Set III, were recorded by the combination of pragmatic, prosodic and idiosyncratic features. The missing of features that represent content (lexical) and structure (syntactic) could be

the reason why such performance was recorded. It is conjectured that the inclusion of either lexical or syntactic features could produce better results.

In set IV, the combination of syntactic, prosodic, pragmatic and idiosyncratic features performed the best with an F_{avg} score of 0.848 using the top 50% features. The results is conform with the trend of results produced in Set I to IV where better performances were produced whenever syntactic was used over lexical. However, the performance of sarcasm detection using combination of features in Set IV is lower than Set III. This could be due to inclusion of idiosyncratic feature, where some of them can be found in both sarcastic and non-sarcastic comment. Comparison of syntactic and lexical features is shown in Table 4.13.

Table 4.13: Comparison of Syntactic and Lexical Effectiveness

Feature Group	Best Result (F_{avg})			
	Set I	Set II	Set III	Set IV
	Single feature	Two combinations	Three combinations	Four combinations
Syntactic combination	0.847	0.851	0.852	0.848
Lexical combination	0.840	0.836	0.831	0.827

Finally in set V for five combinations (combination of all features), the performance recorded was F_{avg} score of 0.825, lower than baseline of lexical single feature category. This is might be due to repetition of same features, where lexical feature from bilingual text duplicated in syntactic representation (after translation and syntactic feature extraction). For example lexical feature of *"naik naik letih ooi nak follow semuanya terpaksa"* also appeared in syntactic feature as "up_ADV up_ADV tired_ADJ going_NOUN follow_VERB forced_ADJ". This had increased the size of feature considered for classification and consequently effect the ranking of feature produced for feature selection. As a result, some of the features selected may be redundant. Table

4.14 shows an example of cases for Set III and V where occurrence of similar features may affect the detection performance, compared to actual label (human annotation). The best combination of set III able to predict sarcasm for some comment such as *"Pi mai pi mai tang tu ja.."* ("Go let's go let it alone") and *"bajet cool"* ("cool budget"), while set V failed to predict. Deeper observation found that, the former was represented in set III as "go_VERB lets_ADJ go_VERB lets_ADJ just_ADV", while in Set IV it was represented as "pergi mari pergi mari sahaja go_VERB lets_ADJ go_VERB lets_ADJ just_ADV". The repetition of similar features (although they are in different language) may affect the detection performance.

Table 4.14: Examples of Prediction by Set III and Set V

No.	Comment	Translation	Actual Label	Prediction	
				Set III	Set V
1	*"Pi mai pi mai tang tu ja.."*	"Go let's go let it alone"	Sarcastic	Correct	Wrong
2	*"bajet cool"*	"cool budget"	Sarcastic	Correct	Wrong

The idiosyncratic feature category did not perform well compared to the other features. It is conjectured that the method used to extract the idiosyncratic features could have affected the performance. Two possible issues identified. First, the idiosyncratic feature was extracted from translated bilingual text using the phrase of "noun-adposition-noun" where phrases that are not idiosyncratic could be incorrectly identified as idiosyncratic. Examples for such case are "budget on paper" and "people of Malaysia". To address this, each phrase should be manually annotated by annotators to identify the peculiar and odd phrase. Second, the limitation of automatic translation from Malay to English may hamper the translation process. Hence, the creation of corpus and use of a Malay idiosyncratic feature category might resolve this issue. It is conjectured that if the abovementioned strategies are conducted, better detection of sarcasm can be recorded using the idiosyncratic feature.

4.5 Summary

A process to extract features to detect sarcasm in bilingual texts using combinations of categories of NLP based features has been presented. The process extracts the features from dataset in bilingual and translated form. Five categories of NLP feature were considered: lexical, pragmatic, prosodic, syntactic and idiosyncratic. A non-linear SVM was used for classification purposes with respect to sarcasm detection, to evaluate the performance of the features to detect sarcasm in text. Comparison with a baseline feature demonstrated that the proposed features performed better. The comparison also concluded that the best combination of feature categories for sarcasm detection in the context of Malay social media data was the combination of syntactic, pragmatic and prosodic feature categories. This feature combination will be used as part of the framework to support SA described in the following chapter.

CHAPTER 5

SENTIMENT ANALYSIS WITH SARCASM DETECTION AND CLASSIFICATION FRAMEWORK

5.1 Introduction

This chapter describes a framework of SA where sarcasm detection and classification is considered. The concept framework to extract opinionated information from text proposed by Cambria et al. (2015), coupled with sarcasm detection framework proposed by Muresan et al. (2015) and Bharti et al. (2015) used as a basis of the framework presented in this chapter. Muresan et al. (2015) produced final output of positive, negative, or sarcasm while Bharti et al. (2015) produced final output of actual positive, actual negative, or sarcastic. The significant differences of the proposed framework over the available ones found in the literature are the two sentiment classification steps: the first is 'initial' sentiment classification and the second is 'actual' sentiment classification. The latter considers sarcasm detection and classification, with which the features identified in Chapter 4 is employed, before sentiment label is predicted for the texts. The rest of this chapter is organized as follows. The detail of the proposed framework is presented in Section 5.2. The experimental setup is explained in Section 5.3. Section 5.4 presents the results obtained and discussion of results. Summary of this chapter is presented in Section 5.5.

5.2 The Framework for Sentiment Analysis with Sarcasm Detection and Classification

In this thesis, the proposed framework to support SA consists of six modules: (i) preprocessing, (ii) feature extraction, (iii) feature selection, (iv) initial sentiment classification, (iv) sarcasm detection and classification, and (v) actual sentiment classification. Figure 5.1 shows the proposed framework. Initial sentiment refers to sentiment classification performed without considering sarcasm detection that is common in other work found in the literature. Actual sentiment on the other hand refers to sentiment classification that considers the sarcastic content of the texts.. In this framework, the sarcasm detection and classification module is the most critical. The ability of the proposed module to identify sarcasm texts accurately will results in better actual sentiment classification over the initial sentiment classification. To identify sarcasm in texts or comments, the features presented in Chapter 4 was employed.. The details of initial sentiment classification, sarcasm detection and classification, and actual sentiment classification modules are explained in Sub-section 5.2.1 to Sub-section 5.2.3. The details of preprocessing, feature extraction and selection modules have been previously described in Chapter 3 and 4 respectively.

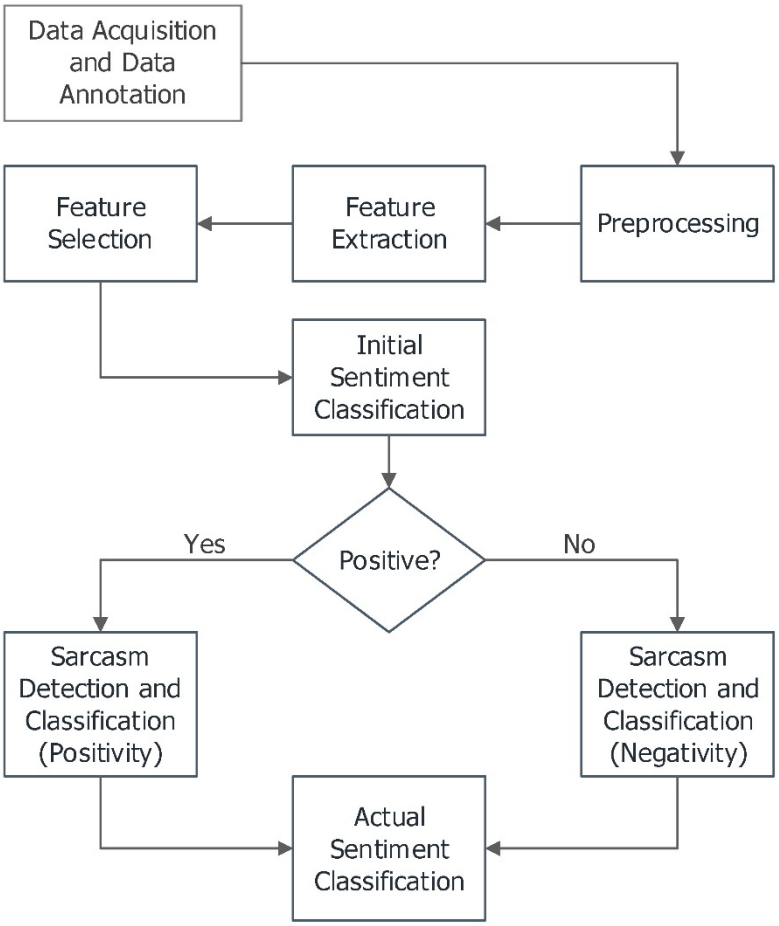


Figure 5.1: The Framework to Support SA Using Sarcasm Detection and Classification

5.2.1 Initial Sentiment Classification

Initial sentiment classification module classifies an opinion or text as either positive or negative. Figure 5.2 shows the classification process for initial sentiment module. In the work presented in this thesis, non-linear SVM has been used to generate the classifier, because it has been shown to perform well in the context of supervised classification (Muresan et al., 2015). In this module, the same parameter setting as described in Chapter 4 was used (see Sub-section 4.3.2 for detail).

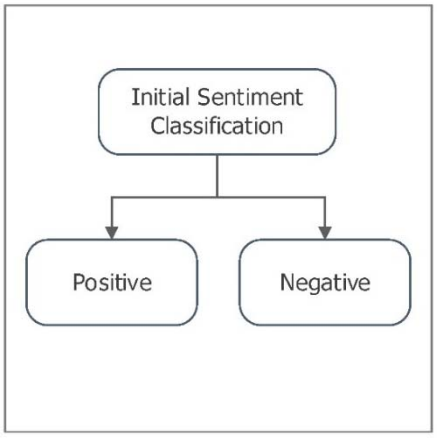


Figure 5.2: Initial Sentiment Classification Module

From the given comment, initial sentiment classification will classify it as either positive or negative. Table 5.1 shows some examples of comment extracted from the dataset used in this thesis.

Table 5.1: Initial Sentiment Prediction of Comments

No.	Comment	Translation	Initial Sentiment Prediction
1	*"Alhamdulillah…"*	"thank god"	Positive
2	*"Bodoh paluii"*	"stupid stupid"	Negative
3	*"woww igt byk ke 500 tu"*	"wow remember much into it"	Positive
4	*"Hahaha..... Rakyat dibadutkan....."*	"haha people in clown"	Negative

5.2.2 Sarcasm Detection and Classification

The aim of this module is to identify comments containing sarcastic features. This module has two sub-processes, which are sarcasm detection and sarcasm classification. Figure 5.3 shows the process of sarcasm detection and sarcasm classification for each comment after initial sentiment classification is performed.

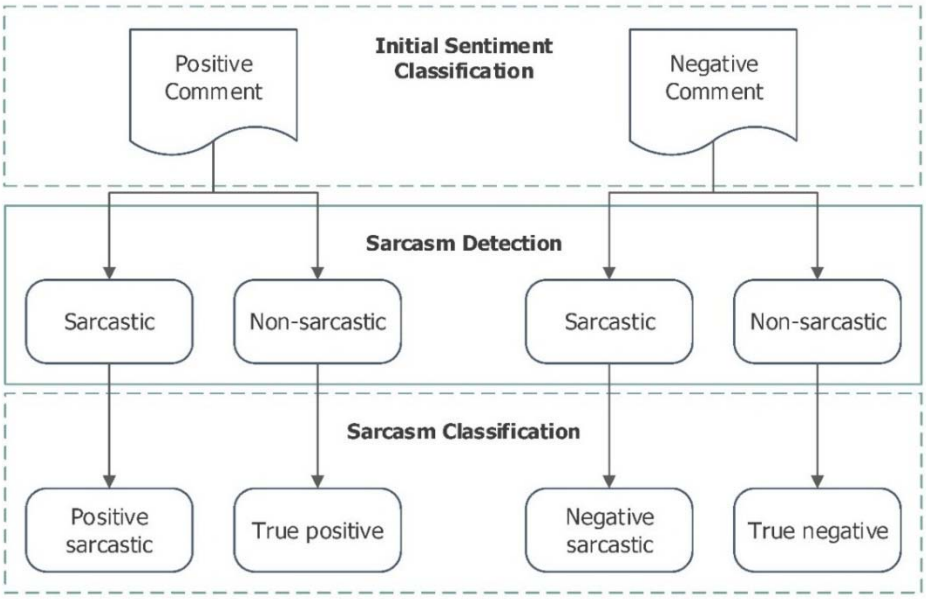


Figure 5.3: Sarcasm Detection and Sarcasm Classification of the Sentiment After Initial Sentiment Classification

Comments or texts that have been identified as positive or negative by the initial sentiment classification will be further classified as either containing sarcastic or non-sarcastic contents. Table 5.2 shows example of sarcasm detection process performed on the same comments as shown in Table 5.1. The process will assign sarcastic or non-sarcastic labels to the comments.

Table 5.2: Sarcasm Detection on Comments

No.	Comment	Translation	Sarcasm Detection
1	*"Alhamdulillah…"*	"thank god"	Non-sarcastic
2	*"Bodoh paluii"*	"stupid stupid"	Non-sarcastic
3	*"woww igt byk ke 500 tu"*	"wow remember much into it"	Sarcatic
4	*"Hahaha..... Rakyat dibadutkan....."*	"haha people in clown"	Sarcastic

Sarcasm classification process consists of two sub-processes; (i) sarcasm positivity classification, and (ii) sarcasm negativity classification. Comments that have been identified as positive by initial sentiment classification and are sarcastic/ non-sarcastic, will be further classified as positive sarcastic/ true positive. Similarly, comments that have been identified as negative by initial sentiment classification and are sarcastic/ non-sarcastic, will be further classified as negative sarcastic/ true negative. This process is named as sarcasm negativity classification. Table 5.3 shows the same examples shown in Table 5.2 after sarcasm classification is performed.

Table 5.3: Sarcasm Classification of Comments

No.	Comment	Translation	Sarcasm Classification
1	*"Alhamdulillah…"*	"thank god"	True positive
2	*"Bodoh paluii"*	"stupid stupid"	True negative
3	*"woww igt byk ke 500 tu"*	"wow remember much into it"	Positive Sarcastic
4	*"Hahaha..... Rakyat dibadutkan....."*	"haha people in clown"	Negative Sarcastic

5.2.3 Actual Sentiment Classification

As mentioned in the foregoing sections, sarcastic content tends to reverse the actual sentiment of the comments. Therefore, once sarcasm is detected and labeled (positive or negative sarcastic), actual sentiment classification can be performed using polarity flip (Burgers, 2010; Kunneman et al., 2015). This polarity flip is employed to reverse the initial sentiment classification results based on linguistic hypothesis. Two strategies are considered in this thesis: (i) to flip all comments with sarcastic label (positive and negative sarcastic) or (ii) to flip only positive comments with sarcastic label (positive sarcastic). Each is described in further details below.

i. Flip both positive sarcastic and negative sarcastic

The first strategy employs the hypothesis that considers sarcasm as something that is opposite of what the speaker means (R. Gibbs, 2007; Raymond W Gibbs, 1986). When sarcastic content is used in a positive statement, the actual sentiment is negative, and vice versa (Liebrecht et al., 2013; B. Liu, 2015). Based on these, the polarity of positive sarcastic comments will be flipped to negative, and negative sarcastic comments will be flipped to positive. Figure 5.4 shows the polarity flip of sentiment based on this hypothesis.

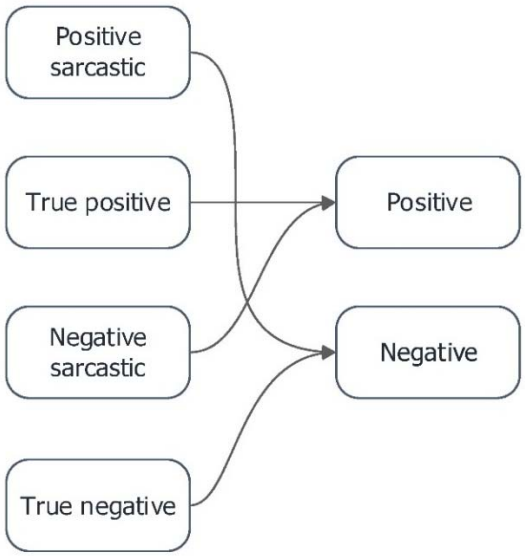


Figure 5.4: Polarity Flip of Both Positive Sarcastic and Negative Sarcastic Based on the First Hypothesis

Using example of comments shown in the Table 5.3, the output of actual sentiment classification is shown in Table 5.4. Comments that are labeled as true positive and positive sarcastic by the sarcasm detection and classification module are labeled as positive. On the other hand, comments that are labeled as true negative and positive sarcastic by the sarcasm detection and classification module are labeled as negative

Table 5.4: Actual Sentiment Classification of Comments (Flip Both Positive and Negative Sarcastic)

No.	Comment	Translation	Sarcasm Classification		Actual Sentiment (Flip Both)
1	*"Alhamdulillah…"*	"thank god"	True positive	→	Positive
2	*"Bodoh paluii"*	"stupid stupid"	True negative	→	Negative
3	*"woww igt byk ke 500 tu"*	"wow remember much into it"	Positive Sarcastic	→	Negative
4	*"Hahaha..... Rakyat dibadutkan....."*	"haha people in clown"	Negative Sarcastic	→	Positive

ii. Flip positive sarcastic only

The second hypothesis states that sarcasm in a negative statement does not always deliver the opposite of what the speaker meant. This hypothesis is derived from sarcasm linguistic study (Attardo, 2000; Roger J Kreuz & Glucksberg, 1989) and computational experiments (Bouazizi & Ohtsuki, 2015; Riloff et al., 2013). Bouazizi and Ohtsuki (2015) focused on identifying method to resolve the misclassification. They managed to increase the success identification of negative comments when sarcasm is considered. However, this is not the case for positive comments. This is due to most sarcastic comments are basically negative comments that have been missclassified as positive. Another work of Riloff et al. (2013) produced similar results where sarcastic comments is less common used in negative comments. Using this hypothesis, only the polarity of positive sarcastic comments will be flipped to negative, while negative sarcastic comments remain negative. Figure 5.5 shows the polarity flip based on the second hypothesis. Table 5.5 shows example of actual sentiment classification, where flip positive sarcastic only hypothesis is used, applied on the same comments used in the previous modules.

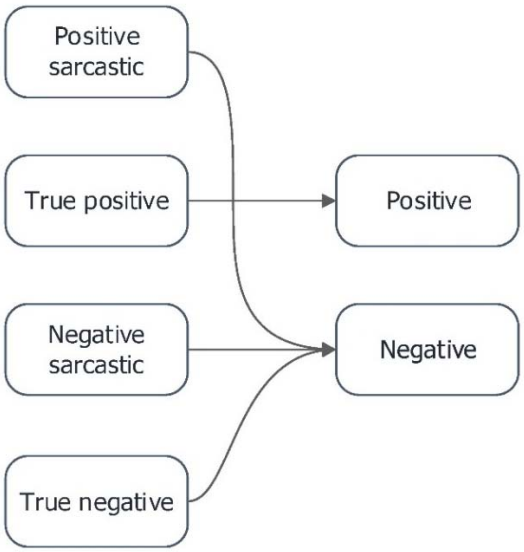


Figure 5.5: Polarity Flip of Positive Sarcastic Only, Based on Second Hypothesis

Table 5.5: Actual Sentiment Classification of Comments (Flip Positive Sarcastic Only)

No.	Comment	Translation	Sarcasm Classification	Actual Sentiment (Flip Positive Only)
1	*"Alhamdulillah…"*	"thank god"	True positive →	Positive
2	*"Bodoh paluii"*	"stupid stupid"	True negative →	Negative
3	*"woww igt byk ke 500 tu"*	"wow remember much into it"	Positive Sarcastic →	Negative
4	*"Hahaha..... Rakyat dibadutkan....."*	"haha people in clown"	Negative Sarcastic →	Negative

5.3 Experimental Setup

This section describes the objective of the experiment and the adopted parameter setting to evaluate the proposed framework.

5.3.1 Experiment Objective

The aim of the experiment was to evaluate the effectiveness of the proposed framework of sentiment analysis that considers sarcasm detection and classification. To achieve this, four experiment objectives were identified. The first was to identify the performance of initial sentiment classification on the dataset and features used (details of the features used in the experiments are described in Sub-section 5.3.3). The second and third were conducted to evaluate the performances of sarcasm positivity and negativity classifications. The final objective was to evaluate the performance of actual sentiment classification, in which a polarity flip was used to reverse the initial sentiment of comments where sarcasm occurred. The adopted evaluation metric F_{avg} was used to evaluate the results. All experiments were conducted using 10-CV and the Weka Knowledge Flow (Witten, Frank, & Hall, 2011). The performance of sarcasm detection was not measured in this chapter as it has been performed and reported in Chapter 4.

5.3.2 Parameter Setting

The same parameter settings as described in Sub-section 4.3.2 was used in this chapter. The non-linear SVM was used as the classification model to conduct the experiment. The variation of non-linear SVM used was LibSVM (Chang & Lin, 2011). The kernel function used is RBF with parameters of C = 3.0 and λ = 0.03 (this is the same setting as described in Sub-section 4.3.2).

5.3.3 Preprocessing, Feature Extraction and Feature Selection

The dataset used for evaluation is as described in Chapter 3. The dataset was tokenized, spellchecked, and stopword removed.

For feature extraction, the best categories of feature for sarcasm detection identified in Chapter 4 (see Section 4.2 for detail) were used, which are syntactic, pragmatic and prosodic. For comparison purpose, lexical features (the common features used in the literature for SA) were also used. All the extracted features were than vectorized and normalized to document length.

With respect to feature selection, the top 25%, 50% and 75% of the features were selected based on the Pearson's correlation coefficient ranking. Details of the number of features for each set of experiment and the size of the dataset are given in Table 5.6.

Table 5.6: The Number of Features Used for Experimentation

Experiment	Dataset size	% Feature size			
		25%	50%	75%	Full
Initial sentiment classification (Baseline)	1970	681	1362	2042	2723
Initial sentiment classification		941	1883	2824	3765
Sarcasm detection					
Sarcasm positivity classification	802	514	1028	1542	2056
Sarcasm negativity classification	1168	686	1372	2058	2744
Actual sentiment classification	1970	941	1883	2824	3765

5.4 Result and Discussion

This section presents the results of the experiment to evaluate the proposed framework. Evaluation and comparison are presented in Sub-section 5.4.1 and analysis of the result is presented in Sub-section 5.4.2.

5.4.1 Evaluation and Comparison

Sub-section 5.4.1.1 presents the results of initial sentiment classification. Sarcasm positivity and negativity classification results are presented in Sub-subsection 5.4.1.2. Sub-subsection 5.4.1.3 discusses the actual sentiment classification results.

5.4.1.1 Initial Sentiment Classification

In this set of experiment, the comments were classified as either positive or negative. Table 5.7 shows the comparison of results of initial sentiment classification using the lexical features (baseline) and the features identified in Chapter 4 (see Section 4.2 for detail). The best sentiment classification performance was recorded when using the top 25% features of combination of syntactic, pragmatic and prosodic, with an F_{avg} score of 0.839. The worst was recorded when all features were used for classification (F_{avg} = 0.611).

Table 5.7: Results of Initial Sentiment Classification

% Feature size	25%	50%	75%	Full
Experiment	F_{avg}			
Initial sentiment classification (Baseline)	0.819	0.810	0.757	0.713
Initial sentiment classification	**0.839**	0.623	0.754	0.611

5.4.1.2 Sarcasm Positivity and Negativity Classification

Table 5.8 shows the results for sarcasm classification. For sarcasm positivity classification, the comments were classified as either positive sarcastic or true positive. The best result was recorded when using the top 25% of the features, with an F_{avg} score of 0.942. Sarcasm negativity classification that classify the comments into negative sarcastic or true negative produced the best F_{avg} score of 0.909, which is lower than to the performance of sarcasm positivity classification. This is also true in most cases of different feature sizes. This may be due to difficulties to recognize the negative sarcastic features from true negative comments compared to sarcasm positivity classification. This is supported by Giora (1995) that found some sarcastic negative forms are more difficult to identify.

Table 5.8: Results of Sarcasm Classification

% Feature size	25%	50%	75%	Full
Experiment	F_{avg}			
Sarcasm positivity classification	**0.942**	0.787	0.776	0.767
Sarcasm negativity classification	0.909	0.797	0.614	0.593

5.4.1.4 Actual Sentiment Classification

The results of the actual sentiment classification are shown in Table 5.9. In this experiment, the comments were classified as either positive or negative. For the first strategy, where both positive and negative sarcastic comments polarity were flipped, the best F_{avg} score recorded was 0.899 using the top 25% features. On the other hand, the second strategy that flipped the polarity of positive sarcastic comments only recorded the best F_{avg} of 0.905 (also using the top 25% features). Table 5.9 also shows that the second strategy always performed better than the first.

Table 5.9: Results of Actual Sentiment Classification

% Feature selection size	25%	50%	75%	Full
Experiment	F_{avg}			
Actual sentiment classification (Flip both positive sarcastic & negative sarcastic)	0.899	0.715	0.671	0.666
Actual sentiment classification (Flip positive sarcastic only)	**0.905**	0.903	0.903	0.900

Comparing the actual sentiment classification result towards initial sentiment classification result (Table 5.7), the actual sentiment classification (either flips both positive sarcastic and negative sarcastic or flip positive sarcastic only) results outperformed the initial sentiment classification results in all cases. This indicates that

considering sarcasm detection and classification in sentiment classification produced better SA.

5.4.2 Analysis of Result

Based on the results shown in Table 5.7 to Table 5.9, the performance of sentiment classification was improved by 6.6% after considering sarcasm contents (the initial sentiment classification produced the best F_{avg} of 0.839 while the actual sentiment classification produced the best F_{avg} of 0.905). The actual sentiment performance was also improved by 8.6% after considering sarcasm contents for initial sentiment classification using lexical feature as baseline. The results also indicate that, based on the dataset used in this thesis, the polarity only reverses in positive sarcastic comments; sarcasm in negative comments does not always reverse the polarity, hence the results. Table 5.10 shows the improvement of best actual sentiment classification (flip positive sarcastic only) against initial sentiment classification.

Table 5.10: Actual Sentiment Classification Comparison Against Initial Sentiment Classification

Experiment	Best Result (F_{avg})	Actual Sentiment Improvement
Actual sentiment classification (Flip positive sarcastic only)	**0.905**	
Initial sentiment classification (Baseline)	0.819	8.6%
Initial sentiment classification	0.839	6.6%

In most cases, the actual sentiment classification that considers sarcasm detection and classification able to correct the misclassification of sentiment produced by the initial sentiment classification. Table 5.11 shows examples of such cases found in the experiments conducted. The comments were incorrectly classified as positive in

the initial sentiment classification module. This may be due to the existence of positive words in the comments. Using the features presented in the foregoing chapter, sarcasm was detected and classified accordingly. Finally the actual sentiment correctly reclassify the comments as negative, similar to actual label (human annotation).

Table 5.11: Example of Actual Sentiment Classification Over Initial Sentiment Classification

No.	Comment	Syntactic, Pragmatic, and Prosodic Feature	Predicted Initial Sentiment	Predicted Sarcasm Detection/ Classification	Predicted Actual Sentiment	Actual Label
1	*"tsk tsk tsk. xlme lg yuran nek la tu. good job"*	isk isk isk, soon_ADV fee_NOUN will_VERB rise_VERB good_ADJ job_NOUN isk isk isk lah	Positive	Positive sarcastic	Negative	Negative
2	*"Tk yah kenakan cukai langsung lahh.Senang crita"*	need_NOUN impose_VERB direct_ADJ taxation_NOUN is_VERB happy_ADJ story_NOUN lah	Positive	Positive sarcastic	Negative	Negative

However, there are minor cases that prediction is unable to correctly classify the comments. The comments: *"Yeyyyy!!!! bestnya dpat brim..xsbar nk tggu thun dpan.."* extracted as "best_ADJ got_VERB brim_NOUN can_VERB not_ADV wait_VERB next_ADJ year_NOUN yay !!!", and *"Wow, rakyat malaysia kita masyukkkk… hahahaha...."* extracted as "malaysian_NOUN earning_VERB wow haha", were misclassified by the system as negative. An annotator believed they were negative since the stylistic comment delivered was intended to reverse from literal meaning. While two annotators believe that positive sentiment was intended as the words used in the comment show the happy and joy mood. This might be due to different method used by speaker to

deliver sarcasm in their message. It is conjectured that by considering contextual feature of user's profile will determine the sentiment of the comment (Amir, Wallace, Lyu, & Silva, 2016; Bamman & Smith, 2015). User embedded profile such as user information and historical comments can be used in conjunction with other feature category for feature extraction that could better identify sarcastic contents. Table 5.12 shows some error found in misclassification for actual sentiment prediction.

Table 5.12: Examples of Actual Sentiment Misclassification

No.	Comment	Syntactic, Pragmatic, and Prosodic Feature	Predicted Initial Sentiment	Predicted Sarcasm Detection/ Classification	Predicted Actual Sentiment	Actual Label
1	*"Yeyyyy!!!! bestnya dpat brim..xsbar nk tggu thun dpan.."*	best_ADJ got_VERB brim_NOUN can_VERB not_ADV wait_VERB next_ADJ year_NOUN yay !!!	Positive	Positive sarcastic	Negative	Positive
2	*"Wow, rakyat malaysia kita masyukkkk… hahahaha...."*	malaysian_NOUN earning_VERB wow haha	Positive	Positive sarcastic	Negative	Positive

Another example is shown in Table 5.13, where errors occur in actual sentiment classification due to failure in initial and sarcasm classification. The phrase: *"hidup segan mati xnk.."* translated as "feel don't want to live or dead" was extracted from the comment. The original comment was translated to "just like feel alive dead do not want to", then preprocessed and extracted as "VERB just_ADJ feel_VERB alive_ADJ dead_ADJ do_VERB not_ADV want_VERB". The predicted initial sentiment was negative, as negative words can be found in the comments. After sarcasm detection, it is found sarcastic and classified to negative sarcastic. The actual sentiment classification later predicts the comment as negative. The misclassification might be due to bad translation and used of negative words to convey positive message, the latter is the rarest case. It is conjectured that using the contextual valence shifter will return sentiment orientation

of the comment in initial sentiment (Kennedy & Inkpen, 2006; Valitutti & Veale, 2015). Valence shifter such as negations, intensifiers, and diminishers could be used to identify sentiment polarity for actual sentiment by means of computing polarity score for each comment.

Table 5.13: Examples of Misclassification by the Proposed Framework

Comment	Syntactic, Pragmatic, and Prosodic Feature	Predicted Initial Sentiment	Predicted Sarcasm Detection/ Classification	Predicted Actual Sentiment	Actual Label
"Sbnarnye senang je ipta ni nk cari dana y bnyk.. Naikkn yuran jela.. Mmg sblm ni pun mcm tu.. Koperasi sendiri ada tp mcm hidup segan mati xnk.."	actually_ADV easy_ADJ ipta_NOUN want_VERB raise_VERB funds_NOUN just_ADJ increased_VERB fee_NOUN indeed_ADV same_ADJ own_VERB cooperatives_NOUN had_VERB just_ADJ feel_VERB alive_ADJ dead_ADJ do_VERB not_ADV want_VERB lah	Negative	Negative sarcastic	Negative	Positive

5.5 Summary

This chapter presents a framework to support SA by utilizing sarcasm detection and classification. Six modules proposed; preprocessing, feature extraction, feature selection, initial sentiment classification, sarcasm detection and classification, and actual sentiment classification. The most significant contribution of this framework is the sarcasm detection and classification module and the actual sentiment classification module. The former is used to identify comments that contain sarcastic features. The latter is used to revise the initial sentiment predicted of those comments. A non-linear SVM was used for classification purpose. Comparison of SA without sarcasm detection and classification (initial sentiment classification) against with sarcasm detection and

classification (actual sentiment classification) shows that the latter produced better classification performance.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Introduction

This chapter summarizes the research work on SA on bilingual social media data that considers sarcasm detection and classification. The main findings, contributions of the research, and some possible future research directions are presented in Sections 6.2, 6.3 and 6.4 respectively.

6.2 Research Summary

The motivation of the research presented in this thesis was to produce an approach for social media SA on bilingual text that considers sarcasm detection and classification to make sentiment prediction. To achieve this, the research work focuses on (i) the process of identification of features for sarcasm detection and classification, and (ii) the framework of SA where sarcastic content is considered.

The proposed feature extraction process to detect sarcasm in bilingual texts used combination of different categories of NLP based features. The process extracts the features from dataset in both bilingual and English translated form. Five categories of NLP feature were considered: lexical, pragmatic, prosodic, syntactic and idiosyncratic. A non-linear SVM was used for classification purpose to evaluate the proposed features in sarcasm detection. The reported evaluation indicates that the combination of syntactic, pragmatic and prosodic feature categories produced the best results for sarcasm detection and classification in the context of the dataset used.

The proposed framework of SA includes sarcasm detection and classification to support SA. It consists of six modules: preprocessing, feature extraction, feature selection, initial sentiment classification, sarcasm detection and classification, and actual sentiment classification. A non-linear SVM was used for classification purpose in a series of experiment. The reported evaluation indicates that using the proposed framework better performance produced by actual sentiment classification (sarcasm detection and classification was considered) compared to initial sentiment classification (without sarcasm detection and classification).

6.3 Main Finding and Contributions

This section describes the main findings and contributions of the reported research in the context of the research question and objectives identified in Section 1.3. Each identified research objective is considered in turn, and the manner in which the proposed research work addresses each objective considered.

1. To investigate and identify features for sarcasm detection on bilingual social media data.

   The first objective was derived from the research question: *"What are the features that can be extracted from social media containing bilingual data that can better identify sarcasm features?"*. The proposed feature extraction process presented in Chapter 4 concluded that three categories of feature can be extracted from bilingual social media data and able to identify sarcasm feature better as shown in the evaluation section of Chapter 4. The features are syntactic, pragmatic, and prosodic.

2. To investigate and implement a framework for SA with sarcasm detection and sarcasm classification to produce better sentiment classification's performance.

   The second research objective was derived from the second research question: *"How the sarcasm detection and classification can be employed into SA system?"*. According to the work presented in Chapter 5, sarcasm detection and

classification can be incorporated into SA system using the proposed framework. The proposed framework consists of six modules: preprocessing, feature extraction, feature selection, initial sentiment classification, sarcasm detection and classification, and actual sentiment classification. The sarcasm detection and classification module will identify comments that contain sarcastic features, which will eventually cause misclassification of sentiment produced by initial sentiment classification. Once sarcasm is detected in comments, the actual sentiment classification module will correct the initial sentiment prediction. The evaluation conducted shows the superiority of the proposed framework over other framework found in the literature with respect to SA.

3. To evaluate and compare the result of the proposed approach in (1) and (2). The evaluation of the proposed features and framework in (1) and (2) can be found in Chapter 4 and 5. The evaluation and comparison showed that the identified features can be used to accommodate the sarcasm detection and classification module of the proposed framework. Both work conducted in (1) and (2) should be used together in order to produced better SA results, as shown in the evaluation sections in Chapter 5.

Highlighting the main research questions in Chapter 1: "*What is the appropriate approach to classify sentiment using sarcasm detection and classification for bilingual social media data?"*, the work described in Chapter 3, and particularly Chapters 4 and 5, provide the answer to this question as they have showed that (i) detecting sarcasm in comments by means of identifying features that are indicator of sarcastic content and (ii) using the output of (i) to reconsider the initial sentiment classification able to produce improved sentiment classification.

The main contributions of the work described in this thesis are thus summarized as follows:

1. A process to identify sarcasm features for sarcasm detection on bilingual social media data.
2. A framework for SA that considers sarcasm detection and classification.

6.4 Future Work

In this section, some identified potential future research directions are listed.

1. Additional feature category and diversification of feature types. To better detect sarcasm and improve SA, various feature types can be added to the existing three categories. Pragmatic features such as slang and emoticon may be used to provide more meaningful features. Contextual features such as user's profile also can be considered to be extracted.
2. To address the issue of preprocessing. Spellchecking and translation can be improved to provide better cleaned text for feature extraction. Building large corpus of common misspelling words made by social media user, and building translation tool specifically for bilingual text can overcome those issues.
3. Application on an alternative bilingual dataset. The proposed approaches were evaluated using a Malay bilingual social media dataset. It would be of interest to investigate if the presented work can be applied to other bilingual dataset.

REFERENCES

Agarwal, B., & Mittal, N. (2016). *Prominent Feature Extraction for Sentiment Analysis*: Springer.

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*: Springer-Verlag New York.

Al-Moslmi, T., Gaber, S., Al-Shabi, A., Albared, M., & Omar, N. (2015). Feature Selection Methods Effects on Machine Learning Approaches in Malay Sentiment Analysis.

Alsaffar, A., & Omar, N. (2014). Study on feature selection and machine learning algorithms for Malay sentiment classification. *Information Technology and Multimedia (ICIMU), 2014 International Conference on*.

Altrabsheh, N., Cocea, M., & Fallahkhair, S. (2015). Detecting sarcasm from students feedback in Twitter. *Design for Teaching and Learning in a Networked World*, 551-555.

Amir, S., Wallace, B. C., Lyu, H., & Silva, P. C. M. J. (2016). Modelling context with user embeddings for sarcasm detection in social media. *arXiv preprint arXiv:1607.00976*.

Attardo, S. (2000). Irony as relevant inappropriateness. *Journal of pragmatics, 32*(6), 793-826. doi:10.1016/s0378-2166(99)00070-3

Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., & Varma, V. (2012). Mining sentiments from tweets. *Proceedings of the WASSA, 12*.

Balahur, A., & Jacquet, G. (2015). Sentiment analysis meets social media–Challenges and solutions of the field in view of the current information sharing context: Elsevier.

Balahur, A., & Turchi, M. (2013). *Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data*. In RANLP (pp. 49-55).

Balahur, A., & Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language, 28*(1), 56-75. doi:http://dx.doi.org/10.1016/j.csl.2013.03.004

Bamman, D., & Smith, N. A. (2015). *Contextualized sarcasm detection on Twitter*. In Proceedings of the 9th International Conference on Web and Social Media (pp. 574-577): AAAI Menlo Park, CA.

Barbieri, F., Ronzano, F., & Saggion, H. (2015). *UPF-taln: SemEval 2015 Tasks 10 and 11 Sentiment Analysis of Literal and Figurative Language in Twitter*. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (pp. 704).

Bharti, S. K., Babu, K. S., & Jena, S. K. (2015). *Parsing-based sarcasm sentiment recognition in Twitter data*. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (pp. 1373-1380).

Bikel, D., & Zitouni, I. (2012). *Multilingual Natural Language Processing Applications: From Theory to Practice*: IBM Press.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*: " O'Reilly Media, Inc.".

Blinov, P., Klekovkina, M., Kotelnikov, E., & Pestov, O. (2013). Research of lexical approach and machine learning methods for sentiment analysis. *Computational Linguistics and Intellectual Technologies, 2*(12), 48-58.

Bouazizi, M., & Ohtsuki, T. (2015). *Opinion mining in Twitter: How to make use of sarcasm to enhance sentiment analysis*. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. Paris, France (pp. 1594-1597). Paris, France.

Burgers, C. F. (2010). *Verbal irony: Use and effects in written discourse*: [Sl: sn].

Buschmeier, K., Cimiano, P., & Klinger, R. (2014). *An impact analysis of features in a classification approach to irony detection in product reviews*. In Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 42-49).

Cambria, E., Poria, S., Bisio, F., Bajpai, R., & Chaturvedi, I. (2015). *The CLSA model: A novel framework for concept-level sentiment analysis.* Paper presented at the International Conference on Intelligent Text Processing and Computational Linguistics.

Carvalho, P., Sarmento, L., Silva, M. J., & de Oliveira, E. (2009). *Clues for detecting irony in user-generated contents: Oh...!! it's "so easy" ;-).* Paper presented at the Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion - TSA '09, New York, New York, USA.

Castelvecchi, D. (2016). Deep learning boosts Google Translate tool. Retrieved from http://www.nature.com/news/deep-learning-boosts-google-translate-tool-1.20696 doi:doi:10.1038/nature.2016.20696

Chaffey, D. (2016, Feb 16). Global social media research summary 2016. *Smart Insights (Marketing Intelligence) Ltd.* Retrieved from http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

Chandrakala, S., & Sindhu, C. (2012). Opinion mining and sentiment classification: A survey. *ICTACT journal on soft computing, 3*(1), 420-425.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST), 2*(3), 27.

Chew, P. A. (2013). *Critiquing text analysis in social modeling: best practices, limitations, and new frontiers*. In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (pp. 350-358): Springer.

Dang, Y., Zhang, Y., & Chen, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems, 25*(4), 46-53.

Daniel, J., & James, H. (2009). Speech and Language processing: An introduction to natural language processing. *Computational Linguistics and Speech Recognition, 2nd Ed., Prentice Hall*.

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., & Zhou, Q. (2016). Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive Computation, 8*(4), 757-771.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. In Proceedings of the 12th international conference on World Wide Web (pp. 519-528): ACM.

Davidov, D., Tsur, O., & Rappoport, A. (2010). *Semi-supervised recognition of sarcastic sentences in Twitter and Amazon*. In CoNLL 2010 - Fourteenth Conference on Computational Natural Language Learning, Proceedings of the Conference (pp. 107-116).

de Freitas, L. A., Vanin, A. A., Hogetop, D. N., Bochernitsan, M. N., & Vieira, R. (2014). *Pathways for irony detection in tweets.* Paper presented at the Proceedings of the 29th Annual ACM Symposium on Applied Computing - SAC '14, New York, New York, USA.

Dress, M. L., Kreuz, R. J., Link, K. E., & Caucci, G. M. (2008). Regional variation in the use of sarcasm. *Journal of Language and Social Psychology, 27*(1), 71-85.

Farzindar, A., & Inkpen, D. (2015). *Natural Language Processing for Social Media* (Vol. 8): Morgan & Claypool Publishers.

Fersini, E., Pozzi, F. A., & Messina, E. (2015). *Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers*. In Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on (pp. 1-8).

Filatova, E. (2012). *Irony and sarcasm: Corpus generation and analysis using crowdsourcing*. In LREC 2012 - Eighth International Conference on Language Resources and Evaluation. 55-57, Rue Brillat-Savarin, Paris, 75013, France (pp. 392-398). 55-57, Rue Brillat-Savarin, Paris, 75013, France: EUROPEAN LANGUAGE RESOURCES ASSOC-ELRA.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin, 76*(5), 378.

Fleiss, J. L., Nee, J. C., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological bulletin, 86*(5), 974.

Forman, G. (2007). Feature selection for text classification. *Computational methods of feature selection, 1944355797*.

Gao, Y., & Sun, S. (2010). *An empirical evaluation of linear and nonlinear kernels for text classification using support vector machines*. In Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on (pp. 1502-1505): IEEE.

Ghorbel, H., & Jacot, D. (2011). Further experiments in sentiment analysis of french movie reviews *Advances in Intelligent Web Mastering–3* (pp. 19-28): Springer.

Ghosh, D., Guo, W., & Muresan, S. (2015). *Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words*. In Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing (pp. 1003-1012): Association for Computational Linguistics (ACL).

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR), 49*(2), 28.

Gibbs, R. (2007). Irony in talk among friends. In R. Gibbs & H. Colston (Eds.), *Irony in language and thought: A cognitive science reader* (pp. 339-360). London: Taylor & Francis Group.

Gibbs, R., & Colston, H. (2007). The future of irony studies. In R. Gibbs & H. Colston (Eds.), *Irony in language and thought: A cognitive science reader* (pp. 339-360). London: Taylor & Francis Group.

Gibbs, R. W. (1986). On the psycholinguistics of sarcasm. *Journal of Experimental Psychology: General, 115*(1), 3.

Gibbs, R. W. (2000). Irony in talk among friends. *Metaphor and Symbol, 15*(1-2), 5-27.

Giora, R. (1995). On irony and negation. *Discourse Processes, 19*(2), 239-264. doi:10.1080/01638539509544916

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). *Identifying sarcasm in Twitter: A closer look*. In ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon (pp. 581-586). Portland, Oregon.

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2006). *Feature extraction: Foundations and applications* (Vol. 207). New York: Springer.

Hailong, Z., Wenyan, G., & Bo, J. (2014). *Machine learning and lexicon based methods for sentiment classification: A survey*. In Web Information System and Application Conference (WISA), 2014 11th (pp. 262-265): IEEE.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*(1), 10-18.

Hei, K. C. (2009). Moves in Refusal: How Malaysians say 'No'. *China Media Research, 5*, 31-44.

Indurkhya, N., & Damerau, F. J. (2010). *Handbook of natural language processing* (2nd ed.): CRC Press.

Isa, N., Puteh, M., & Kamarudin, R. (2013). Sentiment Classification of Malay Newspaper Using Immune Network (SCIN). *Proceedings of the World Congress on Engineering, Jul*, 3-5.

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2016). Automatic sarcasm detection: A survey. *arXiv preprint arXiv:1602.03426*.

Joshi, A., Sharma, V., & Bhattacharyya, P. (2015). *Harnessing context incongruity for sarcasm detection*. In ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference (pp. 757-762): Association for Computational Linguistics (ACL).

Jurafsky, D. (2000). Speech and language processing: An introduction to natural language processing. *Computational linguistics, and speech recognition*.

Justo, R., Corcoran, T., Lukin, S. M., Walker, M., & Torres, M. I. (2014). Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems, 69*, 124-133. doi:10.1016/j.knosys.2014.05.021

Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence, 22*(2), 110-125.

Kim, Y., Do Young Kwon, S. R. J., & Jeong, S. R. (2015). Comparing Machine Learning Classifiers for Movie WOM Opinion Mining. *TIIS, 9*(8), 3169-3181.

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research, 50*, 723-762.

Korayem, M., Aljadda, K., & Crandall, D. (2016). Sentiment/subjectivity analysis survey for languages other than English. *Social Network Analysis and Mining, 6*(1), 75.

Kreuz, R. J., & Glucksberg, S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General, 118*(4), 374.

Kreuz, R. J., & Roberts, R. M. (1993). On satire and parody: The importance of being ironic. *Metaphor and Symbolic Activity, 8*(2), 97-109. doi:10.1207/s15327868ms0802_2

Kunneman, F., Liebrecht, C., van Mulken, M., & van den Bosch, A. (2015). Signaling sarcasm: From hyperbole to hashtag. *Information Processing & Management, 51*, 500-509. doi:10.1016/j.ipm.2014.07.006

Lakoff, G., & Johnsen, M. (2003). Metaphors we live by. London: The university of Chicago press. *Prieiga per internetą: http://shu.bg/tadmin/upload/storage/161.pdf [žiūrėta 2012 09 24]*.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Lane, P. C., Clarke, D., & Hender, P. (2012). On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data. *Decision Support Systems, 53*(4), 712-718.

Liebrecht, C. C., Kunneman, F. A., & van den Bosch, A. P. J. (2013). *The perfect solution for detecting sarcasm in tweets# not*. In 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Atlanta, Georgia (pp. 29-37). Atlanta, Georgia: New Brunswick, NJ: ACL.

Liu, B. (2011). *Web data mining: Exploring hyperlinks, contents, and usage data* (Second Edition ed.): Springer Science & Business Media.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies, 5*(1), 1-167.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*: Cambridge University Press.

Liu, H., & Motoda, H. (1998). *Feature extraction, construction and selection: A data mining perspective* (Vol. 453): Springer Science & Business Media.

López, G. J., & Ruiz, I. M. (2016). Character and word baselines systems for irony detection in Spanish short texts. *Procesamiento del Lenguaje Natural, 56*, 41-48.

Lunando, E., & Purwarianti, A. (2013). *Indonesian social media sentiment analysis with sarcasm detection*. In 2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 195-198): IEEE.

Madhoushi, Z., Hamdan, A. R., & Zainudin, S. (2015). *Sentiment analysis techniques in recent works*. In Science and Information Conference (SAI), 2015 (pp. 288-291): IEEE.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1): Cambridge university press Cambridge.

Mansor, N., Ahmad, F., & Yaakub, Y. (2010). Kesantunan bahasa dalam kalangan pelajar IPT: Satu kajian perbandingan etnik.

Maria Alcaide, J., Justo, R., & Ines Torres, M. (2015). *Combining Statistical and Semantic Knowledge for Sarcasm Detection in Online Dialogues*. In Pattern Recognition and Image Analysis (IBPRIA 2015). Heidelberger Platz 3, D-14197 Berlin, Germany (pp. 662-671). Heidelberger Platz 3, D-14197 Berlin, Germany: Springer-Verlag Berlin.

McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*: " O'Reilly Media, Inc.".

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093-1113.

Medhat, W., Yousef, A. H., & Korashy, H. (2014). *A Framework of preparing corpora from Social Network sites for Sentiment Analysis*. In Information Society (i-Society), 2014 International Conference on (pp. 32-39): IEEE.

Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications, 40*(2), 621-633.

Muresan, S., Gonzalez-Ibanez, R., Ghosh, D., & Wacholder, N. (2015). Identification of nonliteral language in social media: A case study on sarcasm. *Journal of the Association for Information Science and Technology*.

Nagwanshi, P., & Veni Madhavan, C. E. (2014). *Sarcasm detection using sentiment and semantic features*. In KDIR 2014 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (pp. 418-424): INSTICC Press.

Newell, A., Potharaju, R., Xiang, L., & Nita-Rotaru, C. (2014). *On the practicality of integrity attacks on document-level sentiment analysis*. In Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop (pp. 83-93): ACM.

Ng, V., Dasgupta, S., & Arifin, S. (2006). *Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews*. In Proceedings of the COLING/ACL on Main conference poster sessions (pp. 611-618): Association for Computational Linguistics.

O'Keefe, T., & Koprinska, I. (2009). *Feature selection and weighting methods in sentiment analysis*. In Proceedings of the 14th Australasian document computing symposium. Sydney (pp. 67-74). Sydney: Citeseer.

Ozdemir, C., & Bergler, S. (2015). *CLaC-SentiPipe: SemEval2015 Subtasks 10 B, E, and Task 11.* Paper presented at the Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).

Pang, B., & Lee, L. (2004). *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts*. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (pp. 271): Association for Computational Linguistics.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up?: sentiment classification using machine learning techniques*. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86): Association for Computational Linguistics.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. *UT Faculty/Researcher Works*.

Pham, S. B. (2014). Sentiment Classification on Polarity Reviews: An Empirical Study Using Rating-based Features. *ACL 2014*, 128.

Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters *Computing attitude and affect in text: Theory and applications* (pp. 1-10): Springer.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*: " O'Reilly Media, Inc.".

Ptácek, T., Habernal, I., & Hong, J. (2014). *Sarcasm detection on Czech and English Twitter*. In COLING 2014 - 25th International Conference on Computational Linguistics. Dublin, Ireland (pp. 213-223). Dublin, Ireland.

Pustejovsky, J., & Stubbs, A. (2012). *Natural language annotation for machine learning*: " O'Reilly Media, Inc.".

Ramteke, A., Malu, A., Bhattacharyya, P., & Nath, J. S. (2013). *Detecting turnarounds in sentiment analysis: Thwarting*. In ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (pp. 860-865): Association for Computational Linguistics (ACL).

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems, 89*, 14-46.

Read, J. (2005). *Using emoticons to reduce dependency in machine learning techniques for sentiment classification*. In Proceedings of the ACL student research workshop (pp. 43-48): Association for Computational Linguistics.

Reyes, A., & Rosso, P. (2012). Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems, 53*, 754-760. doi:10.1016/j.dss.2012.05.027

Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering, 74*, 1-12. doi:10.1016/j.datak.2012.02.005

Riloff, E., Patwardhan, S., & Wiebe, J. (2006). *Feature subsumption for opinion analysis*. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 440-448): Association for Computational Linguistics.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). *Sarcasm as contrast between a positive sentiment and negative situation*. In EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (pp. 704-714): Association for Computational Linguistics (ACL).

Salvetti, F., Lewis, S., & Reichenbach, C. (2004). Automatic opinion polarity classification of movie. *Colorado research in linguistics, 17*(1), 2.

Samsudin, N., Puteh, M., & Hamdan, A. R. (2011). Bess or xbest: Mining the Malaysian online reviews. *Data Mining and Optimization (DMO), 2011 3rd Conference on*.

Samsudin, N., Puteh, M., Hamdan, A. R., & Nazri, M. Z. A. (2013a). Immune based feature selection for opinion mining. *Proceedings of the World Congress on Engineering, 3*, 3-5.

Samsudin, N., Puteh, M., Hamdan, A. R., & Nazri, M. Z. A. (2013b). Mining Opinion in Online Messages. *IJACSA) International Journal of Advanced Computer Science and Applications, 4*.

Selvi, C., Ahuja, C., & Sivasankar, E. (2015). A Comparative Study of Feature Selection and Machine Learning Methods for Sentiment Classification on Movie Data Set *Intelligent Computing and Applications* (pp. 367-379): Springer.

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: a review and comparative analysis of web services. *Information Sciences, 311*, 18-38.

Shany-Ur, T., Poorzand, P., Grossman, S. N., Growdon, M. E., Jang, J. Y., Ketelle, R. S., . . . Rankin, K. P. (2012). Comprehension of insincere communication in neurodegenerative disease: Lies, sarcasm, and theory of mind. *Cortex, 48*(10), 1329-1341. doi:http://dx.doi.org/10.1016/j.cortex.2011.08.003

Sharma, A., & Dey, S. (2012). *A comparative study of feature selection and machine learning techniques for sentiment analysis.* Paper presented at the Proceedings of the 2012 ACM Research in Applied Computation Symposium.

Strapparava, C., & Valitutti, A. (2004). *WordNet Affect: an Affective Extension of WordNet.* Paper presented at the LREC.

Sulis, E., Irazú Hernández Farías, D., Rosso, P., Patti, V., & Ruffo, G. (2016). Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*. doi:http://dx.doi.org/10.1016/j.knosys.2016.05.035

Tepperman, J., Traum, D., & Narayanan, S. (2006). *"Yeah right": Sarcasm recognition for spoken dialogue systems*. In INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP (pp. 1838-1841).

Tsonkov, T. V., & Koychev, I. (2015). *Automatic detection of double meaning in texts from the social networks*. In CEUR Workshop Proceedings (pp. 33-39): CEUR-WS.

Tungthamthiti, P., Shirai, K., & Mohd, M. (2014). *Recognition of Sarcasms in Tweets Based on Concept Level Sentiment Analysis and Supervised Learning Approaches*. In PACLIC (pp. 404-413).

Valitutti, A., & Veale, T. (2015). *Inducing an ironic effect in automated tweets*. In Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on (pp. 153-159): IEEE.

Wallace, B. C., Choe, D. K., & Charniak, E. (2015). *Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment*. In ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference (pp. 1035-1044): Association for Computational Linguistics (ACL).

Wallace, B. C., Choe, D. K., Kertz, L., & Charniak, E. (2014). *Humans require context to infer ironic intent (so computers probably do, too)*. In 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference (pp. 512-516): Association for Computational Linguistics (ACL).

Wang, X., Zhang, C., & Wu, M. (2015). Sentiment classification analysis of chinese microblog network *Complex Networks VI* (pp. 123-129): Springer.

Wang, Z., Wu, Z., Wang, R., & Ren, Y. (2015). Twitter sarcasm detection exploiting a context-based model. *Web Information Systems Engineering-WISE 2015, 9418*, 77-91. doi:10.1007/978-3-319-26190-4

Weiss, S. M., Zhang, T., & Indurkhya, N. (2015). *Fundamentals of predictive text mining* (2nd ed.): Springer London.

Weitzel, L., Prati, R. C., & Aguiar, R. F. (2016). The comprehension of figurative language: What is the influence of irony and sarcasm on NLP techniques? *Sentiment Analysis and Ontology Engineering* (pp. 49-74): Springer.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (Third ed.). Boston: Morgan Kaufmann.

Xia, R., & Zong, C. (2010). *Exploring the use of word relation features for sentiment classification*. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 1336-1344): Association for Computational Linguistics.

Xu, H., Santus, E., Laszlo, A., & Huang, C.-R. (2015). *LLT-PolyU: Identifying Sentiment Intensity in Ironic Tweets*. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Colorado, USA (pp. 673-678). Colorado, USA.

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications, 36*(3), 6527-6535.

Yusof, N. N., Mohamed, A., & Abdul-Rahman, S. (2015). Reviewing Classification Approaches in Sentiment Analysis *Soft Computing in Data Science* (pp. 43-53): Springer.

Zhang, D., Xu, H., Su, Z., & Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and SVM perf. *Expert Systems with Applications, 42*(4), 1857-1863.

APPENDIX A

MALAY STOPWORD LIST

acap	bahawa	jika	minta	semoga
acap kali	bahawasanya	jikalau	misal	semua
adakala	bahkan	jua	mungkin	seperti
adakalanya	barangkali	juga	nampaknya	serba
adalah	beberapa	kadang	namun	serta
adapun	belum	kadangkala	nan	sesungguhnya
aduhai	benar	kalakian	nian	setelah
agak	berapa	kalau	nun	setiap
agaknya	bila	ke	oleh	sewaktu
agar	bilamana	kemudian	pabila	siapa
ah	buat	kenapa	pada	sila
ahai	dalam	kendatipun	padahal	sudah
ai	dan	kepada	paling	sungguh
aja	dapat	kerana	para	sungguhpun
akan	dari	ketika	pelbagai	supaya
alah	darihal	kian	perlu	syahadan
alamak	daripada	lagi	pernah	tatkala
alhasil	demi	lagikan	pun	telah
alkisah	dengan	laksana	saja	tentang
amat	di	laksana	sambil	terhadap
andai	ee	lalu	sampai	terlalu
andai kata	eh	macam	sangat	tetapi
aneka	enggan	maha	sebab	tiap
antara	entah	mahu	sebagai	tolong
apa	entahkan	mahupun	sebagaimana	umpama
apabila	gamaknya	maka	sebermula	umpama
apakala	haah	malah	sedang	untuk
apalagi	hanya	malahan	segala	usah
arkian	harapnya	mana	sejak	wah
atau	harus	manakala	sekali	wahai
au	hatta	manalagi	sekali peristiwa	walau
auh	hendak	masih	sekalian	walaupun
ayuhai	hingga	masing	sekiranya	yang
bagai	ialah	memang	selalu	
bagaimana	ini	mengapa	seluruh	
bagaimanapun	itu	meskipun	semasa	
bagi	jemput	mesti	sementara	

APPENDIX B

ENGLISH STOPWORD LIST

a	each	most	then
about	few	mustn	there
above	for	my	these
after	from	myself	they
again	further	needn	this
against	had	now	those
ain	hadn	o	through
all	has	of	to
am	hasn	off	too
an	have	on	under
and	haven	once	until
any	having	only	up
are	he	or	ve
as	her	other	very
at	here	our	was
be	hers	ours	we
because	herself	ourselves	were
been	him	out	what
before	himself	over	when
being	his	own	where
below	how	re	which
between	i	s	while
both	if	same	who
but	in	shan	whom
by	into	she	why
can	is	should	will
couldn	isn	so	with
d	it	some	won
did	its	such	y
didn	itself	t	you
do	just	than	your
does	ll	that	yours
doesn	m	the	yourself
doing	ma	their	yourselves
don	me	theirs	
down	mightn	them	
during	more	themselves	

APPENDIX C

MALAY INTERJECTION LIST

aduh	hmmm	yahoo
aduhai	hoi	yay
ah	hoorey	zzz
ahh	isk	
aii	kahak	
aik	lah	
aiya	nah	
alahai	oh	
amboi	ooi	
boo	oops	
celaka	pehh	
cih	puii	
cis	puik	
coi	syok	
eh	tuih	
ehh	uwek	
haha	wah	
ha	wahai	
hai	weii	
haish	wow	

APPENDIX D

ENGLISH INTERJECTION LIST

absolutely	aww	gosh	what
achoo	bah	gracious	whiz
ack	bam	hallelujah	woah
agreed	behold	hey	woops
aha	bingo	hi	yes
ahem	blah	huh	yikes
ahoy	bless	indeed	
alack	bravo	jeez	
alas	cheers	no	
alright	crud	now	
alrighty	dang	ouch	
amen	darn	phew	
anyhoo	doh	please	
anyhow	drat	rats	
anytime	duh	shoot	
argh	eek	shucks	
as if	gee	there	
attaboy	geepers	tut	
attagirl	golly	uggh	
awful	goodness	waa	