

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

**MOHD SUHAIRI BIN MD SUHAIMIN**



PERPUSTAKAAN  
UNIVERSITI MALAYSIA SABAH

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

UNIVERSITI MALAYSIA SABAH

**FACULTY OF COMPUTING AND INFORMATICS  
UNIVERSITI MALAYSIA SABAH**

**2017**

**UNIVERSITI MALAYSIA SABAH**

**BORANG PENGESAHAN STATUS TESIS**

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

IJAZAH: **SARJANA SAINS (SAINS KOMPUTER)**

Saya **MOHD SUHAIRI BIN MD SUHAIMIN**, Sesi **2015-2017**, mengaku membenarkan tesis Sarjana ini disimpan di Perpustakaan Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:-

1. Tesis ini adalah hak milik Universiti Malaysia Sabah.
2. Perpustakaan Universiti Malaysia Sabah dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. Sila tandakan ( / ):

SULIT

(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA 1972)

TERHAD

(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/ badan di mana penyelidikan dijalankan)

TIDAK TERHAD



**MOHD SUHAIRI BIN MD SUHAIMIN**  
**MI1511003T**

Tarikh: 09 September 2017

Disahkan oleh,

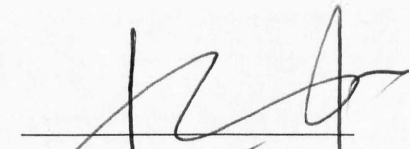
**NURULAIN BINTI ISMAIL**

LIBRARIAN

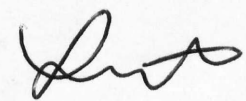
UNIVERSITI MALAYSIA SABAH



(Tandatangan Pustakawan)



(Dr. Mohd Hanafi bin Ahmad Hijazi)  
Penyelia




(Prof. Madya Dr. Rayner Alfred)

Penyelia Bersama

## 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 Suhain bin Md Suhaimin

MI1511003T



UMMS  
UNIVERSITI MALAYSIA SABAH

# 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 : **25 AUGUST 2017**



**CERTIFIED BY**

UNIVERSITI MALAYSIA SABAH

Signature

**1. SUPERVISOR**

Dr. Mohd Hanafi bin Ahmad Hijazi

**2. CO-SUPERVISOR**

Assoc. Prof. Dr. Rayner Alfred

## 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 sarkasme. 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

	Page
<b>DECLARATION</b>	ii
<b>CERTIFICATION</b>	iii
<b>ACKNOWLEDGMENT</b>	iv
<b>ABSTRACT</b>	v
<b>ABSTRAK</b>	vi
<b>TABLE OF CONTENTS</b>	vii
<b>LIST OF TABLES</b>	xii
<b>LIST OF FIGURES</b>	xiv
<b>LIST OF ABBREVIATIONS</b>	xvi
<b>LIST OF SYMBOLS</b>	xviii
<b>LIST OF APPENDICES</b>	xix
<b>CHAPTER 1: INTRODUCTION</b>	1
1.1 Overview	1
1.2 Research Motivation	3
1.3 Research Objective	3
1.4 Research Scope	4
1.5 Research Methodology	4
1.6 Evaluation Criteria	6
1.7 Research Contribution	7
1.8 Published Work	7
1.9 Thesis Organization	8
<b>CHAPTER 2: LITERATURE REVIEW</b>	9
2.1 Introduction	9
2.2 Framework of Sentiment Analysis	9



2.2.1	Overview of Framework	10
2.2.2	Sentiment Analysis Using Supervised Learning Approach for English Language	13
2.2.2.1	Dataset Acquisition and Annotation	13
2.2.2.2	Preprocessing	14
2.2.2.3	Feature Extraction	14
2.2.2.4	Feature Selection	16
2.2.2.5	Classifier Generation and Classification	17
2.2.3	Sentiment Analysis for Non-English Language	20
2.2.4	Sentiment Analysis for Malay Language	21
2.3	Framework of Sarcasm Detection and Classification	23
2.3.1	Sarcasm Definition	23
2.3.2	Overview of Sarcasm Framework	24
2.3.3	Sarcasm Detection and Classification Using Supervised Learning Approach for English Language	25
2.3.3.1	Dataset Acquisition and Annotation	25
2.3.3.2	Preprocessing	27
2.3.3.3	Feature Extraction and Selection	28
2.3.3.4	Classifier Generation and Classification	29
2.3.4	Sarcasm Detection and Classification for Non-English Language	29
2.4	Support Vector Machines	35
2.5	Summary	38
<b>CHAPTER 3: THE DATASET AND PREPROCESSING</b>		<b>40</b>
3.1	Introduction	40
3.2	The Dataset	40
3.3	Data Acquisition	40

3.4	Data Annotation	42
3.5	Data Preprocessing	46
3.5.1	Tokenization	46
3.5.2	Spellchecking	47
3.5.3	Stopword Removal	48
3.6	Summary	49

## **CHAPTER 4: FEATURE FOR SARCASM DETECTION ON BILINGUAL SOCIAL MEDIA DATA**

		50
4.1	Introduction	50
4.2	The Proposed Feature Extraction Process	50
4.2.1	Extraction from Bilingual Dataset	52
4.2.1.1	Lexical Feature Extraction	53
4.2.1.2	Pragmatic Feature Extraction	53
4.2.1.3	Prosodic (Malay) Feature Extraction	54
4.2.2	Extraction of Features from English Translated Dataset	54
4.2.2.1	Dataset Translation to English	55
4.2.2.2	Prosodic (English) Feature Extraction and Combination	55
4.2.2.3	Syntactic Feature Extraction	55
4.2.2.4	Idiosyncratic Features Extraction	56
4.2.3	Feature Selection	57
4.3	Experimental Setup	57
4.3.1	Experiment Objective	57
4.3.2	Parameter Setting	59
4.4	Result and Discussion	60
4.4.1	Evaluation and Comparison	60
4.4.2	Analysis of Result	60

4.5	Summary	66
<b>CHAPTER 5: SENTIMENT ANALYSIS WITH SARCASM DETECTION AND CLASSIFICATION FRAMEWORK</b>		67
5.1	Introduction	67
5.2	The Framework for Sentiment Analysis with Sarcasm Detection and Classification	67
5.2.1	Initial Sentiment Classification	69
5.2.2	Sarcasm Detection and Classification	70
5.2.3	Actual Sentiment Classification	72
5.3	Experimental Setup	74
5.3.1	Experiment Objective	75
5.3.2	Parameter Setting	75
5.3.3	Preprocessing, Feature Extraction and Feature Selection	75
5.4	Result and Discussion	76
5.4.1	Evaluation and Comparison	76
5.4.1.1	Initial Sentiment Classification	76
5.4.1.2	Sarcasm Positivity and Negativity Classification	77
5.4.1.4	Actual Sentiment Classification	77
5.4.2	Analysis of Result	78
5.5	Summary	82
<b>CHAPTER 6: CONCLUSION AND FUTURE WORK</b>		83
6.1	Introduction	83
6.2	Research Summary	83
6.3	Main Finding and Contributions	84
6.4	Future Work	85
<b>REFERENCES</b>		87



**UMS**  
UNIVERSITI MALAYSIA SABAH

# LIST OF TABLES

	Page
Table 2.1: Summary of Previous Work Using Supervised Learning Approach	18
Table 2.2: Summary of Previous Work on Malay Sentiment Analysis Using Supervised Learning Approach	22
Table 2.3: Sarcasm Detection and Classification Using Supervised Learning Approach	31
Table 3.1: Dataset Distribution of Sentiment and Sarcasm Annotation	43
Table 3.2: Dataset Distribution for Sarcasm in Positive and Negative Class (Positivity and Negativity)	43
Table 3.3: Examples of Output Annotated Dataset	45
Table 3.4: Kappa Statistic of Agreement Measures for Categorical Data	45
Table 4.1: Types of Feature Extracted	51
Table 4.2: Examples of Lexical Feature	53
Table 4.3: Examples of Pragmatic Feature	54
Table 4.4: Examples of Prosodic Feature (Malay)	54
Table 4.5: Examples of Prosodic Feature (English)	55
Table 4.6: Examples of Syntactic Feature	56
Table 4.7: Examples of Idiosyncratic Feature	57
Table 4.8: Experimental Combination of Feature	58
Table 4.9: Result of CV Grid Search Test	59
Table 4.10: The Number of Features Used for Experimentation	61
Table 4.11: Sarcasm Detection Performance	62
Table 4.12: Feature Performance Ranking for Experiment Set I	63

Table 4.13: Comparison of Syntactic and Lexical Effectiveness	64
Table 4.14: Examples of Prediction by Set III and Set V	65
Table 5.1: Initial Sentiment Prediction of Comments	70
Table 5.2: Sarcasm Detection on Comments	71
Table 5.3: Sarcasm Classification of Comments	71
Table 5.4: Actual Sentiment Classification of Comments (Flip Both Positive and Negative Sarcastic)	73
Table 5.5: Actual Sentiment Classification of Comments (Flip Positive Sarcastic Only)	74
Table 5.6: The Number of Features Used for Experimentation	76
Table 5.7: Results of Initial Sentiment Classification	77
Table 5.8: Results of Sarcasm Classification	77
Table 5.9: Results of Actual Sentiment Classification	78
Table 5.10: Actual Sentiment Classification Comparison Against Initial Sentiment Classification	79
Table 5.11: Example of Actual Sentiment Classification Over Initial Sentiment Classification	80
Table 5.12: Examples of Actual Sentiment Misclassification	80
Table 5.13: Examples of Misclassification by the Proposed Framework	81

# LIST OF FIGURES

	Page
Figure 1.1: Methodology Phases for Research	5
Figure 2.1: General Framework for SA	11
Figure 2.2: The Linear Non-Separable Case Allowing Data Points Error	36
Figure 2.3: Non-Linear SVM Transformation from Input Space into Feature Space	37
Figure 3.1: Facebook Graph API Explorer GUI	41
Figure 3.2: Comment Characteristic of Malay Social Media Dataset	42
Figure 3.3: Annotation of Sarcasm Positivity and Sarcasm Negativity	44
Figure 3.4: Word Tokenization	46
Figure 3.5: Spellchecking of Tokenized Word	47
Figure 3.6: Returned Word After Stopword Removal	49
Figure 4.1: Feature Extraction Process	52
Figure 4.2: Screenshot of TF-IDF Vectorization and Normalization to Document Length	58
Figure 5.1: The Framework to Support SA Using Sarcasm Detection and Classification	68
Figure 5.2: Initial Sentiment Classification Module	69
Figure 5.3: Sarcasm Detection and Sarcasm Classification of the Sentiment After Initial Sentiment Classification	70
Figure 5.4: Polarity Flip of Both Positive Sarcastic and Negative Sarcastic Based on the First Hypothesis	72





## LIST OF ABBREVIATIONS

<b>AIS</b>	Artificial Immune System
<b>ADJ</b>	Adjective
<b>ADV</b>	Adverb
<b>ANN</b>	Artificial Neural Network
<b>API</b>	Application Programming Interface
<b>BOW</b>	Bag-Of-Words
<b>CHI</b>	Chi-Square
<b>CNB</b>	Complement Naïve Bayes
<b>DF</b>	Document Frequency
<b>DT</b>	Decision Tree
<b>FS-INS</b>	Feature Selection Immune Network System
<b>FQL</b>	Facebook Query Language
<b>GI</b>	Gini Index
<b>GR</b>	Gain Ratio
<b>GUI</b>	Graphical User Interface
<b>IG</b>	Information Gain
<b>IMDB</b>	Internet Movie Database
<b>k-CV</b>	k-fold Cross Validation
<b>k-NN</b>	k-Nearest Neighbors
<b>LIWC</b>	Linguistic Inquiry and Word Count
<b>LogR</b>	Logistic Regression
<b>ME</b>	Maximum Entropy
<b>MI</b>	Mutual Information
<b>MM</b>	Markov Models
<b>MPQA</b>	Multi-Perspective Question Answering
<b>NB</b>	Naïve Bayes

<b>NBM</b>	Naïve Bayes Multinomial
<b>NLP</b>	Natural Language Processing
<b>NLTK</b>	Natural Language Toolkit
<b>OneR</b>	One Rule
<b>PCA</b>	Principal Component Analysis
<b>POS</b>	Part-Of-Speech
<b>RBF</b>	Radial Basis Function
<b>RF</b>	Relief-F
<b>SA</b>	Sentiment Analysis
<b>SMO</b>	Sequential Minimal Optimization
<b>SVM</b>	Support Vector Machines
<b>TF-IDF</b>	Term Frequency - Inverse Document Frequency



UMS  
UNIVERSITI MALAYSIA SABAH

## LIST OF SYMBOLS

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



UMS  
UNIVERSITI MALAYSIA SABAH

# LIST OF APPENDICES

	Page
Appendix A: Malay Stopword List	99
Appendix B: English Stopword List	100
Appendix C: Malay Interjection List	101
Appendix D: English Interjection List	102



**UMS**  
UNIVERSITI MALAYSIA SABAH

# 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.

Previous works proposed the approach on sarcasm detection or sarcasm classification (Bharti, Babu, & Jena, 2015; Lunando & Purwarianti, 2013; Muresan, Gonzalez-Ibanez, Ghosh, & Wacholder, 2015) separately from SA (Medhat, Hassan, & Korashy, 2014; Medhat, Yousef, & Korashy, 2014). 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), factors that influence sarcasm also vary according to geographical area, race and cultural; thus the study and result might be slightly different from the English or others. 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) first proposed by Boser, Guyon, and Vapnik (1992), 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, *et al.*, 2014; Muresan *et al.*, 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. 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.