FEATURE SELECTION METHOD FOR ARABIC TWEETS CLASSIFICATION USING ARTIFICIAL BEE COLONY ALGORITHM

MOHAMMED SABIH ALSHARARI

PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE MASTER IN INFORMATION SCIENCE

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

DECLARATION

I hereby declare that the work in this dissertation is my own except for quotations and summaries which have been duly acknowledged.

07 March 2018

MOHAMMED SABIH ALSHARARI
GP04728

ACKNOWLEDGEMENT

First and foremost praise be to Almighty Allah for all His blessings for giving me patience and good health throughout the duration of this master study.

I am very fortunate to have Associate Professor Dr Mohammed Faidzul Nasrudin as the master project supervisor.

I would like to thank all my colleagues for their help, friendship, and creating a pleasant working environment throughout my years in UKM.

Moreover, I am grateful to my beloved parents, brothers, and sisters; and also to my dearest wife, sons, and daughters for their continuous support during my whole life.

ABSTRACT

This is the era of social websites, and Twitter is considered among the most prevalent social media. Twitter is a micro blogging internet website where a user may share his view in the form of short messages, which are called tweets. In fact, research studies on tweets classification usually face many challenges, such as colloquialism in tweets, spelling variation, use of special characters, violating regular grammar rules, etc. In addition, the short message size does not give enough word occurrences. As a result, many areas of research studies are conducted such as classifying users' tweets into predefined and well-known topics (e.g., sport, politics, economic, and culture) for easier retrieval of information. However, most of the studies are mainly on text of English language. That was due to the lack of substantial resources (like available corpuses) for use in other languages such as Arabic. Moreover, Arabic users use their own local dialect during the writing of their tweets. Furthermore, the morphological complexity of Arabic adds more difficulties to Arabic short text classification subtasks, particularly the feature selection task. However, existing feature selection and machine learning approaches achieve good performances with long documents and with large training data sets are available. Hence, this research proposed a feature selection method based on the artificial bee colony algorithm that are capable of performing well with Twitter short message, the morphological complexity of Arabic, and with the small amount of data. The final experimental results show that SVM classifier (with the average result obtained is 83.85% F-measure) outperformed NB and KNN classifiers (with the average results obtained are 76.79% and 73.76% F-measure, respectively) for Arabic tweets classification on politics, sport, culture and economic domains. Moreover, the highest result yield by the proposed classification model that combines SVM classifier with TF-IDF and the proposed artificial bee colony algorithm is 87.97% F-measure for economic domain, and the lowest result is 79.31% F-measure for culture domain.

ABSTRAK

Sekarang adalah era laman web sosial dan Twitter dianggap antara media sosial yang digunakan secara meluas. Twitter adalah laman web mikro blog yang penggunanya boleh berkongsi pendapat dalam bentuk mesej pendek yang dipanggil tweet. Kajian ke atas pengelasan tweet biasanya menemui banyak cabaran seperti gaya bahasa, kepelbagaian ejaan, penggunaan aksara khas, tidak mengikut tatabahasa dan sebagainya. Selain itu, mesej pendek tidak membenarkan penggunaan perkataan yang banyak. Oleh itu, banyak penyelidikan dijalankan untuk permasalahan ini seperti pengelasan tweet kepada topik tertentu yang diketahui umum (seperti sukan, politik, ekonomi dan kebudayaan) untuk memudahkan dapatan semula maklumat. Namun, kebanyakan penyelidikan adalah untuk teks bahasa Inggeris. Ini kerana kurangnya sumber penting bagi bahasa lain seperti bahasa Arab. Ini penting, pengguna Arab menggunakan dialek mereka semasa menulis tweet. Di samping itu, kekompleksan morfologi bahasa Arab menambahkan kesukaran pengelasan teks pendek terutamanya untuk pemilihan fitur. Oleh itu, kajian ini menilai beberapa model pembelajaran mesin seperti Naïve Bayes (NB), Jiran Utama-Terdekat (JUT) dan Mesin Sokongan Vektor (MSV) untuk mengendalikan pengelasan tweet bahasa Arab. Pengelasan fitur dan model pembelajaran mesin sedia ada mendapat keputusan yang baik untuk dokumen panjang dengan set data latihan yang besar. Kajian ini juga mencadangkan satu pemilihan fitur hibrid iaitu kombinasi algoritma Koloni Lebah Buatan dan kaedah statistik yang berfungsi baik dengan mesej pendek Twitter, kekompleksan bahasa Arab dan set data yang sedikit. Keputusan eksperimen menunjukkan pengelas MSV (dengan purata keputusan F-Measure 83.85%) lebih baik daripada NB dan JUT (dengan keputusan F-Measure masing-masing 76.79% dan 73.76%) untuk pengelasan tweet dalam domain politik, sukan kebudayaan dan ekonomi. Keputusan tertinggi diperoleh oleh model pengelas cadangan yang menggabungkan pengelas MSV dan algoritma Koloni Lebah Buatan dengan F-Measure sebanyak 87.97% untuk domain ekonomi. Sebaliknya, keputusan F-Measure paling buruk adalah 79.31% bagi domain kebudayaan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABC	Artificial Bee Colony algorithm
ATC	Arabic Tweets Corpus
BoW	Bag-of-Words
KNN	Key-Nearest Neighbor
ML	Machine Learning
NB	Naïve Bayes
NN	Neural Networks
SVM	Support Vector Machines
TC	Text Classification
UKM	Universiti Kebangsaan Malaysia
WSD	Word Sense Disambiguation

CHAPTER I

INTRODUCTION

1.1 INTODUCTION

This is the era of social websites, and Twitter considers among the most prevalent social media. Twitter is a microblogging internet website where a user may share his view in the form of a short message, which is called tweets. Every day, millions of tweets are created (Suh et al. 2010). This massive amount of produced data has introduced new tracks of study works, which have both academic and business value. Twitter message classification and Twitter sentiment classification are the two major domains where many researchers are working around the world (Ghiassi, Skinner & Zimbra 2013). Twitter message classification is a process of classifying the messages (tweets) based on the subjects using the words (terms) of the tweets as the feature. While the second domain classification is the process to classify tweets according to the sentiment expressed in them, usually positive or negative, and sometimes positive, negative, or neutral (Kouloumpis, Wilson & Moore 2011; Rosenthal, Farra & Nakov 2017; Verma et al. 2011).

In fact, it is easy to access Twitter data and cover a wide scope of topics. The number of research studies is plenty even though the classification of subjects in Twitter data may create different challenges due to the restricted size of tweets (Rao et al. 2010). The classification criteria may differ according to message content or on user aim.

Researchers utilize certain types of algorithms called classification algorithms in order to classify text in documents. The aim is to classify a targeted document into

one (or more) categories, according to document's content. Typically, humans select carefully these categories (Khan et al. 2010).

Text documents are usually massive and have many contents. Classic methods like Bag-of-Words (BoW) perform well-enough with those types of data sets because the word occurrence is high. While the arrangement of words is lost, word frequency is sufficient to interpret the semantics of the document's content. However, classic techniques usually do not perform well-enough with Twitter messages because those types of tweets are short (Lin, Khade & Li 2012). Since those messages are so short, they do not present enough information about the text itself.

Message created on Twitter is known as a tweet and restricted by Twitter to be equal or shorter than 140 characters. Thus, Twitter users need to write a concise message and that's why they use word shortenings and abbreviations. Interestingly enough, set of abbreviations are surprisingly harmonic across user groups and among other media groups such as such as SMS and chat platforms. Because of Twitter constraint on the message size, users need to express their intentions within the restricted size of tweet message. This makes Twitter a challenging medium to work with (Go, Bhayani & Huang 2009; Mehak 2017). Oftentimes, when Twitter users incapable to express their view using tweet message, they may utilize hyperlinks to redirect readers to other external resources; such as text, audio or video files. Those external resources are known as “Artifacts”.

In fact, research studies on tweets classification usually face many challenges, such as colloquialism in tweets, spelling disparity, utilizing of unusual fonts, breaking systematic syntax rules, etc. In addition, the short message size does not give enough word occurrences (Selvaperumal, & Suruliandi 2014).

On the other hands, supervised learning methods succeed-well to classify text, including tweets classification, and utilized on different domains and particularly for the English language. However, the need for large labeled in-domain data for training is the drawback of such methods (Mandal, & Sen 2014). Since users have different languages and domains, the value of collecting data groups for certain domain in all

languages becomes costly (Indurkhya, & Damerau 2010). Thus, this project aims to design supervised learning models that competent of acting fine with small amount of data and to deal with short text (i.e., tweet message). Moreover, the main contributions of this research as follows:

i. To propose a feature selection method based on the artificial bee colony algorithm.

ii. To propose a new enhanced Arabic tweets classification model based on a feature selection method and machine learning models.

iii. To build an Arabic tweet corpus that contains a number of labelled tweets that are distributed equally between four domains namely political, sport, economic, and culture for Arabic tweets classification.

1.2 RESEARCH MOTIVATION

Nowadays, many users utilizing the web as a method of connection and collaboration. Based on those actions, the data value, which grows, is outside the conception.

Based on the research, Twitter is among the largest community media platforms, it ranks after the Facebook, produces more than three hundreds thousands messages each minute, or more than twenty million per hour (Berners-Lee 2014). The vast amount of data value is considered as the main challenge for designers and developers to introduce applicable proposals for useful data style and handling abilities to utilize data mining.

Hence, numerous fields of research are set up for the meaning analysis in the Internet. Classifying users' tweets into predefined and well-known topics is considered among those fields. It is earning impact since the number of users are growing day by day. Every day millions of Twitter user tweets their views on various topics using short messages of 140 characters length (Go et al. 2009; Mehak 2017). In fact, Twitter offers a list of utmost common subjects users tweet about, which identified as trending subjects in real time. However, it is usually difficult to realize

the main ideas of those trending subjects. Hence, it is vital and essential to categorize those subjects into broad groups with high precision for well information repossession. These general categories can include music, politics, religion, science, sport, technology, movies, and other. The aim here is to help people seeking for information on Twitter to get only slighter subset of trending subjects by categorizing those subjects into broad groups, such as sports, politics, and books, for better searching of information.

Until now, the studies is typically focused on English language. However, web purpose is likewise extended amongst other languages speakers. In addition, since the varied choice of languages and possible domains, the value of collecting data groups for every target domain in all languages are excessively costly. Consequently, this research aim to design supervised learning models that could able of accomplishment better with small amount of data. Moreover, there are only few research studies that are implemented to evaluate classification of Arabic tweets in the Arabic language (Abdul-Mageed, Diab & Kübler 2014; Abdulla, Al-ayyoub & Al-Kabi 2014; S. Ahmed, Pasquier & Qadah 2013; Al-Ayyoub, Essa & Alsmadi 2015; Al-Subaihin, Al-Khalifa & Al-Salman 2011). In addition to the other usages, the Arabic language is very well represented in Twitter; there were about 11.1 million active Arabic users on Twitter as of the beginning of 2017. There are about 22 Arab countries and millions of people who understand Arabic, since it is the language of the holy Quran. Despite extensive research, the researcher could not find considerable work published for Arabic tweets classification.

1.3 PROBLEM OF RESEARCH

Classification of Twitter messages (tweets) is a very challenging task as they are short messages since they are restricted to be within 140 characters by Twitter (Go et al. 2009; Mehak 2017). Short messages does not have enough content while a group of manuscript tends to cover a varied kind of words. This makes it really difficult to build a feature space directly for texts classification (Rao et al. 2010). Scale short text classification is poor due to the data sparseness (Cheng et al. 2014). Traditional methods do not perform as well as they do performed on large texts due to the data

sparsity problem (Jin et al. 2011). Short Twitter messages (tweets) make these issues even more serious, due to the use of rich set of abbreviation, spelling mistakes and tweets can be prone to syntactic errors (Eryiğit, & Torunoğlu-Selamet 2017). There are few researches on supervised learning methods for classification of Twitter messages in the past years and have achieved considerable success. However, most of the studies are mainly on text of English language (Go et al. 2009). That was due to the shortage of enough resources in other languages such as Arabic (Abdul-Mageed, Diab & Kübler 2014; S. Ahmed, Pasquier & Qadah 2013; Al-Ayyoub, Essa & Alsmadi 2015; Al-Subaihin, Al-Khalifa & Al-Salman 2011). Besides, Arabic users have their own native dialog for scripting their messages. Because the existing of numerous Arabic dialects that vary in meaning, the precision in meaning could not be realized. Furthermore, the morphological complexity of Arabic adds more difficulties to Arabic short text classification subtasks particularly, feature selection task (Darwish, Magdy & Mourad 2012). All of these problems and the restriction of accessible language resources in Arabic aid to encourage more studies in this field. Earlier research such as (Mesleh 2011) on feature sub-set selection for Arabic text only focuses on filter-based feature selection. However, many useful techniques like wrapper-based feature selection have not been explored widely.

1.4 RESEARCH QUESTIONS

The following describes the research questions of this study:

Q 1: How are the performance of existing classification methods for Arabic tweets classification?

Q 2: Which features to be selected? How to enrich the features and to build an effective feature selection method?

Q 3: How do a feature selection method affects the performance of the Arabic tweets classification?

1.5 OBJECTIVES

The objective of this project are illustrated as follows:

i. To evaluate several machine learning models namely Naïve Bayes, Key-Nearest Neighbor, and Support Vector Machines classifiers for handling the problem of Arabic tweets classification.

ii. To propose a feature selection method based on the artificial bee colony algorithm.

iii. To propose a new enhanced Arabic tweets classification model based on a feature selection method and machine learning techniques, and evaluate the proposed model.

1.6 METHODOLOGY OF RESEARCH

This project will study the problem of Arabic tweets classification. To deal with this issue, the study will propose a new model that will contribute to handle this problem.

Phase 1: Literature Review and Identifying the Problem

- **State of the art**

This step will present the main relevant approaches that is considered as being closely related to this research. Additionally, to come out with state-of-the-art for the methods used to deal with this problem. Furthermore, to have information about the results have been achieved and what is the methods that outperform others.

- **Identify the problem**

This step is to identify and determine the research problem based on state-of-the-art that has been done in the first step in this phase.

- **Determine objectives and hypothesis**

The objectives and hypothesis of this research will be obtained and determined in this step. The proposal of this research including research objectives, hypothesis, and problem statement will be one of this phase output. Moreover, this phase will offer the knowledge about the techniques and methods have been used to solve this problem.

Phase 2: Designing and Proposing the Model

- **Designing and testing initial model**

In this step, initial model will be designed to test the baseline methods performance. Furthermore, the most related methods and approaches will be applied and tested to come out with conclusion of their performance and ability to adopt research problem.

  - Text Representation
  - Baseline Classifiers

- **Designing the extended model**

Unlike all proposed methods so far, this step aims at taking advantage of different view levels to learn models of Arabic tweets classification. This work aims on developing enhanced Arabic tweets classification model based on:

i. An enhanced method which combines a feature selection method and a machine learning model.

ii. Word distributed representation to enrich the representation of document and solve the data sparsity problem.

1.7 OUTLINE OF THE THESIS

This thesis is divided into five chapters. It commences with this introductory chapter and then followed by these chapters:

Chapter II: Literature Review - Emphasis on reviewing and analysing related literature. First, it covers the concepts and the state-of-the-art in tweets classification. Second, it deals with the understanding of the challenges and problems of Arabic tweets classification.

Chapter III: The Methodology - Introduce the proposed extended techniques that would be applied in this project.

Chapter IV: Experimental Results - The discussion of experimental results of the proposed extended techniques.

Chapter V: Conclusion and Future Work - This chapter presents the summary of this research. It describes the contribution and the suggestions for future work of this research.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter presents a brief review of the literature on the application of text classification. Firstly, it briefly presents different techniques for Arabic language as well as its challenges. Then, it discusses text classification and its types, techniques, and application. This chapter also offers a brief section about feature selection methods. Then, this chapter provides a summary about Arabic text classification, and presents available Arabic text corpora used in previous studies. Finally, the last section concludes this chapter.

2.2 ARABIC LANGUAGE

Arabic is the formal language of more than twenty countries, and more than three hundred million people speak Arabic. It is a Semitic language and considers among the utmost broadly used languages. It has a composite and many morphology than the English language, and thus it is a very transformed language (Al-Harbi, Almuhareb & Al-Thubaity 2008; M. K. Saad, & Ashour 2010).

Arabic characters appear in different styles according to their position within the Arabic words, either at the start, middle or tail of a word and on whether the character can be connected to its neighbor characters or not (Duwairi 2007). For instance, the styles of character (س) as follows: at the beginning must be written as (سـ), at the middle must be written as (ـسـ), and at the tail of the word must be written as (ـس).

Diacritics is considered as one of the most distinctive features of the Arabic language. They are signals usually placed below or above the written characters. Table 2.1 lists the common diacritics used in Arabic language (Hammo 2009). They contain shada, dama, fatha, kasra, sukon, double dama, double fatha, and double kasra. The masculine name differs from feminine in Arabic. Arabic words have three types of numbers include singular, dual, and plural, and have three grammatical types: nominative, accusative, and genitive. The first type when the noun is subject of the sentence, the accusative case when the noun is the object of a verb, and the genitive case if the noun is the object of a preposition. There are three types of Arabic words: nouns, verbs, and particles. All verbs and some nouns are originated from a group of roots.

Table 2.1 Popular diacritics utilized in Arabic manuscript

Diacritic	Shape description	Example	Sound
fatha	A small diagonal line appears above a letter	دَ	**da**
kasra	A small diagonal line appears below a letter	دِ	**di**
damma	A small "comma-like" diacritic placed above a letter	دُ	**du**
tanwin	A double vowel diacritic appears only at the end	دً دٍ دٌ	**dan**, **din**, **dun**
sukun	A small circle shape above the letter indicating that the consonant is not followed by a vowel	دْ	d
shadda	A small diacritic (ّ) indicating a doubled consonant	دّ	dd
madda	A diacritic appears on top of alif indicating a long alif	آ	aa

Source: (Hammo 2009)

Usually, in classical poetry and kid's literature, writers use diacritics if that manuscript is unclear to speak. For example in Table 2.2, the word (كتب), which means "write" in English and consists of three letters, can have different interpretations while using diacritics (Hammo 2009).

Table 2.2 Diverse reading of the Arabic word كتب (katab) while using the diacritics

Arabic word	Transliteration	Part of speech	English meaning
كَتَبَ	*kataba*	3PSNG (verb)	Wrote
كُتُب	*kutub*	Noun	Books
كُتِبَ	*kutiba*	Passive (verb)	Written
كَتَّبَ	*kattaba*	Verb	Make someone to write

Source: (Hammo 2009)

2.3 ARABIC LANGUAGE CHALLENGES

Researchers identified a number of challenges to study the Arabic language. Those challenges includes the following (Ayedh et al. 2016; M. K. Saad, & Ashour 2010):

1. Some set of Arabic letters can be written in various forms. For example, occasionally in using Alhamza with Alalef character (أ), sometimes Alhamza is not written and it appears like this (ا). This results the letter unclear as if the "hamza" is there or not.

2. Compare to the English language, the Arabic language has a very complex morphology. For instance, to express the possessive things using the Arabic language, the word ends with character (ي). Arabic does not have certain term like "my" in englsih for that purpose.

3. Arabic words are extracted: Arabic words are typically extrcted from their roots. Word root is a plain abstract verb form, which comprises of three characters. Many of times, searching for the root of a certain derived word could be difficult task.

4. Arabic language has much of broken plurals, similar English plurals. However, they frequently are not like the singular form as closely as irregular plurals resemble the singular in English. Usually, such Arabic plurals do not follow ordinary morphological rules.

5. Arabic language has short vowels that give diverse pronunciation. They are essential but misplaced in printed Arabic manuscripts.

6. Synonyms of Arabic words are common and thus Arabic language considered as one of the richest languages. Therefore, retrieving or classifying Arabic texts by utilizing exact keyword match is considered inappropriate.

7. Many Arabic words have different meanings and the suitable meaning could be specified based on its presence in the paragraph.

8. In Arabic language, there are no capital or small letters like the English Language have. This makes distinguishing proper names, acronyms, and abbreviations is difficult.

9. The number of free Arabic datasets (especially large-scale Arabic datasets) that suitable to classify Arabic documents is very limited.

10. A number of Arabic words cannot drive their roots because they have been taken from other languages, such as (برنامج) that means program, and (انترنت) that means internet in English.

From all the above challenges, it is obvious that the Arabic language has various and hard strict structure. These variations make it difficult for language processing techniques, which proposed for other languages, to be directly applied to the Arabic Language (Ababneh et al. 2012). Hence, for conducting a required manipulation, the study needs a set of preprocessing tasks such as normalization and stemming that considered hard to maintain.

2.4 TEXT MINING

Currently, text is the utmost common and useful approach to share information. The significance of this approach has motivate a number of studies to propose suitable methods for study natural language texts to elicit beneficial information. Researchers define text mining as the procedure or task of discovering significant and interesting linguistic patterns from natural language texts (Hotho, Nürnberger & Paaß 2005; Patel, & Soni 2012). More precisely, it is utilizing algorithms and approaches from machine learning and statistics to natural language texts to elicit significant information for further use (Allahyari et al. 2017).

In general, data mining is an automatic procedure of realizing beneficial and helpful patterns in a huge volume of data. In contrast to data saved in organized layout (such as databases), data saved in text files is unstructured and hard to manage and use. To utilize unstructured data, a preprocessing is essential to convert textual data into a proper layout for manipulating (Hotho et al. 2005).

In data mining, information or possibly valuable patterns are typically concealed and anonymous, so automatic procedures are essential for enabling the extraction of targeted data. In text mining, the information is obvious in the texts; however, the challenge is that this targeted information is not appeared in an approach proper for processing by a computer. The main objective of text mining field is to demonstrate data saved in texts into a shape proper for automatic processing. Studies in the text mining field comprises dealing with issues like text summarization, text classification, document clustering, text representation, and information retrieval. Those issues could be summarized as follows (Abbas, Zhang & Khan 2014; El-Halees 2007; Gray, & Debreceny 2014; Hotho et al. 2005; Khan et al. 2010; Witten 2005):

- Text summarization is an automatic discovery of the utmost significant expressions in a specified text file and produce a summarized form of the input text to be used by targeted users. It could be conducted for a one file or a collection of files. Most methods in this field emphasis on getting informative clauses from texts and structuring abstracts established from the acquired information.

- The ides of text classification is to assign text files into single or other pre-identified classes according to their content. Text classification is a supervised learning method where the classes are recognized in advance. There are many widely known methods for this problem, such as NB, KNN, and SVM algorithm.

- Document grouping is a machine learning technique to classify the match between text files according to their content. The difference between this method and text classification is that the document grouping is an unsupervised process where there are no pre-identified classes. The idea of this

technique is to generate relations between related files in a file group in order to retrieve them together.

- Text representation is dealing with the issue of how to demonstrate text in proper layout for automatic handling. In fact, text files could be demonstrated in two methods. This first method as a Bag-of-Words where the context and the text arrangement are ignored. While the second method is to discover shared expressions in text and consider them as a particular expression.

- For dealing with the last mentioned issue (i.e., the information retrieval), the information required to be extracted is denoted as request and the function of the information retrieval systems is to discover and provide documents that hold the utmost related information to the specified request. Hence, text mining methods are utilized to process text data and match the recovered information and the given requests to discover the documents that contain responses.

2.5 TEXT CLASSIFICATION (TC)

Due to the fast development of natural language text files existing in automated format, automatic text classification (TC) becomes a significant tool, which could be utilized to execute the task of classifying this data. It is the task of classifying natural language texts to single or more pre-identified classes according to their content. Before using this approach, users were using other approach like knowledge engineering, which is based on set of rules defined by human to determine how to categorize text files into given classes. However, machine learning methods became the utmost used method for the TC issue since the early 1990s. For dealing with TC issue, machine learning could be defined as a general inductive process, which makes a text classifier by learning the structures of the text classes from a group of pre-categorized text files. The difference between this approach and knowledge engineering approach is that machine learning classifier is automatic and does not require the designers to prepare the definition manually (Sebastiani 2002).

The core three phases of TC are text preprocessing, feature selection, and building the classification model on training data. In the first phase (i.e., preprocessing phase), a tokenization process is accomplished to eliminate non-informative terms such as punctuation symbols and numbers. In addition, terms that occur commonly and do not afford informative data, such as stop terms and rare terms, that assists in recognition between text classes are often removed. A group of the utmost informative terms is reserved and then utilized to characterize text file as vectors of features. Then, feature selection is executed. The aim of this process is to increase the classification precision and computational effectiveness by eliminating unrelated and noisy features that not contain adequate information to help the text classification process. Finally, building a classification model on training data using the best subset of features designated by the second process, and then assessing its accomplishment on a single test data. Figure 2.1 illustrates the general steps of the TC process.

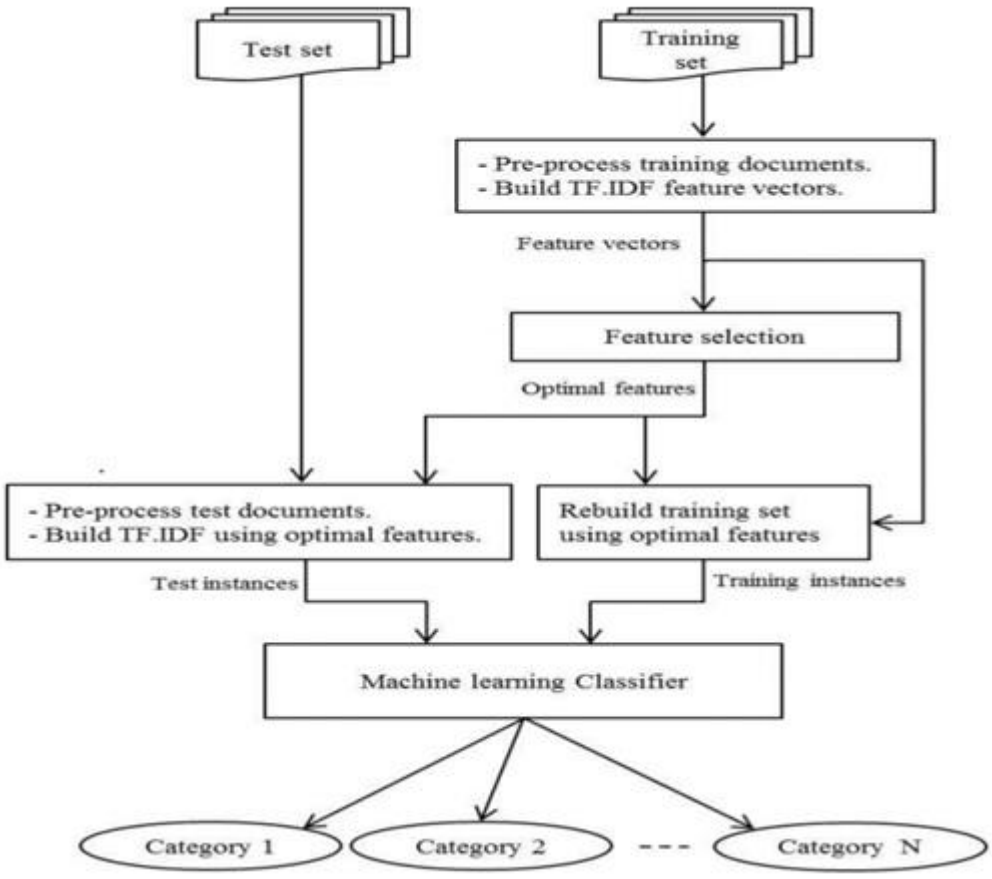

Figure 2.1 Text classification process

Source: (Chantar 2013)

For instance, a research discovered the categories of triggers that stimulus chats on Twitter (Zubiaga et al. 2015). Specifically, they have discovered the highest chats shown as trending subjects on Twitter. They have presented a classification method to unify the site's trending subjects by the category of event that triggered them. They have presented fifteen features that not depending of the used language to characterize trending subjects. They have utilized SVM classifiers to examine the effectiveness of these features to distinguish categories of trending subjects.

In (Dai, & Bikdash 2015), the study introduces a cross classification technique aiming on a small domain that could differentiate Twitter messages representing flu infection from those messages simply linked with flu, but unrelated to flu infection. It expands the classification procedure since it utilizes benefit of the several methods. Their outcomes illustrate that the cross classification techniques accomplished better outcomes than any single technique. The main drawback of this approach is that it is difficult to adapt this classifier as to make it appropriate to alternative illness or to alternative query.

Another research was directed to assess the usage of hidden subjects as a text demonstration method for multi-label boosting algorithms (Al-Salemi, Ab Aziz & Noah 2015). The subject model was utilized to get the hidden subjects is latent Dirichlet allocations with Gibbs sampling. Latent subject demonstration assessed in contrast with the classical BoW text demonstration and multi-label instance-based algorithms was utilized to validate the boosting-based algorithms achievement assessment.

Furthermore, some research define two types of text classification: flat and hierarchal text classification. Where the subtitles are ignored in flat classification. If the number of files in the class is huge, the exploration over such class becomes hard and the classification precision of the text classification method that utilizes this data could be influenced. Thus, researchers prose to utilize the hieratical classification where the connection between files could be utilized by separating every central class