AN IMPROVED K-NEAREST NEIGHBORS APPROACH USING MODIFIED TERM WEIGHTING AND SIMILARITY COEFFICIENT FOR TEXT CLASSIFICATION

by

AMMAR ISMAEL KADHIM

Thesis submitted in fulfillment of the requirements
for the degree of
Doctor of Philosophy

March 2016

ACKNOWLEDGEMENTS

In the name of Allah the most gracious the merciful. First and foremost, Praise be to Allah, Lord of the worlds; and prayers and peace be upon the master of messengers, the Prophet Mohammed (PBUH), our leader in this life until "the end".

I would like to thank Allah for granting me health and patience to finish this research. I would also like to express my sincere gratitude to my main supervisor, Associate Prof. Dr. Cheah Yu-N for his valuable guidance and support throughout these years of study. Without his comments and continuous encouragement, this dissertation would not have been possible. I thank him for having his door open every time I needed help. I would also like to thank my co-supervisor Dr. Nurul Hashimah Ahamed Hassain Malim for her scholarly guidance and support throughout this study.

I also owe a huge debt of gratitude to my mother for her prayers for me during this important time. My heart felt gratitude also goes to my family members: to my wife and my children**:** Mareem, Adraa, Karar and Hia. I would like to thank my parents, my brothers and sisters for their continuous prayers, and endless support when I needed it. Finally, many thanks to all my friends and colleagues who supported and helped me at the School of Computer Sciences, Universiti Sains Malaysia.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AI	Artificial intelligence
API	Application Programming Interfaces
BOW	Bag-of-words
DAN	Dynamic artificial neural
DM	Direct message
DR	Dimension reduction
FE	Feature extraction
FS	Feature selection
HTML	Hyper text markup language
IDF	Inverse document frequency
IR	Information retrieval
K-NN	k-Nearest neighbors
LDR	Local dimension reduction
LM	Language Model
LSI	Latent semantic indexing
MI	Mutual information
ML	Machine learning
MTML	Multi-task multi-label
NLP	Natural language processing
NN	Neural Networks
PCA	Principal component analysis
PMI	Pointwise mutual information
SDR	Sufficient dimension reduction

SGML	Standard Generalized Markup Language
SML	Supervised machine learning
SOM	Self-organizing maps
SSDR	Semi-supervised dimension reduction
STW	Supervised term weighting
SVD	Singular value decomposition
SVM	Support vector machine
TC	Text classification
TF	Term frequency

PENDEKATAN K-JIRAN TERDEKAT DIPERBAIK MENGGUNAKAN PEWAJARAN KATA DAN PEKALI KESERUPAAN TERUBAH SUAI UNTUK PENGELASAN TEKS

ABSTRAK

Pengelasan teks automatik adalah penting kerana peningkatan bilangan dokumen digital dan oleh itu ia perlu diurus. Kaedah pemodelan statistik terkini tidak memberi maklumat berguna yang mencukupi tentang topik untuk setiap ciri dan kategori. Tambahan pula, penyarian sifat menggunakan frekuensi kata-frekuensi dokumen songsang (TF-IDF) tradisional menghasilkan pengenalan kategori yang terlalu banyak untuk sesuatu dokumen. Dalam usaha pengelasan pula, kaedah k-jiran terdekat (k-NN) sedia ada dengan jarak Euclid dan skor keserupaan kosinus menghasilkan julat varians yang besar dalam prestasinya. Untuk menangani isu ini, kajian ini mengelaskan topik untuk teks pendek dan panjang dengan menggunakan pendekatan baharu untuk tahap-tahap utama pengelasan teks (iaitu penyarian sifat dan pengelasan teks). Kajian ini juga memperkenalkan TD-IDF dengan logaritma dan k-NN dengan skor keserupaan kosinus yang baharu untuk penyarian sifat dan pengelasan masing-masing. Lagipun, faktor yang memberi kesan terhadap prestasi pembelajaran mesin berselia juga dikenalpasti. Untuk teks pendek, tiga saiz set data yang berbeza dipungut menggunakan antara muka pengaturcaraan aplikasi (API) (iaitu setiap satu dengan 2,196; 5,534; dan 10,186 *tweet*). Untuk teks panjang, pungutan ujian Reuters-21578 digunakan. Eksperimen menunjukkan bahawa TF-IDF dengan logaritma menambahbaik prestasi penyarian sifat dengan purata ukuran F1 (*F1-measure*) 92.36%, 93.04%, dan 93.60% untuk set data 2,196, 5,534, dan 10,186 *tweet* masing-masing untuk teks pendek dan 92.53% untuk teks panjang. Untuk

pengurangan dimensi (DR), empat kes berbeza digunakan untuk setiap set data teks pendek dan panjang. Kemudian, untuk pengelasan teks, pendekatan k-NN dengan skor keserupaan kosinus baharu (*k-NN-CSNew*) yang dicadangkan menunjukkan prestasi yang lebih baik berbanding k-NN dengan jarak Euclid (*k-NN-ED)* dan k-NN dengan skor keserupaan kosinus tradisional (*k-NN-CSOld*) berdasarkan bilangan jiran k yang berbeza.

AN IMPROVED K-NEAREST NEIGHBORS APPROACH USING MODIFIED TERM WEIGHTING AND SIMILARITY COEFFICIENT FOR TEXT CLASSIFICATION

ABSTRACT

Automatic text classification is important because of the increased availability of digital documents and therefore the need to organize them. The current state-of-the-art statistical modeling approaches do not provide sufficient useful information on the topics for each feature and category. Furthermore, feature extraction using traditional term frequency-inverse document frequency (TF-IDF) results in the identification of too many categories for a particular document. In terms of classification, current k-NN approaches with Euclidean distance and cosine similarity score produce a wide range of variance in performance. To address these issues, this study classifies topics for short and long texts using a new method for the main stage (i.e., feature extraction and text classification). The study also introduces TF-IDF with logarithm and k-NN with a new cosine similarity score for feature extraction and classification, respectively. Moreover, the factors that affect the performance of supervised machine learning (ML) are also identified. For short texts, three different dataset sizes are collected using API (i.e., each with 2,196; 5,534; and 10,186 tweets). For long texts, the Reuters-21578 test collection is used. The experiments show that TF-IDF with logarithm improves the performance of feature extraction with an average F1-measure of 92.36%, 93.04%, and 93.60% for the 2,196-; 5,534-; and 10,186-tweet datasets, respectively, for short text, and 92.53%, for long texts. For dimension reduction (DR), four different cases are applied for each dataset in short and long texts. Subsequently, for text classification, the

proposed k-NN approach with new cosine similarity score (k-NN-CSNew) outperforms k-NN with Euclidian distance (k-NN-ED) and k-NN with traditional cosine similarity score (k-NN-CSOld) based on different k number of neighbors.

CHAPTER 1

INTRODUCTION

1.1 Overview

Automatic text classification, which is a crucial task in information management applications, involves the automatic assignment of a given text to one or more predefined categories. This particular task can be performed in many information management applications, such as an indexing mechanism for text retrieval and a component of information filtering. Moreover, text classification (TC) helps users capture their areas of interest, thereby allowing them to easily filter out documents that are not relevant to their interest by automatically grouping the documents based on their contents. These groups or topics can then be used to improve certain tasks, such as obtaining search results, or can serve as a means of improving user experience in exploring the underlying document dataset. TC is the task in which the topics are classified into one or more predefined categories based on their contents Sadiq and Abdullah (2012). The supervised machine learning (SML) technique has been utilized to generate the automatic topic classifier with the training set of documents. In other words, the automatic classification of electronic documents is realized through the Internet and other channels, such as news reports and articles.

The increasing development of multimedia technologies, storage capacity, and computational power, together with the growth and diversification of telecommunication technologies, has allowed us to receive any kind of information supported by any forms of media (Macdonald & Ounis, 2006) and therefore constitutes further motivation for effective TC.

Rule-based approaches for TC fall into two basic kinds. The first rule-based approach represents the classification rules, which are usually generated manually by experts in the domain of texts. Although this rule-based approach can achieve high accuracy, it is costly in terms of labor and time. The second approach includes ML techniques in which the classification rules are automatically produced using the information from labeled (i.e., already categorized) texts. SML saves cost because it requires only the labeled texts (Pappuswamy, Bhembe, Jordan, & VanLehn, 2005).

Automatic TC, such as that for short texts (e.g., Twitter) and long texts (e.g., Reuters-21578), is crucial for users of social networking sites because of the increased availability of documents in digital structures that must be organized on these sites. Many users participate in social awareness streams, including microblogging services and social networks, to post and spread information throughout the network.

Twitter is among the extremely popular online social networking sites and microblogging services that enable users to share short messages and communicate public opinion about events in real time, which are worthy of extensive public attention. These messages collectively specify the interests and attention of local and global communities and particularly form the temporal trends on Twitter. Numerous events and topics are discussed on Twitter. Some of these topics receive substantial attention, thereby becoming a trend, whereas others do not. Detecting these trends in online social networking sites has become an important problem that has drawn the attention of both the industry and the research community over the past few years.

The Reuters-21578 test collection has been one of the datasets that are most widely used as a standard benchmark for the TC technique over the last 10 years. Nevertheless, given that different researchers have derived various subsets out of this

collection and tested their categories on one of these subsets only, categories that have been tested on different Reuters-21578 subsets are therefore not readily comparable (Debole & Sebastiani, 2005).

Thus, specific preprocessing methods and techniques are required to extract useful information effectively (e.g., word frequency). Text mining is generally defined as the process of extracting interesting and non-trivial features and knowledge from unstructured text documents (Sadiq & Abdullah, 2012).

Considerable research focuses on the analysis of textual media, which is obviously more prominent than other methods at a time. Accordingly, substantial research aims to classify the text documents or messages posted as short text, such as those on Twitter (Sadiq & Abdullah, 2012).

Computational approaches to TC may be extremely useful in different fields, such as the analysis of public opinion among other users. A relevant class is represented by approaches for classification. In the same context, certain techniques have been used in a particular class of probabilistic processes that are called topic models. Such models are imported from the text analysis area as workhorses in several scientific fields. In particular, topic models are generated models that are regarded as the basic idea to describe a document as a mixture of various topics (Steyvers & Griffiths, 2007).

1.2 Automatic Text Classification

The general problem of TC can be further divided into multiple sub-problems, such as subject classification, sentiment classification, functional classification, and other types of classification (Bijalwan, Kumar, Kumari, &

Pascual, 2014). However, this study focuses on subject classification, or simply, the classification of texts into different topics or categories.

Automatic TC is a supervised learning problem in which a set of labeled text documents is used to train a classifier. The classifier is then employed to assign one or more predefined category labels to future trending topics on Twitter. The topics can be classified into various domains through different approaches.

One approach is to employ a supervised classification technique using a set of pre-classed text documents provided as a training model. TC has a crucial role in various areas, such as information retrieval, word sense disambiguation, and Web page classification, as well as in any application that requires text document organization (Wanas, Said, Hegazy, & Darwish, 2006). TC is also used to identify texts on the same topic. However, one of the major problems in using this kind of application is the exceedingly difficult classification and specification of texts for the same topic (Fiaidhi, Mohammed, et al., 2013b).

For short texts, social network media (e.g., Twitter) are frequently used, whereas for long texts, texts from the Internet (e.g., news stories) are mostly used. In the case of short texts, Twitter allows users to write or post anything in a spontaneous manner. Twitter specifies the trending topics, and the site automatically provides an algorithm that attempts to detect currently trending topics. Thus, the trend list is designed to aid people in discovering the "most current happenings" (i.e., news) in real time across the world. This list not only captures the most popular topics but also the new and emerging ones. Several studies have produced an estimate of trending topics for classification (Fiaidhi, Mohammed, et al., 2013b). Only 140 characters are allowed for each tweet because of the high volume of tweets posted daily on Twitter.

For long texts, Reuters-21578 consists of a set of 21,578 news stories that emerged in the Reuters newswire in 1987, which are classified with respect to 135 topic categories, which are mostly related to business and economy (Debole & Sebastiani, 2005).

1.3 Motivations

The motivations of this study can be specified into three: document organization, format types, and classification.

Given the rapid development of information on the Internet, people search for topics regarding some emerging events by adopting social media tools (e.g., Twitter) or by reading Internet articles. Some research groups have successfully developed TC methods based on the content of data streams. These groups, however, do not provide information on the commonly discussed topics for selection. As a result, the development of automated TC systems has become a serious issue in organizing documents.

Social media provides people with a platform to express their viewpoints and to share messages over the Internet. People write comments about what they hear, post about their activities and plans for the future, or contribute their expertise and opinions on various topics. However, Twitter and Reuters-21578 use formats that are different from those of microblogging sites, blogs, and other social networking sites. The format of the other sites includes different types of messages that satisfy the various needs of users.

For classification, users require only few labeled documents as much as possible because labeling documents by hand is remarkably expensive and time

consuming. The current research classifies the documents that are required to analyze topics in public opinion.

Numerous techniques (e.g., SML) are used for TC depending on the situation. These techniques predefine the target category labels to unlabeled documents.

1.4 Statement of Problem

The study investigates how differences between two documents pose a significant challenge in TC, considering that a word belonging to a particular category may contain keywords that may belong to another category. The document files that are not identified with the user also generate problems such as the following:

1. The difficulty of statistical modeling involves obtaining useful information on a topic for each category because current approaches use only either bag-of-words (BOW), word co-occurrence, or mutual information (MI).
2. The difficulty in determining the topic category using the traditional TF-IDF method (i.e., without logarithm) results in the identification of too many categories for a particular document.
3. The traditional k-nearest neighbor (k-NN) approaches using Euclidean distance and cosine similarity score produce a wide range of variance in the performance evaluation of possible category results and are therefore less accurate than a new cosine similarity.

1.5 Objectives of the Research

This study primarily aims to classify topics for short and long texts by extracting and analyzing the content of document features from the text. The specific objectives of this study are as follows:

1. To enhance each stage of the TC process (i.e., feature extraction, Dimension reduction, and supervised machine learning).
2. To improve the performance of TC using TF-IDF with logarithm and k-NN with a new cosine similarity score.
3. To identify factors that affect the performance of supervised machine learning such as the training dataset sizes and number of features.

1.6 Scope of Research

A TC model with good performance can be developed based on the content of documents and the data stream through a well-specified domain. In this study, only documents in English are examined given that English is the language normally used on Twitter and Reuters-21578. The research on TC also focuses on short texts (Twitter) and long texts (Reuters-21578). The short and long TC area process involves assigning different input texts to one or more predefined categories based on their contents. The characteristics considered for the Twitter and Reuters-21578 datasets are as follows:

- The Twitter data collection is used to analyze short texts, which are limited to a few words only (i.e., 20 words) because each tweet consists of not more than 140 characters. The dataset collected consists of only 12 general categories (e.g., politics, education, health, marketing, music, news and media, recreation and sports, computers and technology, pets, food, family, and others) for short texts.

- Reuters-21578 is used to classify long texts that comprise the 10 largest classes (e.g., earnings, acquisitions, money-fx, grain, crude, trade, interest, ship, wheat, and corn).

1.7 Contributions

This study makes the following original contributions:

1. For each category, the BOW, word co-occurrence, and MI (i.e., all of these) are used in feature extraction, instead of only one of these three, as in related work.
2. The feature for each category is extracted by performing weighting based on log (TF-IDF).
3. The K-NN technique with a new cosine similarity score is adopted to automatically classify documents into one or more categories.

1.8 Thesis Outline

The present study consists of the following chapters:

Chapter 2 introduces Twitter and Reuters-21578. A backgrounder on topic discovery and TC is also presented. The background of TC and some definitions of related terms are also provided. The chapter also compares single-label and multi-label TC. Finally, it introduces advancements in k-NN in two subsections: feature extraction and DR, which are used in classifying and summarizing this chapter.

Chapter 3 describes the methodology and techniques used in this thesis.

Chapter 4 presents and discusses the performance evaluation for each

technique and relevant discussion.

Finally, **Chapter 5** presents the conclusions and future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter provides an overview of Twitter and Reuters-21578 and presents the relevant concepts and technologies, particularly the following: SML techniques for TC, issues in improving k-NN, and issues in large datasets. This chapter introduces the most common techniques used to find relevant information. They serve as the inspiration and implementation building blocks of the method. This chapter provides a basis to understand TC SML techniques, which can be applied to create a classifier model for text documents. Automatic TC has repeatedly been used in different applications. Some of the previous studies propose techniques that include a combination of text mining techniques and social network analysis techniques. Some of them recommend mathematical formulas to compute public opinion analysis scores. However, they differ in the details of the techniques used and the virtual environment in which their experiments are conducted.

2.2 Overview of Twitter

Microblogging platforms are important real-time information resources. Twitter is one such microblogging platform, particularly for social networking, which allows people to post a broad range of different topics. The messages exchanged through Twitter are defined as microblogs (short text) because each message is limited to 140 characters. This limitation allows users to write any information with only a few words (i.e., around 20 words). Thus, Twitter messages, or "tweets," are usually focused. Several other social networking sites, such as

LinkedIn, Facebook, and Orkut, present the concept of "status" messages, and some originated much earlier than on Twitter. Figure 2.1 shows the homepage of Twitter.

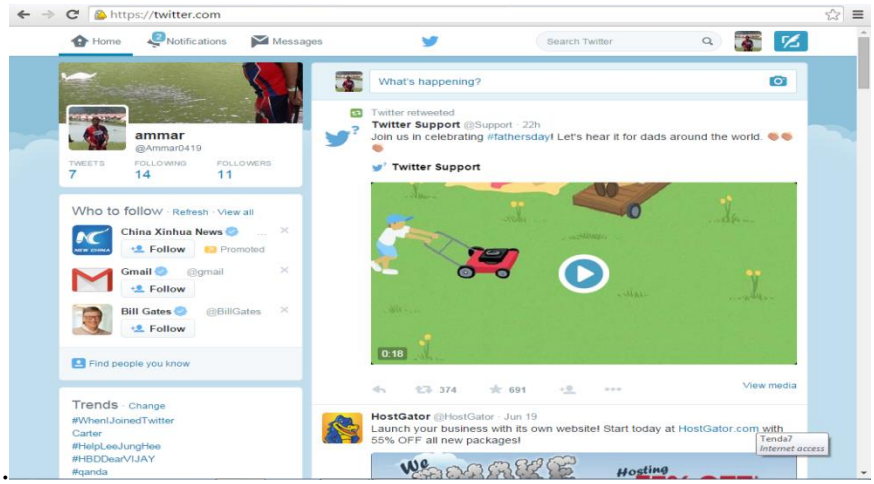


Figure 2.1: Twitter homepage.

2.2.1 Application Programming Interface on Twitter

The Twitter platform is well suited for social media analysis given that people openly share their opinions publicly. This condition is in contrast to Facebook and others where social communications are often private. Thus, topic classification and analysis is made possible using Twitter's Search and Streaming Application Programming Interfaces (API), which capture the most current and discussed terms on Twitter at any given moment.

The Twitter search and streaming API is based on the representational state transfer (REST) API and has been modified to function as a streaming API retrieval tool. The streaming API is distinct from the RESTAPI, given that the streaming type produces long-lived connections. The streaming API returns a collection of relevant tweets that match a specific query (or search keywords) accompanied by seven main parts: tweet, retweet, feed, profile, reply and mention (@), direct message (DM), and

hashtags (#). Some details on the key components of a tweet can be summarized as follows (see Figure 2.2):

Tweet: Twitter is a platform that users can use to share news, thoughts, links, and information in messages that consist of 140 characters only. These messages are called "tweets." Users "follow" one another to keep tabs on and converse with specific people only.

Retweet: A retweet can be defined as re-publishing on one's own Twitter account a post that another Twitter user has written. This term is easy to remember because it sounds like "repeat." Thereafter, the tweet shows up on Twitter posts along with the signifier "retweet (RT)." This term substitutes for "tweet," which signifies an original post.

Feed: A constant stream of Twitter messages (tweets).

Profile: A profile can be called a "handle," which includes a profile icon, a brief bio, a larger background photo, and the tweets, which a Twitter user can create.

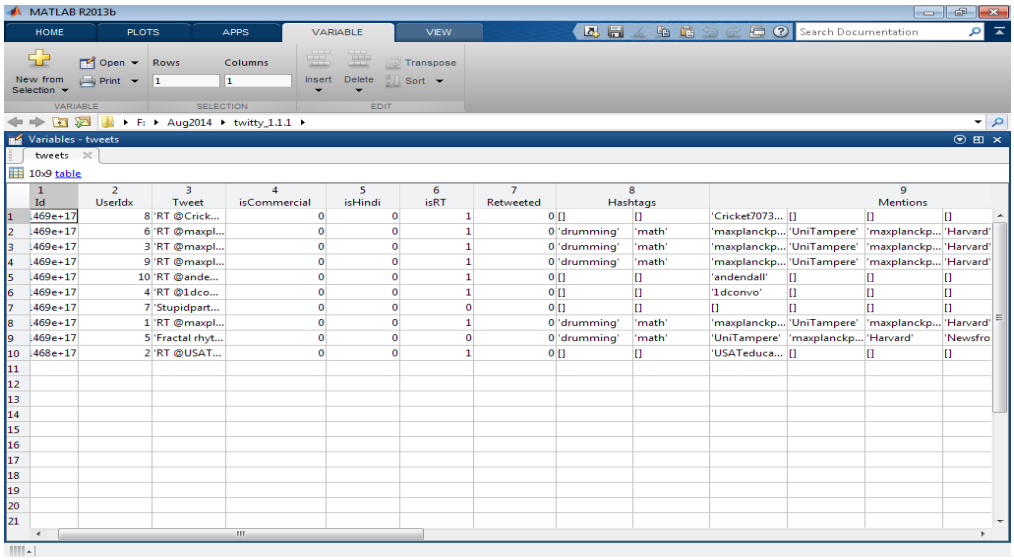
Reply: A reply can be defined as a response to another user's tweet, which starts with @ username of the person to whom one is replying. A reply button is available, which is used to respond to a particular tweet. Any tweet that is a reply starts with @ username and appears on one's notifications tab.

Mention @: A mention can be defined as a tweet, which includes another user's @ username anywhere in the body of the tweet. Anyone can read the messages, which are collected in the notifications tab for replies.

Direct messages (DM): DM is utilized to send a private tweet (Twitter update) to a person you are clicking the button followers.

Hashtag (#): A hashtag is a kind of metadata tag or label that is used on social network and microblogging services. It allows users to easily discover messages with a specific topic or content. It can also be clicked to find all the tweets that mention it in real time, even from users that you do not follow.

However, the API must be analyzed to know the limitations that are present when working with Twitter data because this denotes the access point used by researchers.



	1 Id	2 UserIdx	3 Tweet	4 isCommercial	5 isHindi	6 isRT	7 Retweeted	8 Hashtags		9 Mentions			
1	.469e+17	8	'RT @Crick...	0	0	1	0	[]	[]	'Cricket7073...	[]	[]	[]
2	.469e+17	6	'RT @maxpl...	0	0	1	0	'drumming'	'math'	'maxplanckp...	'UniTampere'	'maxplanckp...	'Harvard'
3	.469e+17	3	'RT @maxpl...	0	0	1	0	'drumming'	'math'	'maxplanckp...	'UniTampere'	'maxplanckp...	'Harvard'
4	.469e+17	9	'RT @maxpl...	0	0	1	0	'drumming'	'math'	'maxplanckp...	'UniTampere'	'maxplanckp...	'Harvard'
5	.469e+17	10	'RT @ande...	0	0	1	0	[]	[]	'andendall'	[]	[]	[]
6	.469e+17	4	'RT @1dco...	0	0	1	0	[]	[]	'1dconvo'	[]	[]	[]
7	.469e+17	7	'Stupidpart...	0	0	0	0	[]	[]	[]	[]	[]	[]
8	.469e+17	1	'RT @maxpl...	0	0	1	0	'drumming'	'math'	'maxplanckp...	'UniTampere'	'maxplanckp...	'Harvard'
9	.469e+17	5	'Fractal rhyt...	0	0	0	0	'drumming'	'math'	'UniTampere'	'maxplanckp...	'Harvard'	'Newsfro
10	.468e+17	2	'RT @USAT...	0	0	1	0	[]	[]	'USATeduca...	[]	[]	[]

Figure 2.2: Key components of a tweet.

However, it is necessary to analyze the API to know the limitations present when working with Twitter data since this denotes the access point used by researchers.

2.2.2 Feature of Short Text

Features can be drawn from a number of natural language processing approaches to text analysis and depend on both single word and term. Each of these semantic, morphological, and syntactic features differs. In practice, the features of short texts are as follows (Jin-Shu, Bo-Feng, & Xin, 2006; Yan, Cao, & Li, 2009):

- Sparseness: A short text is limited only to a dozen words with a few features and does not present sufficient word co-occurrence or shared data for a good association similarity measure. Thus, its proper language features are difficult to extract.
- Immediacy: Short texts are immediately sent and messages are received in real time. Their quantity is very large.
- Non-standardability: The description of a short text is concise, with several misspellings, noise, and non-standard terms.
- Expanding the coverage of the classifier: Texts that originate from the external data include many words/terms that do not need to exist in a small labeled training dataset. This is extremely valuable in handling future data, particularly in tokenization, and usually includes many previously unknown features.
- Flexible SML: This approach can also be considered as a supervised technique because it can use a predefined category to enhance the classifier. Nevertheless, unlike in traditional SML techniques (Ikonomakis, Kotsiantis, & Tampakas, 2005), the global data and the training/test data do not need to have the same format.
- Easy implementation: Given a classification process, preparation involves collecting large-scale data for use as global data and annotating a small training dataset.

2.3 Overview of Reuters-21578

Reuters-21578 test collection is a collection of news articles and is considered as a resource in SML and other corpus-based research. The Reuters-21578 distribution 1.0 test collection can be downloaded from http://www.daviddlewis.com/resources/testcollections/reuters21578. This dataset is based on an earlier version called the Standard Generalized Markup Language (SGML)-tagged collection by Finch. The new collection has only 21,578 text documents and is called the Reuters-21578 collection (Debole & Sebastiani, 2005).

2.3.1 Formatting on Reuters-21578

The Reuters-21578 test collection is distributed in 22 files. The first 21 files (from reut2-000.sgm to reut2-020.sgm) consist of 1000 text documents, and the last file (reut2-021.sgm) consists of 578 text documents.

All these files are in SGML format. These files describe how the SGML tags are utilized to classify each file, and each text document, into sections. The 22 files start with the following text document type declaration line:

<!DOCTYPElewis SYSTEM “lewis.DTD”>

2.4 Topic Discovery and Text Classification

The following are some issues that should be considered before addressing the subject of automatic TC on Twitter and Reuters (Desai, 2015):

- A topic can be defined as a subject that is discussed in one or more texts. Examples of topics involve news on world events, such as those concerning the

Iraqi government. Each topic is supposed to be indicated by a multinomial distribution of words.

- Topic category groups can be defined as topics that belong to a common subject area. We suppose that each topic can be specified under a topic category resulting in the use of a fully automatic technique to explore topics from each data collection. It is then used to supervise the ML technique to assign the predefined topic category, in contrast to other kinds of topic that are manually labeled.

2.5 Text Classification

Text classification is the task of automatically classifying a set of text documents into one or more predefined categories based on their contents. The main aim of TC is to derive techniques to classify natural language processing texts (Sebastiani, 2006). The objective is to automatically derive techniques that, given a set of training text documents $D = \{d_1, \ldots, d_n\}$ with predefined categories $C = \{c_1, \ldots, c_q\}$ and a new document q, which is usually denoted as the query, will predict the query's category, which belongs to one or more of the categories in C.

TC techniques are used in several tasks, such as searching for similar documents, classifying topics by text documents from legitimate short text messages on Twitter or long text documents on Reuters-21578, organizing documents in different topics, and others. Thus, the goal of classification is to automatically assign each document the appropriate label.

2.5.1 Comparison between Single-Label and Multi-Label

The TC problem can generally be divided into two important tasks. The first involves determining only one predefined category for each "unknown" text

document, as in the work of Cachopo (2007) and is often denoted as single-label TC task, where exactly one category $c_k \in C$ must be specified for each text document $d_j \in D$. The second task involves determining more than one predefined category for an "unknown" text document, as in the work of Feng, Wu, and Zhou (2005). and is often denoted as multi-label TC task, where any number $0 < n_j \leq |C|$ of categories may be specified for each document $d_j \in D$. "TC is binary text classification considered as a special case of single-label" (Sebastiani, 2002), which in particular specifies neither a predefined category nor its complement for an "unknown" text document. Several studies have been conducted in the past (Joachims, 1998).

The limitations of traditional single-label TC are the following:

(i) The volume of textual data is so large that it poses challenges in terms of experimentation.

(ii) Too many predefined categories are involved.

(iii) The number of words and documents are insufficient for training purposes.

(iv) Some text documents are labeled with a single category, whereas others are labeled with multiple categories.

Thus, implementing TC experiments using a textual dataset in its original form is difficult.

.

2.6 Text Classification Techniques

A classification technique is a systematic approach to build a classification model from an input set of data. The technique requires an ML technique to identify a model that knows the relationship between the feature set and category label of the input data. This technique should fit the input data very well and predict the category labels of previously unknown records. To develop any classification model, a collection of input datasets is utilized. Thereafter, the datasets are classified into training dataset and test dataset (J. T. Wang, Zaki, Toivonen, & Shasha, 2005).

The training dataset consists of the collection of data whose category labels are already known and is utilized to build the classification model. Thereafter, it is applied in the test dataset.

The testing dataset consists of the collection of data whose category labels are known. However, when it is specified as an input to the built classification model, should go back the accurate category labels of the data. It determines the accuracy of the classification model, which depends on the count of positive and negative predictions of the test data (J. T. Wang et al., 2005).

Figure 2.3 displays the general approach to build a classification model to tackle classification problems using the training and test datasets.

Several different techniques can be used to classify short texts into one or more topics based on their contents, such as naïve Bayes (NB), support vector machine (SVM), and k-nearest neighbor (k-NN).

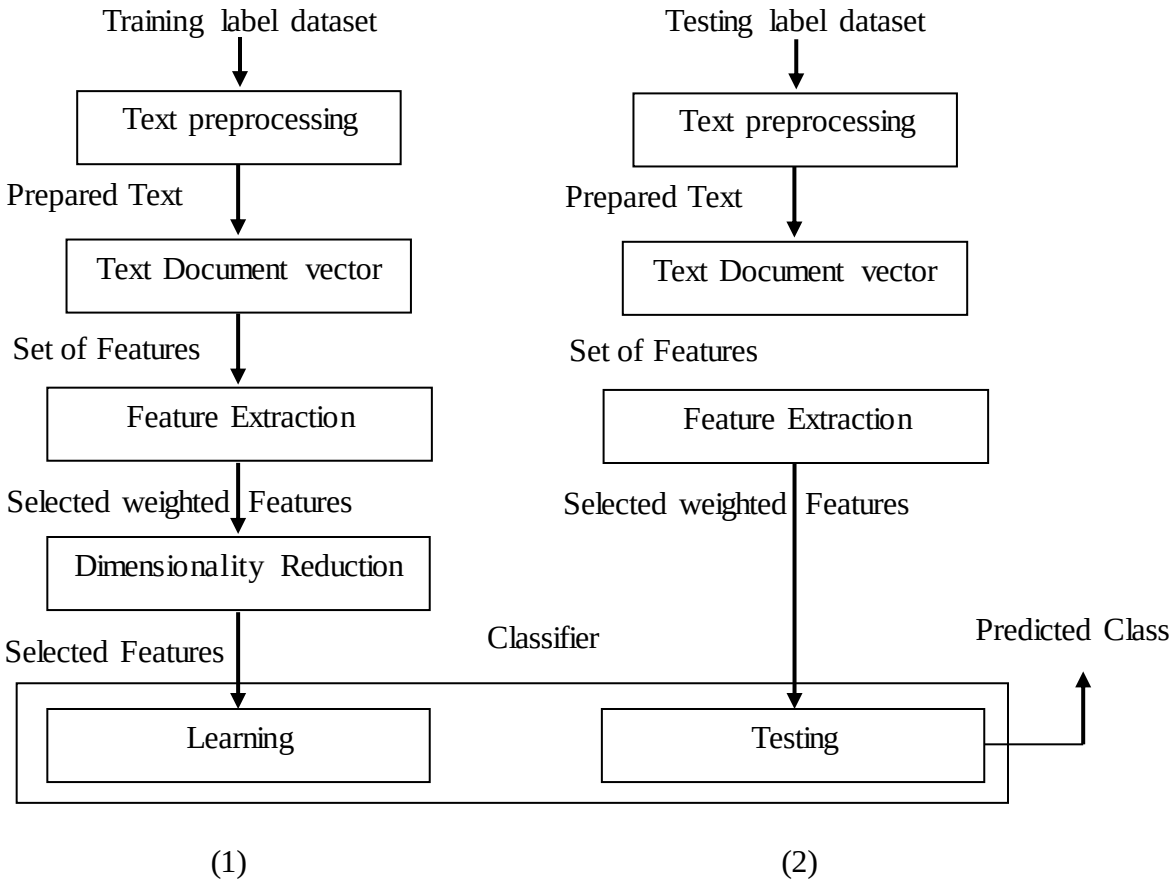


Figure 2.3: The process of TC model (1) training label (2) testing label.

2.6.1 Naïve Bayes

The NB technique is a type of module classifier that falls under different module classifier techniques of priori probability and category conditional probability. A simple probability classifier depends on applying the well-known Bayes' theorem. This theorem depends on strong (naïve) independent bases/assumptions. These assumptions are obviously violated in natural language processing texts: different kinds of dependencies between words induced by the conversational structure of a short text and its syntactic, semantic, and pragmatic characteristics. This type of classifier requires fewer training data to assume the parameters (i.e., means and variances of the variables), which are important for classification. By analyzing and searching for the dependency among the different properties, NB is very easy to implement and compute (Khamar, 2013).

Many researchers have proposed modifications to the way documents are represented to obtain the best fit with the assumptions made by NB. This task involves extracting more complex features, such as syntactic or statistical phrases, and exploiting semantic relations using lexical resources (Sahami, Dumais, Heckerman, & Horvitz, 1998). Thereafter, it is used for preprocessing. Figure 2.4 illustrates the general idea of NB.

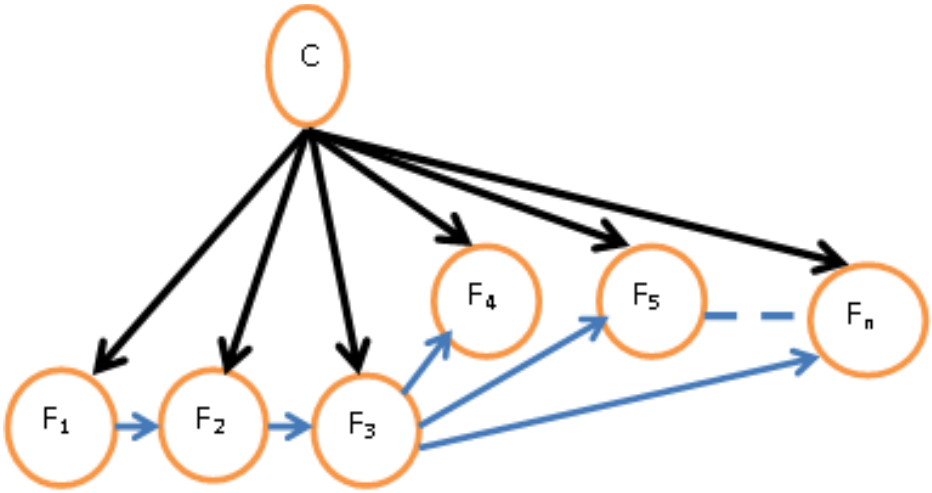


Figure 2.4**:** The general idea of Naïve Bayes classifier.

where (F_1, F_2,…F_n) is the conditional independence assumption that uses the probabilities *p(fi/c)* that are independent given class *c*.

2.6.2 Support Vector Machine

SVM classification techniques, presented by Vapnik (2013) to process two-category problems, depend on searching for a separation between hyper planes denoted by categories of data (Burges, 1996), as depicted in Figure 2.5. This means that the SVM technique can process even large feature datasets, given that its aim is to measure the margin of the separation of the data, which uses rather than matches features. SVM is trained using predefined classified documents.

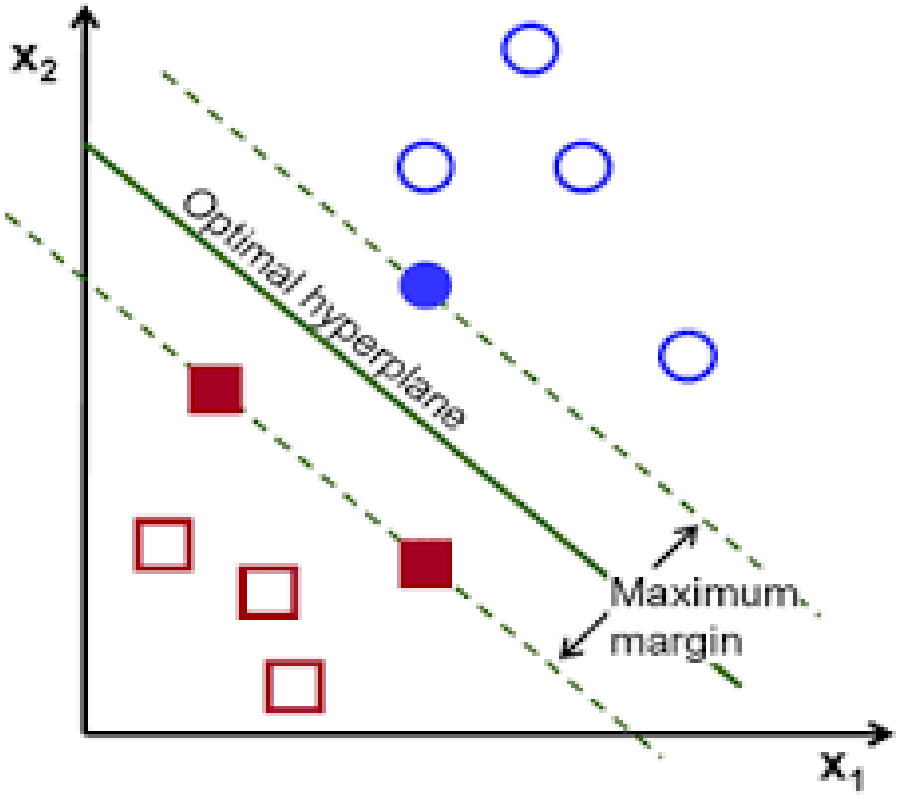


Figure 2.5: Example of SVM classification.

InKwok (1998), has shown that scales well and has good performance on large datasets. Both Bayes and SVM techniques are linear, efficient, and scalable to large document datasets (Rennie & Rifkin, 2001).

2.6.3 K-Nearest Neighbors

presented the first application of k-NN to text classification. The main idea is to determine the category of a given "unknown" document based not only on the document that is nearest to it in the document space but also on the categories of the k documents that are nearest to it. The k-NN technique is a similarity-based learning technique that is very effective for different problems, including text classification (Mooney, 1996). Given a test document to determine the category that it belongs to, the k-NN technique searches the k-nearest neighbors among the training text

documents and utilizes the categories of the k neighbors to weigh the category candidates.

The similarity function score of each neighbor text document to the test document is utilized as the weight of the categories of the neighbor text document. If several of the k nearest neighbors support a category, then the per-neighbor weights of that category are added, and the resulting weighted sum is used as the probability score of candidate categories. A ranked list is then obtained for the test text document. Thus, text document classification depends on the thresholding of these scores, and the binary categories are obtained (Yang & Liu, 1999).

2.7 Supervised Machine Learning Techniques for Text Classification

Currently, many ML techniques are used for text classification. Text classification is an important research area of text mining, where the texts are classified with supervised, unsupervised, and semi-supervised knowledge. Traditionally, this process is solved manually, but such manual classification is expensive to scale, labor intensive, and requires a long time for classification. Thus, researchers have explored the use of ML techniques for automatic text classification (Ikonomakis et al., 2005). Among the different ML techniques used in text classification, the most popular is SML, whose underlying input–output relation is learned using a small number of training data and where the output values for unknown input objects are predicted (Sugiyama & Kawanabe, 2012).

Text classification is the process of assigning a predefined category based on their content. Thus, classification in ML problem is an issue of the efficiency of supervised learning because the learning process is "supervised" using the knowledge of the categories and of the training objects that are relevant to them

(Sebastiani, 2002). In the same context, automatic TC is a form of SML, in which a set of labeled documents is used to train a classifier, which is then employed to assign one or more predefined category labels to new documents (Qi & Davison, 2009).

Han, Karypis, and Kumar (2001) proposed the weight-adjusted k-NN classification algorithm, which depends on the k-NN classification paradigm. The experimental results of many real-life document datasets exhibit the promise of WIND, given that it performs better than state-of-the-art classification techniques, such as C4.5, RIPPER, Rainbow, PEBBLES, and VSM. Pang, Lee, and Vaithyanathan (2002) classified movie reviews depending on whether they are positive or negative. They found that standard ML techniques definitively perform better than human-produced baselines. The three ML methods employed (i.e., NB, maximum entropy classification, and SVMs) do not perform as well on sentiment classification as on normal topic-based classification. Li, Yu, and Lu (2003) modified the k-NN method by changing the bias on larger categories in the normal k-NN algorithm. The method is therefore applied on Chinese texts only, and it must be universally applicable to solve classification problems for data in different languages.

Similar to this study Kwon and Lee (2003) improved the performance of the k-NN technique with a feature selection approach and a term-weighting method that uses markup tags and repaired its document–document similarity measure. They found that the classification of Web pages extends to the classification of the entire Web site. Soucy and Mineau (2005) introduced a new weighting technique that depends on the statistical estimation of the importance of a word for a given classification problem. The experimental result showed that this new weighting

technique significantly enhances the classification accuracy, as measured in several classification tasks.

In the same context, Kurada and Pavan (2013) Kurada and Pavan (2013) studied text classification using AkNN text classifier and kMdd clustering. They used local weighting TF method for feature selection and employed cosine similarity in the AkNN technique. The experimental results conducted on Reuters-21578 were applied, and comparisons with traditional k-NN classifiers showed better results in both clustering and classification.

Most studies in TC have been conducted only in a small number of areas. Go, Bhayani, and Huang (2009) introduced a novel method to automatically classify the sentiments of Twitter messages. ML techniques (e.g., NB, maximum entropy classification, and SVMs) can perform with high accuracy in classifying sentiments when this technique is used. Kamruzzaman and Haider (2010) presented a new methodology for TC that requires fewer documents for training. They used the concept of the NB classifier based on derived features and added the genetic technique for the final classification. The accuracy of existing techniques needs to be enhanced for both data training and calculation time. Suguna and Thanushkodi (2010) improved the k-NN algorithm by combining k-NN with the genetic algorithm, thereby improving the classification performance. The results show that this method improves the accuracy of classification by reducing the complexity of the k-NN. Moreover, the resulting performance is compared with those of the normal k-NN, CART, and SVM classifiers.

In general, TC has an essential role in classifying short texts, such as those on Twitter and Facebook. Sriram, Fuhry, Demir, Ferhatosmanoglu, and Demirbas (2010) classified tweets according to a predefined group of general categories, such