SHORT TEXT CLASSIFICATION USING AN ENHANCED TERM WEIGHTING SCHEME AND FILTER-WRAPPER FEATURE SELECTION

ISSA MOHAMMAD IBRAHIM ALSMADI

UNIVERSITI SAINS MALAYSIA

2018

SHORT TEXT CLASSIFICATION USING AN ENHANCED TERM WEIGHTING SCHEME AND FILTER-WRAPPER FEATURE SELECTION

by

ISSA MOHAMMAD IBRAHIM ALSMADI

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

December 2018

ACKNOWLEDGEMENT

First and foremost, I thank my Creator Almighty Allah for his blessings and generosity.

I would like to express my special appreciation and thanks to my advisor Dr. Keng Hoon Gan, for her constant guidance, encouragement and for valuable feedbacks to improve my research work. I express my sincere appreciation for the recommendations and suggestions received from her.

A special thanks to my family. I thank my parents for their support, prayers, and patience. My deepest gratitude goes to my Wife, and my little son, Mohammad, for unlimited support. And, of course, I thank my brother and sisters, for their patience, and encouragement throughout my Ph.D. journey.

I would also like to extend my appreciation to the School of Computer Sciences, Universiti Sains Malaysia (USM), for making my research study an exciting experience. I am very grateful for the collaboration and friendship I have had in USM, and, particularly, thanks to Omar Alejla, Osama Alomari, and Anas Alshishani.

TABLE OF CONTENTS

LIST OF PUBLICATIONS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

Acc2	Accuracy balanced
ACO	Ant colony optimization
Bat	Bat algorithm
BNS	Bi-normal separation
BOW	Bag of words
CC	Correlation coefficient
Ch2	Chi-squared
CI	Computational intelligence
DT	Decision tree
ES	Evolution strategy
FCD	Feature contribution degree
FN	The number of incorrect predictions that an instance is negative.
FP	The number of incorrect predictions that an instance is positive.
GA	Genetic algorithm
GWO	Grey wolf optimization
IDF	Inverse document frequency
IG	Information Gain
KELM	Kernel extreme learning machine
KNN	K-nearest neighbor

LR	Logistic regression
MBF	Markov blanket filter
MOPSO	Multi-Objective Particle Swarm Optimization
MR	Most relevance
NB	Naïve Bayes
NN	Neural Network
OR	Odd ratio
PBIL	Population-based incremental learning
POS	Position of words
PSO	Particle swarm optimization
RF	Relevance frequency
SVM	Support vector machine
SW	Simple supervised weight
TF	Term frequency
TN	The number of correct predictions that an instance is negative.
TP	The number of correct predictions that an instance is positive.
TS	Term strength
VSM	Vector space model

PENGELASAN TEKS PENDEK MENGGUNAKAN SKIM PEMBERAT TERMA YANG DIPERTINGKATKAN DAN PEMILIHAN CIRI PENAPISAN-PEMBALUTAN

ABSTRAK

Penggunaan rangkaian sosial dalam kehidupan harian telah menyebabkan peningkatan dalam kuantiti dokumen elektronik pendek. Rangkaian sosial, seperti Twitter, merupakan mekanisme biasa dimana orang boleh berkongsi maklumat. Penggunaan data yang tersedia ada melalui media sosial telah meningkat secara beransur-ansur untuk banyak aplikasi. Pertindihan dan kebisingan dalam teks pendek adalah masalah biasa dalam media sosial serta dalam aplikasi berbeza yang menggunakan teks pendek. Walau bagaimanapun, kependekan dan kejarangan tinggi teks pendek menyebabkan prestasi klasifikasi yang lemah. Menggunakan kaedah klasifikasi teks pendek yang berkuasa dapat memberi kesan yang penting kepada banyak aplikasi dari segi peningkatan kecekapan. Penyelidikan ini bertujuan untuk menyiasat dan membangunkan penyelesaian untuk diskriminasi dan pemilihan ciri untuk klasifikasi teks pendek. Untuk diskriminasi ciri, kami memperkenalkan pendekatan penyeliaan terma mudah yang dipanggil kaedah pemberat SW, yang mengambil kira sifat khas teks pendek dari segi kekuatan dan pengedaran terma. Untuk menangani kekurangan pemilihan ciri sedia ada dengan teks pendek, tesis ini mencadangkan pendekatan pemilihan ciri penapisan-pembalutan. Pada peringkat pertama, kami mencadangkan model pemilihan ciri berasaskan penapis adaptif yang berasal dari kaedah nisbah ganjil, yang digunakan untuk mengurangkan dimensi ruang ciri. Di peringkat kedua, algoritma pengoptimuman serigala kelabu (GWO), algoritma carian heuristik baru, menggunakan ketepatan SVM sebagai fungsi kecergasan untuk mencari ciri subset optimum. Untuk mengesahkan kecekapan pendekatan GWO yang dicadangkan, kerja ini menggunakan tiga kaedah metaheuristik biasa (algoritma genetik, kelawar, dan pengoptimalan rawak zarah) untuk masalah pemilihan ciri teks. Penemuan keputusan eksperimen terkawal untuk dua set data pendek berdasarkan masalah yang berbeza menunjukkan keberkesanan teknik yang dicadangkan dalam klasifikasi teks.

SHORT TEXT CLASSIFICATION USING AN ENHANCED TERM WEIGHTING SCHEME AND FILTER-WRAPPER FEATURE SELECTION

ABSTRACT

Social networks and their usage in everyday life have caused an explosion in the amount of short electronic documents. Social networks, such as Twitter, are common mechanisms through which people can share information. The utilization of data that are available through social media for many applications is gradually increasing. Redundancy and noise in short texts are common problems in social media and in different applications that use short text. However, the shortness and high sparsity of short text lead to poor classification performance. Employing a powerful short-text classification method significantly affects many applications in terms of efficiency enhancement. This research aims to investigate and develop solutions for feature discrimination and selection in short texts classification. For feature discrimination, we introduce a term weighting approach namely, simple supervised weight (SW), which considers the special nature of short text in terms of term strength and distribution. To address the drawbacks of using existing feature selection with short text, this thesis proposes a filter-wrapper feature selection approach. In the first stage, we propose an adaptive filter-based feature selection method that is derived from the odd ratio method, used in reducing the dimensionality of feature space. In the second stage, grey wolf optimization (GWO) algorithm, a new heuristic search algorithm, uses the SVM accuracy as a fitness function to find the optimal subset feature. To validate the efficiency of the proposed GWO approach, this work adopts three common metaheuristics (genetic algorithm, bat, and particle swarm optimization) for a text feature selection problem. The controlled experimental results findings on two short text datasets with different problems indicate the effectiveness of the proposed techniques in short text classification.

CHAPTER 1

INTRODUCTION

1.1 Overview

The growth of social networks has provided new ways of accessing information via the internet. As an alternative to traditional and previous online information sources, such as websites, microblogs are significant information portals for different topics ranging from politics to entertainment. Users can select their preferred information and thus increasingly tend to use social networking platforms, such as Twitter, Google+, and Facebook. People also use such tools because of their speed, efficiency, and comprehensive characteristics. Millions of short text are produced daily in the form of posts and comments. These documents tend to have a length of no more than 200 characters; for example, Twitter posts comprise up to merely 140 characters.

However, users of social networks differ in interest and preference. Many people face problems finding appropriate data in reasonable timeframes and thus need a means to identify the most relevant data and classify these on the basis of the main topics covered or other features, such as followers. Consequently, text classification is becoming important in many problems, such as sentiment analysis, tweet personalization, and spam filtering. The short text is classified using the same series of steps used in classifying long text. The classification involves a feature extraction step and a classification step and uses information of labeled data and features from training data. The principal difficulty in short text classification is the brevity of documents and the sparsity of feature space. The general framework of the classification process is summarized in Figure. 1.1.

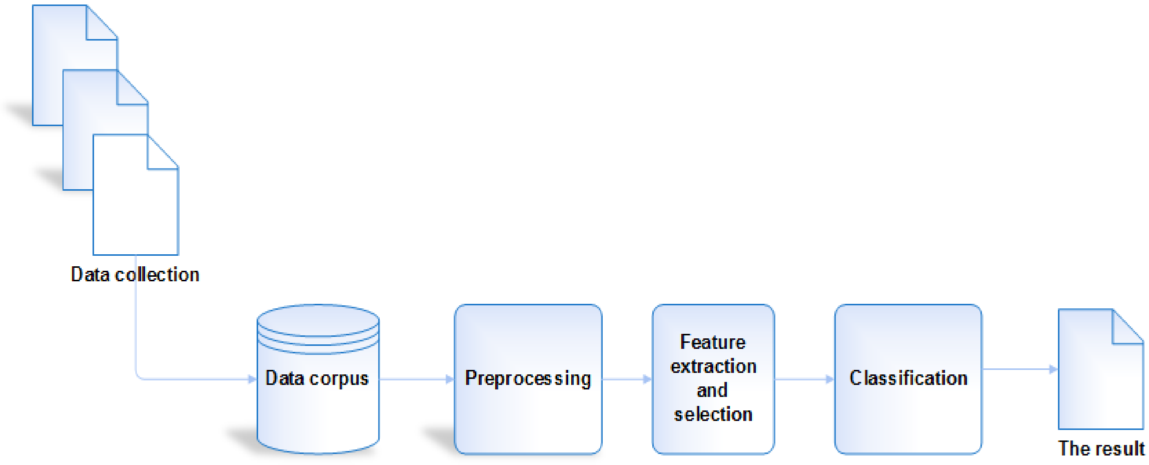

Figure 1.1 General text classification steps (Song et al., 2014).

Text classification is the process of assigning unlabelled text on the basis of its features to certain predefined class labels using induction algorithms. Automatic text classification consists of two essential subprocesses, namely, text representation and classifier learning. The former subprocess transforms textual content into a format that the classification algorithm can process, and the second represents the documents in text classification using the vector space model (VSM), which treats document content as a bag of words (BOW) and disregards the schematic and grammatical structure of text. In the VSM, the term is represented using a term weighting scheme, in which a numerical value is assigned to each word to reflect its contribution to the text classification.

Irani et al. (2010) proposed a method of classifying tweets by trend and improved classification performance by merging tweets with their associated web pages and using information gain (IG) for feature space reduction. Fiaidhi et al. (2013) proposed a hierarchical ensemble classification approach for trending topics. K-nearest neighbor (KNN), term frequency (TF)–inverse document frequency (IDF) (TF–IDF), NB, and language model were combined to classify tweets into relevant labeled classes. The dataset was built by collecting trending topics using T3C and then labeling these topics. The experiment in this work showed that ensemble classification achieved approximately 75% accuracy and was superior to the other classifiers when used alone. Moreover, classifier accuracy was increased by using the language model with N-gram. Weissbock et al. (2013) attempted to solve the problem of shortness by expanding tweets that contained links and accessing the web pages of these links. To increase the number of terms, tweets were extended by appending the titles of web pages and the 10 most frequently used terms on the web pages. An evaluation of an experiment using three familiar classifiers, namely, the support vector machine (SVM), naïve Bayes, and decision tree, demonstrated that expanding tweets with external data enhanced classification accuracy. Wang et al. (2013) proposed a feature selection approach that exploits the data content of short text regardless of external data. The main idea is that the feature to be selected must appear in a sufficient number of documents and thus be representative of the topic. By separating feature selection and feature vector construction, the proposed method attained reasonable results with web snippet dataset via the naive Bayes classifier. However, the classification has considerable time requirements.

Chen et al., (2017) proposed a short text classification approach based on LDA and KNN. The proposed approach builds based on the assumption that says if the words in two texts related to common latent topics, these common topics consider as the third party. Then this topic used to check out if there is a similarity between the short texts to determine if the texts belong to the same class. Selvaperumal and Suruliandi, (2014) proposed a system of classifying tweets to general topics. The classifier is based on other feature with textual content like URL's, retweeted tweets and influential users tweet. If the tweet contains a URL, then the tweet is classified into the same class of the corresponding webpage. If the tweet contains trend topics, then keywords from top five retweeted posts of this trends are collected, and the collected word is classified using conventional classifiers. The tweets that contain only text continents classified based on their text. The proposed method shows a competitive performance compared with common classification algorithms like KNN, Naïve Bayes, SVM.

The high dimensionality of feature spaces is an essential problem in text classification. The presence of many redundant words negatively affects performance because it reduces the accuracy and increases the running time of the classification process (Sebastiani, 2002). Feature selection solves this problem by retaining only items that can contribute to the classification process. Feature selection on textual data has been considerably studied (Bharti and Singh, 2014; Javed et al., 2015; Largeron et al., 2011; Tutkan et al., 2016; Zheng et al., 2004). Recently, many researchers have used metaheuristic search algorithms to solve the feature selection problem. These methods, which perform well in text classification, include those based on genetic algorithm (GA) (Yang and Honavar, 1998; Hamdani et al., 2011; Rostami, 2014), ant colony optimization (ACO) (Ke et al., 2008), and particle swarm optimization (PSO) (Chuang et al., 2011). Mirjalili et al., (2014), developed a new optimization algorithm called grey wolf optimization (GWO), which is inspired by the social nature of wolves during the hunting process. GWO is used in applications such as machine learning (Mirjalili, 2015) and feature selection (Li et al., 2017; Vosooghifard and Ebrahimpour, 2015).

By considering all the fact about the short text classification, this research introduces contributions to perform sufficient term representations and selection for short text classification problems. Moreover, this research aims to address the fundamental issues of determining the capability of term weighting schemes used in traditional text classification to achieve good performance in short text classification and the applicability of an efficient term weighting scheme that considers the special nature of the short text. We intend to propose an enhanced term weighting scheme as well as an adaptive feature selection method to improve the performance of short text classification.

1.2 Motivation

Rapid developments in social networks and their usage have increased the quantity of short electronic documents. Therefore, it is essential to effectively classify these documents into relevant classes on the basis of text content in many applications. A short-text classification is essential in many areas, such as spam filtering, sentiment analysis, Twitter personalization, customer review, and other fields, which are related to social networks.

In the spam filtering domain, the popularity of social networks, as well as their easy use have emerged in the ill-usage of this media. The privacy, undesirable user, and a huge amount of spam post big challenges in social networks. In fact, experts demonstrate that ten of millions of social network accounts are used to propagate spam.

Since not all the message in the social networks is trustworthy (Igawa et al., 2016). Therefore, a technique, which detects spam short text is necessary through extracting meaningful features from the text using Natural Language Processing. Different machine learning techniques can be used for spam detection (Miller et al., 2014; Silva et al., 2017).

Another application, which takes advantage of the precise short text classification, is the sentiment analysis or opinion mining, which can be considered a classification problem (Medhat et al., 2014). It aims to classify opinion texts that represent a positive or negative opinion toward a specific idea or product. The sentiment analysis, in general, considers the text contents' post or the comment in addition to other features of the documents. Therefore, the enhancement in short text classification has an impact on the quality of the sentiment prediction (da Silva et al., 2014; Liu et al., 2017; Michalec et al., 2016; Pak and Paroubek, 2010). Other applications that deal with this type of data and related to e-commerce (Khadjeh Nassirtoussi et al., 2015) or social network include Twitter personalization (Weilin and Hoon, 2015) and customer review.

Due to its special characteristics, organizing and classifying a short text is considered a challenging task. Together with a good term representation, the need for selecting the most appropriate term subset is a critical factor in text classification. The limit in their lengths has led to the poor representation of this short text. However, the sparseness of this data without enough shared context increases the dimension of the feature space, which intensifies the computational cost of the classification task and reduces its performance. Moreover, social media short texts are briefly expressed, and they are informally written, with a lot of misspellings and grammatical mistakes. Therefore, the extraction of valid knowledge has become a hard task.

1.3 Problem Statement

Many classification mechanisms were proposed for short text using the machine learning approach. These solutions (Chen et al., 2017; da Silva et al., 2014; Ma, Aixin Sun, 2013; Quan et al., 2011; Tang and Liu, 2012, 2014; Tommasel and Godoy, 2018a ; Wang et al., 2014; Zheng et al., 2015) differ by means of building classifiers and the features they use to train classifiers. Nonetheless, these models encounter similar challenges. These challenges are related to short text; namely, shortness and sparseness, which reduce classification performance (Song et al., 2014). This substantiated the problem of short text classification, that is, the brevity of the text and the high sparsity of its contents.

Like long text classification, the main issue in short text classification is the weighting scheme, which is used to characterize terms and identify important terms within the text for classification. Binary weight, TF, and TF–IDF are traditional approaches. These methods are implemented to assign weights to terms. Some researchers have used supervised methods, which usually combine term frequency and feature selection measures, such as IG, odds ratio, and Chi2 due to the usefulness of these measures (Debole and Sebastiani, 2003; Erenel et al., 2011; Quan et al., 2011). These approaches, when they are applied to long text classification, have produced high accuracy in most cases. As for short text classification, researchers (Carolina et al., 2015; da Silva et al., 2014; Hassan et al., 2013; Tuarob et al., 2014) applied similar methods that are designed for long text classification to assign weights to terms in the VSM. However, these schemes have not worked properly. This is because of the characteristics of short texts. A short text includes a few words only, and words are rarely repeated in short texts. Moreover, these schemes do not consider the special nature of the short text, i.e., the sparseness of the text.

One of the most critical challenges in text classification is the high dimension of the feature space. It increases the computational cost and reduces the performance of text classification (Tang et al., 2016; Uysal, 2016). Generally, the high sparsity of short text extends the dimensionality of its feature space. The high dimension of the feature space contains features that can be relevant, irrelevant, or redundant. The redundancy of the irrelevant features reduces cost-effectiveness, as well as classification task performance. It enlarges the feature space dimensions and consequently decreases accuracy (Fragoudis et al., 2005; Sebastiani, 2002). This problem can be mitigated by feature selection by retaining the features that are relevant for classification only. Researchers have investigated short text classification via available feature selection techniques, which work well with large documents that are heavy in text and written in standard English. However, these methods may not perform well when used to deal with short texts include a few words, each is much more related to the subject of the short text than any words in long documents. Accordingly, the problem of sparseness must be avoided when selecting sufficient topical words (Wang et al., 2013). In short text, the high sparsity of the term space results in complicating the exploit of the correlation between terms (Liu et al., 2010; Tommasel and Godoy, 2018b).

The filter-based method has become popular for text feature selection. It is considered a low computational method for high dimensional text data. It assesses each feature individually regarding a certain class. However, the selected set contains a set of relevant features that may include redundant terms. This can reduce the classification task accuracy (Javed et al. 2015). Recently, researchers in the text classification domain, have used the evolutionary computation techniques as search strategies in wrapper methods to select a subset of terms, which aims to maximize classification performance. Adaptation of wrapper method alone is indeed high computational. To further enhance the performance of short text classification, a hybrid feature selection method is proposed. GWO is an optimization algorithm, which is developed by (Mirjalili et al., 2014). GWO is considered as a simple and easy method to implement in comparison with other methods (Faris et al., 2017) and, therefore, this algorithm has been used for feature selection (Emary et al., 2015; Emary E, Zawbaa H M, 2015; Vosooghifard and Ebrahimpour, 2015). However, to the researcher’s knowledge, GWO has not been used for feature selection to tackle the current text classification problem.

1.4 Research Objectives

This thesis aims to enhance the overall performance of short text classification in terms of feature selection and term weighting to improve the effectiveness, and accuracy of classification by addressing the brevity and sparse term spaces of short text data. The research objectives are as follows:

- To propose an enhanced weighting scheme with better discriminative capabilities;

- To reduce the dimensions of feature spaces so that text classification efficiency is improved.
- To propose an improved feature selection technique to find a new term subset of more informative terms.

1.5 Research Contribution

The contributions of this study are as follows:

- This study introduces an enhanced supervised term weighting scheme, namely, simple supervised weight (SW), which considers the special nature of short texts in terms of term strength and distribution. This work demonstrates the superior performance of SW in a high-dimensional vector space over the term weighting schemes that are used to represent a baseline term weighting in traditional text classification.
- This study introduces a filter-based feature selection method. It is called 'Most Relevance' (MR) feature selection. MR reduces the dimension of the original feature space by removing noisy, irrelevant, and redundant features.
- The first GWO-based wrapper feature selection for short text classification is proposed in this study. It is combined with MR in the filter-wrapper feature selection approach to select the highly discriminating terms to further enhance short text classification performance. The proposed model is expected to maximize the performance of the classification task. Furthermore, outperforms three other metaheuristic algorithms, i.e., (GA, bat, and PSO) techniques based wrapper algorithms.

1.6 Thesis Outline

The remainder of this thesis is organized into four chapters. Chapter 2 provides a review of related work. Chapter 3 illustrates the methods of this thesis. Chapter 4 explains the experimental evaluation of the proposed contribution on two Twitter datasets, and Chapter 5 concludes the thesis.

Chapter 2 presents a review of the basic concepts of machine learning and classification, term weighting, feature selection, and evolutionary computation, especially the GWO optimization algorithm. This chapter reviews related work in term weighting in text classification using supervised and unsupervised methods and then discusses the background of feature selection using conventional and evolutionary computation approaches. Finally, this chapter provides a critical analysis of the challenges that motivate this research.

Chapter 3 outlines the methodology of this research and the corresponding experiments. The first two sections present the general methodology and the principal natural language preprocessing step. The third section presents our proposed term weighting scheme. Then, the fourth section details the enhanced filter-based feature selection proposed to address the tweet classification problem. The fifth section introduces a wrapper-based feature selection method that uses a new evolutionary computation algorithm known as GWO. The sixth section presents our proposed filter-wrapper feature selection method. The seventh section contains information about the datasets and parameter settings of the algorithms used in the evaluation process. Moreover, the efficiency of the proposed method is compared with three wrappers that are based on common algorithms (GA, bat, PSO).

Chapter 4 discusses the experiments conducted with the proposed work and evaluates each stage of the research. The first section explains the experiments performed to evaluate the capability of the proposed term weighting scheme and compare it with a set of common term weighting schemes. The second section presents the experiments performed to assess the capability of the feature selection method (MR) in dimension reduction and compare it with the common techniques used in text classification. The last section discusses the experiments that reveal the successful performance of the proposed filter-wrapper method in solving the text feature selection problem on short text. It also examines the performance of the proposed combination weighting scheme and all introduced feature selection methods and their capability to maximize the accuracy of the classification task.

Chapter 5 the chapter provides the research conclusion and possible future work.

CHAPTER 2

LITERATURE REVIEW

This chapter provides a review of basic concepts and essential knowledge of the state-of-art related to this research. Section 2.1 outlines of the text classification, and the main characteristics of the short text. 2.2. Review of most common term weighing methods. Feature selection concepts and theory are illustrated in section 2.3. A brief background of evolutionary computation techniques are covered in section 2.4. Section 2.5 contains a detailed description of machine learning algorithms and their application. Section 2.6 includes the critical analysis of our research.

2.1 Short-text Classification

Text classification is the process of assigning documents to correct labels of predefined classes. Short text classification, on the other hand, is used in many applications, such as sentiment analysis, customer review, search, and many other areas in information retrieval. This process is divided into two steps, namely, training and test steps. In training, the training corpus is divided into classes, and feature extraction is used to eliminate noise and redundant terms. Therefore, only the relevant discriminator term remains in the dataset to be used in the test step. When the test document is inputted, the data in the learning step are used to assign the correct class label to this document.

Short text classification and their applications are a new trend in text mining. Short texts involve the type of problem that deals with documents, which are relatively short in their length. A wide range of applications utilize this kind of text, like microblogs, Twitter, mobile messages, and news comments. These texts are short and usually have a maximum of 200 characters. For instance, a tweet has a maximum of 140 characters at most, whereas a short mobile message comprises less than 70 characters. Song et al. (2014) concluded that all short texts have the following attributes.

(1) Shortness: A short text contains a few words, thereby possibly resulting in an inadequate representation of the document.

(2) Sparsity: A short text has limited length. This limited capacity is used to express many different topics, with each user using their own words and writing style. Therefore, a certain topic has diverse content, and obtaining its features precisely is challenging.

(3) Misspellings and informal writing: In most cases, especially in comments in microblogs, a short text is presented briefly and includes many misspellings, noise, and a special language.

According to Faguo et al. (2010), sparsity and shortness significantly affect the performance of machine learning classifiers. The text content of a short text may be highly diverse despite having only a limited number of words. This will complicate the feature space construction of text classification. Furthermore, short texts, which are derived from social networks usually suffer from non-standard ability. The social media platform users usually write their comments or post briefly with many misspelling and grammatical mistakes. Consequently, this adds more challenge to the feature extraction process and may lead to short texts' poor representation. Therefore, some of the existing classification methods may not perform well with short texts. Preprocessing is an essential step in text classification. The preprocessing step is conducted through sequential steps, and it starts by applying several purification operations to real data to remove impurities, represent data in the standard form and prepare them for the application of different machine learning algorithms. Generally, short texts contain a lot of unnecessary data and noise. Cleansing the text of its impurities comprises the preprocessing step. This can alter the performance of the classification task (Haddi et al., 2013). The limited number of keywords results in a classification task with low accuracy. Therefore, a mechanism for increasing the number of the terms inside the texts without changing their semantics is required to enrich the short text representation using additional semantics (Kamath and Caverlee, 2011). Feature enrichment is one of the solutions for texts' shortness. According to Kamath and Caverlee, (2011), three main approaches to feature enrichment are available.

External-based Enrichment: This technique adds features from an external source to increase the terms in the feature set depending on the links inside the short document, especially in microblogs. Features can, therefore, be collected from the web page associated with these links (Klassen, 2013). In addition, we can enrich short texts depending on the context of the short text corpus and semantic similarity or through the search engine to find web pages from trusted sites. We can add the resulting terms directly to the corpus of the short text after processing the collected web pages. Others depend on topic taxonomy, such as using a Wikipedia category, in enriching the short text (Phan et al., 2008; Vicient and Moreno, 2015). Using the parts of speech (POS) is another method; for example, recognizing nouns in short messages provides a reliable understanding of the entire message in many cases (Vaghela and Scholar, 2016).

Lexical-based Enrichment: The lexical approach can be used to solve the sparseness problem of the term. In character-based n-grams, the feature is formed by taking N continuous characters in the document. However, in word n-grams, the feature is built from the consecutive words in the document (Bekkerman and Allan, 2003; Cormack, G. V., Gómez Hidalgo, J. M., & Sánz, 2007; Kang et al., 2012; Lane et al., 2012). The lexical approach is widely used for its simplicity because this approach does not require external data to augment the terms in the text.

Collocation-Based Enrichment: (Kamath and Caverlee, 2011) defines collocation as "two or more words together form an expression that matches the way saying things". Accordingly, the terms can be added by applying the collocation approach. The most important factor in collocation is the association measure, which is a mathematical method used for measuring the closeness of the words in the phrase. The association measure essentially estimates the co-occurrences between the words in the phrase. Examples of this measure are mutual information, log-likelihood ratio, and chi-square. However, this approach may not be effective because it is time-consuming, and it enlarges the dimension of the feature. It is sometimes difficult to find an appropriate source for enrichment. Moreover, the additional features should be compatible with the semantics of the original texts.

Several solutions, which aim to improve short classification, are based on an ensemble classification (da Silva et al., 2014) or based on the semantic analysis (Chen et al., 2017). However, other researchers have focused on enhancing the term representation of the short text by introducing a weighting scheme, which tackles several short text challenges (Quan et al., 2011; Zheng et al., 2015). Some solutions are based on the selection and extraction approaches (Tommasel and Godoy, 2018a; Wang et al., 2013) or on performing the classification using a group of features including the textual features. The rest of this chapter provides a brief review of the state-of-the-art term weighting and feature selection on text classification. The notation used in the theories is first presented. Table 2.1 are defined different notation used in these study**,** based on term t_j in class c_i.

Table 2.1: Description of the notations used in the theories.

Notation	**Description**
$A_{i,j}$	A number of documents belonging to class ci that contain the term tj.
$B_{i,j}$	A number of documents belonging to class ci that do not include the term tj.
$C_{i,j}$	A number of documents that do not belong to class ci and contain the term tj.
$D_{i,j}$	A number of documents that do not belong to class ci and does not contain the term tj.
N	A total number of documents in the corpus N=A+B+C+D.
Np	A number of documents in the positive class Np=A+B.
Nn	A number of documents in the negative class Nn=C+D.
$\boldsymbol{p(t_j)}$	The probability of documents that contain the term t_j.
$\boldsymbol{p(c_i)}$	The probability of the documents in the total collection that belongs to class c_i
$\boldsymbol{p(t_j, c_i)}$	The probability of documents belongs to class c_i and contain the term t_j .
$\boldsymbol{p(t_j^-)}$	The probability of documents that do not contain the term t_j.
$\boldsymbol{p(t_j, c_i^-)}$	The probability of documents do not belong to class c_i and contain the term t_j.
$\boldsymbol{p(t_j^-, c_i)}$	is the probability of documents belong to class c_i that does not contain the term t_j.
$\mathbf{p}(\boldsymbol{t_j^-, c_i^-})$	$\mathrm{p}(t_j^-, c_i^-)$ is the probability of documents not belong to the class c_i, and does not contain the term t_j .

2.2 Weighting schemes in text classification

To classify a text or obtain information from it, determining which words are significant within the text is needed. The most popular way to accomplish this task is by assigning a numeric value to each word to reflect its contribution to document classification. This value is referred to as the weight of the term. The TF-IDF scheme is a weighting scheme that is extensively used today because of its simplicity and efficiency in classification.

To classify a text, a document should be represented as a set of terms; the term itself can be single or have multiple words. Single-word terms are generally used in text representation. Multi-word terms can be one of the following: syntactic phrases, statistical phrases, and term sets. Syntactic phrases (Scott and Matwin, 1999) are a concatenation of words arranged by syntactic relations. Familiar phrases are usually verb, noun, and adjective phrases. Statistical phrases (as n-grams) (Bekkerman and Allan, 2003) are a series of n contiguous words that are employed to define co-occurrence-based features. A term set is a sequence of words in which the co-occurrences of the terms are not necessarily adjacent (Badawi and Altınçay, 2014).

The term weights are calculated based on the statistical information of words within the document, or semantic weights. The semantic weights scheme utilizes the semantics of classes for indexing. (Luo et al., 2011) stated that the semantics of the class is represented by the senses of terms occurring in the class labels as well as the analysis of the terms by WordNet.

Weighting schemes are divided into two categories based on whether or not the scheme requires the class information of training documents in the classification process. Unsupervised weighting schemes do not use class information to discriminate the term, such as TF, TF-IDF, and its variants. On the other hand, supervised weighting schemes use class information, such as TFIG and TFChi2. Many researchers proposed a term-weighting schemes for text classification, and each of the researchers considered a different measure to express the importance of the term within a document.

2.2.1 Unsupervised Term-Weighting Method

Traditionally, text classification employs binary weight that represents the simplest method, or the TF and the term-weighting schemes of its variants. The TF-IDF is considered the most common scheme for weighting because of its simplicity and efficiency.

Unsupervised term weighting approaches do not include the information of training documents in their calculation. Binary weight is based on whether a word appears in a document (Tsai and Kwee, 2011). The common term frequency approach only considers the raw frequency of a term in a document (Tsai and Kwee, 2011). Another version of the term frequency uses the logarithm operation to deal with the unusual large frequency inside a document Log (1+tf). TF-IDF is a common term-weighting scheme used to represent documents in the vector space model (Erenel et al., 2011).

Essentially, this weighting method comprises two factors. The first factor is TF, which provides the frequency of term t_j in the i^{th} document, thus suggesting that the high-frequency term is a good representative of a particular class. The second factor is the inverse document frequency (IDF), which suggests that a term that appears in various documents should be assigned a small weight.

$$w_{ij} = tf_{ij} * \frac{\log \mathrm{N}}{\mathrm{DFI}} \qquad (1)$$

However, TF-IDF does not consider intra-class or inter-class distribution (Cliao, 2010). The term that equally occurs among all classes is difficult to be discriminatory to a particular class. Nevertheless, the term that occurs equally within the same class will be a good discriminator for this class. In short texts, some terms that appear regularly in many documents belong to the same class. These documents are highly related to this class. Hence, the term should be assigned a high score. However, this situation does not occur with IDF, because this approach considers these terms unimportant.

Researchers proposed term-weighting methods derived from TF-IDF. The TF-IDF was updated to meet the requirements of their problem. A variant of TF-IDF (Martineau et al., 2008) is the delta TF-ID, which determines the difference between TF-IDF scores in positive and negative training data. The experimental result on a movie review dataset running on an SVM classifier shows that the delta TF-IDF achieves a good result and enhances classification accuracy. A variant weighting approach of TF-IDF, which has three definitions, namely, frequency, concentration, and depression, was investigated (Shi et al., 2011). Frequency is the number of occurrences of a term, concentration is the number of documents that include the term, and depression is the number of classes in which the term appears. Unlike TF-IDF, this weighting scheme considers the term distribution in the corpus as a primary factor to determine term importance. Another term weight for automated text classification was proposed (Ren and Sohrab, 2013). The proposed method extends TF-IDF to include class inverse named class-indexing term weighting. According to this measure, a term that appears in many classes cannot be considered a useful discriminator. The proposed approach outperformed TF-IDF when it was tested on benchmark datasets Reuters-21578 and 20 Newsgroups using SVM, Naive Bayes, and Centroid classifiers. Table 2.2 summarizes the most common and state of the art term frequency schemes.

Table 2.2 Unsupervised term weighting schemes.

Denoted by	**Mathematical form**	**Description**
Binary weight	1 Term in the document 0 Term not in the document	1 if the term appears in a document, and 0 denote the absence of the term in the document
TF	TF	Number of times the term occur within the document
Log tf	Log (1+tf)	Log of term frequency
TF-IDF	$w_{ij} = \mathrm{tfij} * \log \frac{\mathrm{N}}{df_i}$	Multiply the TF by the inverse document frequency factor
Prob-idf	$\mathrm{tf} * \log\left(\frac{\mathrm{N} - n_i}{n_i}\right)$	Multiply the TF by the Probability idf.(n_i: the number of document contain the t_i)

2.2.2 Supervised Term-Weighting Method

Debole and Sebastiani (Debole and Sebastiani, 2003) used the supervised learning of text classification to develop the concept of supervised term weighting. Thus, the weight reveals whether a particular term in a document belongs to a particular class or not by using the information on the membership of training corpus (Irani et al., 2010). TF-IDF and feature selection techniques are integrated. The IDF value is substituted with the score selection function as TFChi2, TFIG, and TFRF. Supervised weight can usually be generated by one of the following combinations.

(1) Based on Statistical Confidence Interval (Soucy and Mineau, 2005).

(2) Statistical feature selection, such as OR ratio and Chi2; The task of feature selection is to find the best discriminatory terms in a feature space. Thus, selected terms must have the highest scores (Erenel et al., 2011).

(3) Using the classifier score itself (Han et al., 2001).

Various works in the literature are related to supervised weighting for short-text classification. However, numerous researchers developed supervised weighting schemes for text classification. A brief review of the state-of-the-art supervised weighting approaches on text classification will be provided.

Term weighting based on Statistical Confidence Intervals, a term weighting method (ConfWeight) was presented to weigh features in a vector space model for text classification (Soucy and Mineau, 2005). The weighting method is based on the statistical estimation of word importance for a particular classification problem. By using the Wilson proportion estimate p~, for any class c_j, p~- is the ratio of the document that contains the term t_i in the negative class, and the p~+ is the ratio of the document that contains the term t_i in the positive class. The label MinPos is the low range of the confidence interval, and the label MaxNeg is the high range of the confidence interval.

$$\operatorname{str}(t_i, c_j) = \begin{Bmatrix} \log\left(2 - \frac{\text{minpos}}{\text{minpos+maxpos}}\right), & \text{if minpos} > \text{maxneg} \\ 0, & \text{otherwise} \end{Bmatrix} \tag{2}$$

Using a global policy, the improved version of ConfWeight is set as:

$$\text{ConfWeight}(t_i, d_j) = \log(tf_i + 1) * \max(\text{str}^2) \tag{3}$$

The equation of the weight is similar to that in TF-IDF. The ConfWeight method outperforms TF-IDF and GainRatio (Soucy and Mineau, 2005).

Combine with feature selection approach, Feature selection in text mining primarily aims to determine the best discriminating term. These terms obtain a high score in selection measure. Thus, this selection function measure can be utilized to weigh terms (Debole and Sebastiani, 2003). The term frequency is basically multiplied with the feature selection metrics, which were used to select the most powerful terms, such as Chi2, information gain, and gain ratio. In most cases, TF-IDF is better than this method. Researchers (Lan et al., 2009) proposed a supervised term-weighting scheme that focuses on the contribution of a term to a positive document that belongs to a particular class. The proposed method hypothesizes that a term in the positive document has more discriminative power than that in the negative documents. Only the frequency of the relevant document in which a term occurs is considered. Thus, the weight formula is as follows:

$$\mathrm{w} = \mathrm{tf} * \log(2 + \frac{\mathrm{A}}{\max(1,\mathrm{C})}) \qquad (4)$$

where A and C are obtained from Table 2.1

A supervised term-weighting strategy that considers two elements of term significance in a document (ITD) and the term significance for expressing sentiment (ITS) was proposed (Deng et al., 2014). The proposed mechanism is basically represented as:

$$\mathrm{wij} = \mathrm{ITD}\left(\mathrm{f_i}, \mathrm{d_j}\right) * \mathrm{ITS}(\mathrm{f_i}) \qquad (5)$$

Where ITD (f_i, d_j) is the frequency of feature f_i in document d_j, and ITS(f_i) denotes the value of the feature selection function of term f_i. ITD indicates the frequency of a term, which can be binary frequency (presence or absence), term frequency, or the normalization form of term frequency. ITS uses the filter feature selection function. Several functions, such as Chi2 and information gain, are utilized. The experimental results show that this method outperforms state-of-the-art unsupervised methods.

2.2.3 Short-Text Term Weighting.

Most research related to the short-text classification used the same weighting approach in long-text classification without considering the special nature of short texts or the challenges that adversely affect the performance of a classification task. Many researchers focused on developing solutions that can enhance classification performance