Islamic University of Gaza
Deanery of Higher Studies
Faculty of Information Technology
Information Technology Program

An Approach for Detecting Spam in Arabic Opinion Reviews

By:
Ahmad S. J. Abu Hammad

Supervised By:
Dr. Alaa El-Halees

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master in Information Technology

Rabi'ul-Awwal 1434H – January 2013

﴿رَبَّنَا لا تُؤَاخِذْنَا إِنْ نَسِينَا أَوْ أَخْطَأْنَا.....﴾

[سورة البقرة: 286]

﴿مَن ذَا الَّذِي يُقْرِضُ اللَّهَ قَرْضاً حَسَناً فَيُضَاعِفَهُ لَهُ وَلَهُ أَجْرٌ كَرِيمٌ﴾

[سورة الحديد: 11]

Dedication

To my beloved father…
To my beloved mother…
To my wife…
To my sons ……
To sisters and brothers…
To my best friends…

Acknowledgement

*Praise is to **Allah**, the Almighty for having guided me at every stage of my life.*

*Many thanks and sincere gratefulness go to my **supervisor Dr. Alaa El-Halees**, without his help, guidance, and continuous follow-up; this research would never have been.*

*In addition, I would like to extend my thanks to **the academic staff of the Faculty of Information Technology** who helped me during my master's study and taught me different courses.*

*I would like to thank my **colleagues and classmates** for making my study a great experience, useful, enjoyable, and full of a warm atmosphere.*

*Last but not least, I am greatly indebted to my **family** for their support during my course studies and during my thesis work.*

Ahmed S. J. Abu Hammad

January 2013

Abstract

For the rapidly increasing amount of information available on the Internet, there exists the only little quality control, especially over the user-generated content (opinion reviews, Internet forums, discussion groups, and blogs). Manually scanning through large amounts of user-generated content is time-consuming and sometime impossible. In this case, opinion mining is a better alternative. Although, it is recognized that the opinion reviews contain valuable information for a variety of applications, the lack of quality control attracts spammers who have found many ways to draw their benefits from spamming. Moreover, the spam detection problem is complex because spammers always invent new methods that can't be recognized easily. Therefore, there is a need to develop a new approach that works to identify spam in opinion reviews. We have some in English; we need one in Arabic language in order to identify Arabic spam reviews. To the best of our knowledge, there is still no published study to detect spam in Arabic reviews because it has a very complex morphology compared to English.

In this research, we propose a new approach for performing spam detection in Arabic opinion reviews by merging methods from data mining and text mining in one mining classification approach. Our work is based on the state-of-the-art achievements in the Latin-based spam detection techniques keeping in mind the specific nature of the Arabic language. In addition, we overcome the drawbacks of the class imbalance problem by using sampling techniques. Our approach is implemented using RapidMiner; an open-source machine learning tool and exploits machine learning methods to identify spam in Arabic opinion reviews. The experimental results show that the proposed approach is effective in identifying Arabic spam opinion reviews. Our designed machine learning achieves significant improvements. In the best case, our F-measure is improved up to 99.59%. We compared our approach with other approaches, and we found that our approach achieves best F-measure results in most cases.

Keywords: *Opinion Mining, Arabic Opinion Mining, Spam Review, Spam Detection.*

عنوان البحث

طريقة لاكتشاف النص غير المرغوب فيه في مراجعة الآراء العربية

الملخص

تعاني كثير من البيانات المتزايدة الموجودة على شبكة الأنترنت من مشكلة وجود القليل من طرق مراقبة جودة هذه البيانات، لا سيما عبر المحتوى المقدم. المسح يدويا من خلال كميات كبيرة من المحتوى المقدم من المستخدمين تحتاج وقتا طويلا واحيانا مستحيلة. في هذه الحالة، يعد تنقيب الرأي بديلا افضل. وعلى الرغم من أن الآراء تحتوي على معلومات قيمة لمجموعة متنوعة من التطبيقات، وكذلك قلة طرق مراقبة جودة هذه الآراء يجذب المتطفلين (مرسلي الاستعراض النصي غير المرغوب فيه) الذين وجدوا العديد من الطرق للاستفادة من ارسال الاستعراضات النصية غير المرغوب فيها. وعلاوة على ذلك، فإن مشكلة الكشف عن الاستعراضات النصية غير المرغوبة معقدة لأن المتطفلين يبتكرون أساليب جديدة لا يمكن التعرف عليها بسهولة. ولذلك هناك حاجة ماسة إلى تطوير نهج جديد يعمل على كشف النص غير المرغوب فيه في مراجعة الآراء. لدينا البعض في اللغة الإنجليزية، ونحن بحاجة إلى واحدة في اللغة العربية من أجل كشف الاستعراضات النصية غير المرغوب فيها في مراجعة الآراء العربية. ومع ذلك، فإلى حد علمنا، لا يوجد حتى الآن دراسة نشرت للكشف عن الاستعراضات النصية غير المرغوب فيها في مراجعة الآراء العربية لأنها معقدة جدا مورفولوجيا، مقارنة باللغة الانكليزية.

في هذا البحث، نقترح نهج جديد للكشف عن الاستعراضات النصية غير المرغوب فيها في مراجعة الآراء العربية من خلال الجمع بين أساليب من تنقيب البيانات وتنقيب النص. ويستند عملنا على أحدث ما تم التوصل إليه في هذا المجال للغات اللاتينية الأصل والتي تم دراستها باستفاضة من قبل الباحثين. مع الأخذ في الاعتبار طبيعة وخصائص وسيمات اللغة العربية والثقافات العربية المختلفة. كما سنتغلب في هذا البحث على مشكلة عدم التوازن في توزيع البيانات على الفئات التي تنتمي إليها البيانات. سيتم تطبيق النهج المقترح باستخدام (RapidMiner) أداة مفتوحة المصدر، وباستغلال أساليب آلة التعلم لتحديد الاستعراضات غير المرغوب فيها. النتائج التجريبية تبين أن النهج المقترح قد حقق أعلى معدلات الكشف ودقة التصنيف، حيث كشف الاستعراضات النصية غير المرغوب فيها في مراجعة الآراء العربية بمعدل وصل إلى 99.59%. ثم قارنا نهجنا مع أعمال أخرى ووجدنا نهجنا حقق نتائج أفضل وأكثر دقة في التصنيف في معظم الحالات.

الكلمات المفتاحية: تنقيب الآراء، تنقيب الآراء العربية، الاستعراض النصي غير المرغوب فيه، كشف الاستعراض النصي غير المرغوب فيه.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence.
ATBA	Arabic TBA corpus**.**
ATBAH	ATBA Hotels dataset.
DC	Data Classification/Categorization.
DM	Data Mining**.**
IR	Information Retrieval.
K-NN	K-Nearest Neighbor.
KDD	Knowledge Discovery in Databases.
ML	Machine Learning.
NB	Naïve Bayes.
SDAOR	Spam Detection in Arabic Opinion Reviews.
SVM	Support Vector Machine.
TBA	TripAdvisor Booking Agoda dataset.
TC	Text Classification/Categorization.
TM	Text Mining.

Chapter 1

Introduction

This chapter is an introduction to the thesis, first it gives a brief description of opinion mining, and opinion spamming then it discuss spam detection in Arabic opinion reviews. In addition, it states the thesis problem, the research objectives, the significance of the thesis, the scope and limitation of the thesis work, and the research methodology.

1.1 Opinion Mining

The increased use of the Internet has changed people's behavior in the way they express their views and opinions. Nowadays, the public increasingly participates to express their opinions on the web. They can now post their views using Internet forums, discussion groups, product reviews and blogs, which are collectively called user-generated contents. User-generated contents are written in natural language with an unstructured-free-texts scheme. It provides valuable information that can be exploited for many applications. Manually scanning through large amounts of user-generated contents are time-consuming and sometime impossible [22] [49]. In this case, opinion mining is the better alternative which automatically extracts knowledge from various types of user-generated contents.

Opinion mining (also called sentiment analysis, sentiment mining, sentiment classification, subjectivity analysis, review mining or appraisal extraction) is a subtopic of text mining that it automatically extracts opinions, sentiments, and subjectivity from user-generated contents [17]. Basic task in "opinion mining" is to determine the subjectivity, polarity (positive or negative) and polarity strength (weakly positive, mildly positive, strongly positive, etc.) of a piece of text – in other words: What is the opinion of the writer. Opinion mining has a wide range of applications from different domains such as commercial, government, politics, education and others [51]. For example, many applications of opinion mining include detecting movie popularity from multiple online reviews and diagnosing which parts of a vehicle are liked or disliked by owners through their comments in a dedicated site or forum. There are also applications unrelated to marketing, such as differentiating between emotional and informative social media content [17] [51].

In recent years, opinion mining has become a popular topic for the researchers of Artificial Intelligence (AI). The researchers from AI community have developed various sentiment analysis tasks, including sentiment classification [41] [43], opinion retrieval [30], opinion extraction [10], to serve users' need. All the above studies have the

same assumption: their opinion resources are real and trustful. However, in practice, this opinion information may be spammed.

1.2 Opinion Spamming

Since the opinion information can guide the people's behavior, and on the web, any people can write any opinion text, this can let the people by individuals, and organizations give undeserving spam opinions or reviews to promote or to discredit some target products, services, organizations, individuals, and even ideas without disclosing their true intentions. These spammed opinion information is called opinion spam [45].

Opinion spamming can even be frightening as they can warp opinions and effect on the users' experience. It is safe to say that as opinions are increasingly used in practice, opinion spamming will become more and more rampant and also sophisticated, which presents a major challenge for their detection [16]. However, they must be detected to be a trusted source of public opinions, rather than being full of spam opinions, lies, and deceptions [56].

1.3 Spam Detection in Arabic Opinion Reviews

Some new research efforts in the area of opinion spam detection deal with non-Arabic texts are studied and investigated (more about opinion spam detection in Section 2.9), but in Arabic language, to the best of our knowledge, there is still no published study in this area.

Arabic language is the mother tongue of more than 300 million people; it is considered for religious reasons the language of Islam, and it is ranked as the fifth most spoken language around the world [62]. Arabic, in general, is a challenging language because it has a very complex morphology as compared to English. This is due to the unique nature of Arabic morphological principle, which is highly inflectional and derivational [7] [23]. For example, one word may have more than one lexical category in different contexts. In case of user-generated content, it brings another complexity since most writers on the web express their opinion using local accent instead of standard Arabic language. So, we end up with many written accents instead of one formal language. In addition, many times writers misspelled the words either by accident or deliberately (e.g. for short, and repeating letters to insistent in some words) [23].

Thus, spam detection in Arabic opinion reviews is complex. So, there is a great need to develop a new approach that works to detect spam in Arabic spam opinion reviews.

In this research, we propose an approach called **S**pam **D**etection in **A**rabic **O**pinion **R**eviews (SDAOR) that uses combining methods from data mining (more about data mining in Section 2.2) and text mining (more about data mining in Section 2.4) in one mining classification approach. The SDAOR will be designed to detect spam in the Arabic opinion reviews.

1.4 Problem Statement

The problem in this research is how to detect spam in the Arabic opinion reviews that can be valid for some proper domain with accepted accuracy and F-measure. However, there are some challenges in this area.

The sub problems we face are:

1. What are the proper domains that we shall use to show the power of our proposed approach? And, how to collect spam Arabic opinion reviews data sets?
2. What types of spam available in Arabic opinion reviews?
3. What is the proper approach to produce an approach to analyze and detect spam in Arabic opinion reviews?
4. What are the proper preprocessing steps are performed on the data sets before it is used?
5. What are the most relevant features to be extracted that related to reviews, reviewers and products to extract the most appropriate classification?
6. What are the best mining methods to be used to score and classify the spam Arabic opinion reviews?
7. How to evaluate the performance of the proposed approach?

1.5 Objectives

1.5.1 Main Objective

The main objective of this research is to develop an approach that detects spam in the Arabic opinion review that can be valid for some proper domain in an accurate way.

1.5.2 Specific Objectives

The specific objectives of this research are:

1. Search for a domain that can be widely used, and in the important area which can give a good evaluation environment. Build data sets of selected domain.
2. Find language resources can be used or extended to detect spam in the Arabic opinion reviews.
3. Identify different types of spam available in Arabic opinion reviews.
4. Find the most proper approach that can be produced an approach to analyze and detect spam in Arabic opinion reviews.
5. Find the proper preprocessing steps on the spam Arabic opinion data set before use it.
6. Find the most relevant features that related to reviews, reviewers and products to extract the most appropriate classification.
7. Select more than one classification method to find a way to classify the Arabic opinion reviews, whether it should be spam and non-spam.
8. Evaluate the results and try to increase the ratio of precision and recall measurements. In addition, compare our proposed approach with other existing non-Arabic methods.

1.6 Significance of the Thesis

The significance of this thesis is:

1. In the Arab world, individuals and businesses are increasingly using reviews for their decision making. It is critical to detect spammers who wrote spam reviews.
2. Saving efforts and time by helping the user, such as: business to identify spam in Arabic opinion reviews quickly.
3. There are some of researches for opinion spam detection deal with non-Arabic texts, but in Arabic, there is still no published study in this area.
4. Demonstrate that there are different types of spam available in Arabic opinion reviews and identify them.
5. Unfortunately, many data sets in real applications (such as health examination, inspection, spam identification and text mining, and credit fraud detection)

involve imbalanced class distribution problem. In this research, our approach able to deal with an imbalanced data problem.

1.7 Research Scope and Limitation

This research proposes a SDAOR approach. The work is applied with some limitations and assumption such as:

1. Our work is limited to detecting spam in Arabic opinion reviews only. We will not include detecting spam in opinion reviews in other languages, such as: English and European in the same review.
2. We cannot work in all domains such as: commercial, news, sports, and politics, so we will select the most popular one.
3. We cannot apply all classification methods; we will use the most famous ones.
4. The manual evaluation technique is going to be followed because there isn't any automatic evaluation for detecting spam in Arabic opinion reviews.
5. The spammer that registers multiple times (using different user-IDs) at a site by using different machines and write multiple spam reviews under these user-IDs, we will not detect.
6. Spammers write either only positive reviews on his/her own products or only negative reviews on the products of his/her competitors, but not both. This, we will not detect.
7. Members of the group write reviews at random intervals to hide spikes, we will not detect.
8. If the group is sufficiently large, it may be divided into sub-groups so that each sub-group can spam at different websites (instead of only spam at the same site), we will not detect.

1.8 Research Methodology

To accomplish the objectives of the research, the following methodology will be followed (see Figure 1.1):

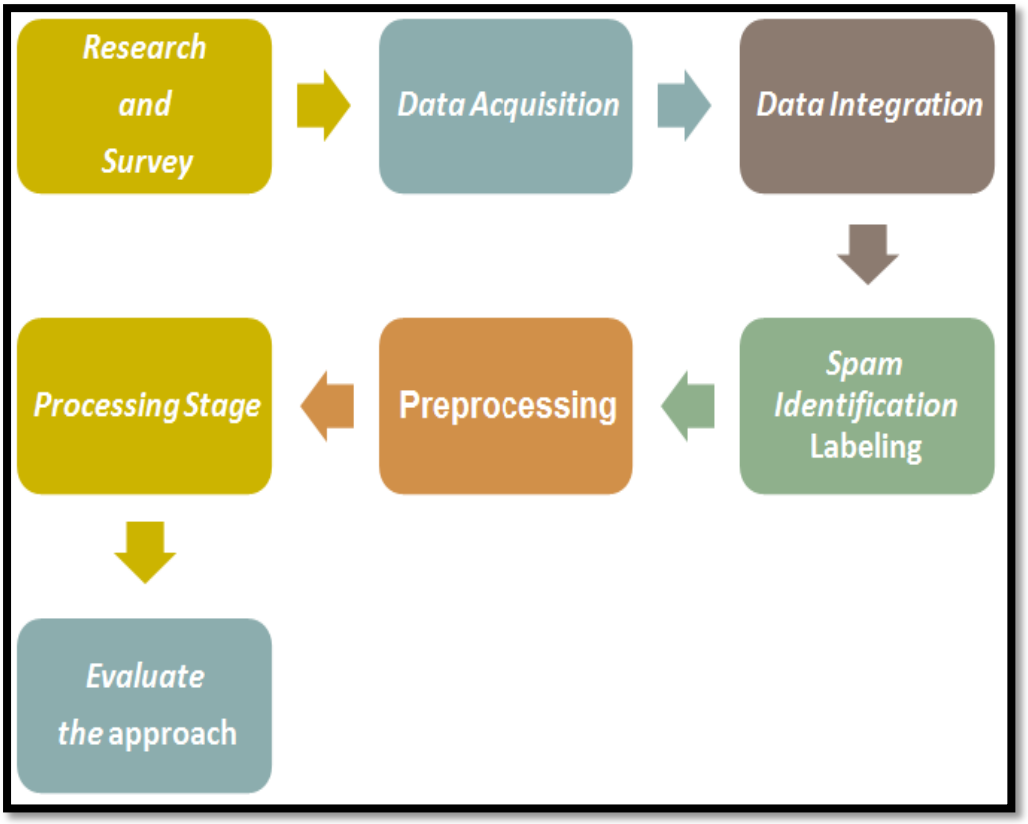


Figure 1.1: Overall Methodology [25].

- ***Research and Survey:*** include reviewing the recent researches closely related in the thesis problem statement and the research question. After analyzing the existing methods in spam detection for non-Arabic opinion reviews, identifying the drawbacks or the lack of existing approaches, we formulate the strategies and solutions how to use or extend in order to be overcome in our research.
- ***Data Acquisition:*** in this step, we will build an in-house data set of spam reviews and reviewers using human collected from online Arabic economic websites, with different characteristics and sizes by crawls. Its records chosen randomly from among any of the records that available on the website.
- ***Data Integration:*** in this step, we will integrate data from multiple source data sets into a coherent form.
- ***Spam Identification Labeling:*** in this step, we will identify different types of the spam in the integrated data set, and manually labeled each record with spam and non-spam. We will explain different types of the spam in detail in Chapter 4.

- ***Preprocessing:*** in this step, we will apply a number of preprocessing techniques to deal with noisy, missing, and inconsistent data. There are a number of preprocessing techniques such as: cleaning, transformations, reduction, tokenization, Arabic stopword removal, and light stemming [14]. We will explain using preprocessing techniques for data features, text features and data-text features in detail in Chapter 4.
- ***Processing Stage:*** to do this step, we implement the following sub-steps:
    - Data mining classification experiments.
    - Text mining classification experiments.
    - Data-Text mining classification experiments.

  Then we apply each previous step by more than one classification method. The structure of our approach, training, testing, and extracting the results will be explained in Chapter 5.
- ***Evaluate the Approach:*** in this step, we will analyze the obtained results and justify the feasibility of our approach by comparing it with other approaches.

1.9 Thesis Structure

This thesis consists of six mainly chapters, which are structured around the objectives of the research. The main points discussed throughout the chapters are listed below:

- ***Chapter 1 Introduction:*** It gives a short introduction about spam detection in Arabic opinion reviews, the thesis problem and objectives.
- ***Chapter 2 Theoretical Foundation:*** It describes the theory needed for thesis work, Knowledge Discovery in Databases (KDD) and its disciplines, namely: Data Mining (DM) and Text Mining (TM), major kinds of classification algorithms which are used in our research: Naïve Bayes (NB), ID3, K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM), and imbalance class distribution problem. In addition, this chapter describes details about opinion mining and opinion spam detection.
- ***Chapter 3 Related Works:*** It presents other works related to the thesis.
- ***Chapter 4 An Approach for Detecting Spam in Arabic Opinion Reviews:*** It includes the methodology steps and the architecture of the SDAOR. An explanation about the data sets used in the experiments, identification data set

label (spam and non-spam), preprocessing of these data sets, and the experiment cases are included as well.

- ***Chapter 5 Experimental Results and Evaluation:*** It gives in detail about the sets of experiments, and analyzes the experimental results. In addition, it gives a discussion for each set experiment. Then, it produces some experiments to comparison goals.
- ***Chapter 6 Conclusions and Future Work:*** It discusses the final conclusions and presents possible future works.

Chapter 2

Theoretical Foundation

In this chapter, the fundamental concepts which represent the basis for understanding of the thesis work are presented. First, the Knowledge Discovery in Databases (KDD) is introduced, followed by Data Mining (DM), Data Classification (DC), Text Mining (TM), Text Classification (TC), major kinds of classification algorithms, which are used in our research: Naïve Bayes (NB), ID3, K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM), the imbalance class distribution problem, and finally discusses opinion mining and opinion spam detection.

2.1 Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) has become a popular area of research and development in computer science [3]. The concept of knowledge discovery was first introduced by ***Frawley et. al.*** in [25] as "knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data". The current interest in KDD is fuelled by the large amount of available data for any particular application-domain considered, such as bioinformatics, e-commerce, marketing and sales financial investments, and geography it currently stored electronically [25].

KDD, with its goal of recognizing patterns within large volumes of data is a tool with the potential to produce new unknown knowledge [14].

The KDD process is a series of interactive steps to achieve the goal of finding useful knowledge from large amounts of raw data. The process is designed to be iterative: any sequence of steps may be refined and re-executed several times [25]. Figure 2.1, illustrates the main stages of the KDD process.

Once a data set is produced, data cleansing and pre-processing occur. The important pre-requisite steps are performed such as removing noisy data, handling missing values and outliers and data type correction, which may impact the performance and quality of the final result [14] [25]. The next step in the process is data reduction, where a subset of the overall data is selected based on its relevance to the data mining task. Data reduction is of critical importance in speeding up mining algorithms to acceptable performance levels, especially in cases where data may contain a large number of attributes, or the data set is in the order of several million records [25] [29].

After data cleansing and reduction, a data set is produced and ready to be mined, and work can begin in choosing a data mining task, based on project goals. Also, exploratory analysis of the data set can take place; this provides further insights on the nature of the data, helps in determining the data mining tasks and provides early feedback on collected data to business users. Any corrective action can take place by re-executing earlier stages in the process. Finally, the data mining stage is executed, where data mining algorithms are applied to the data set based on criteria determined on previous stages. Then, the results are evaluated and interpreted, and it is likely that this will lead to several iterations of the mining stage, so that algorithms can be fine-tuned and hypothesis can be confirmed. Lastly, by reviewing the results of the data mining exercise with business users, these can be used as new knowledge and acted upon in their relevant business context [3] [14] [25] [29].

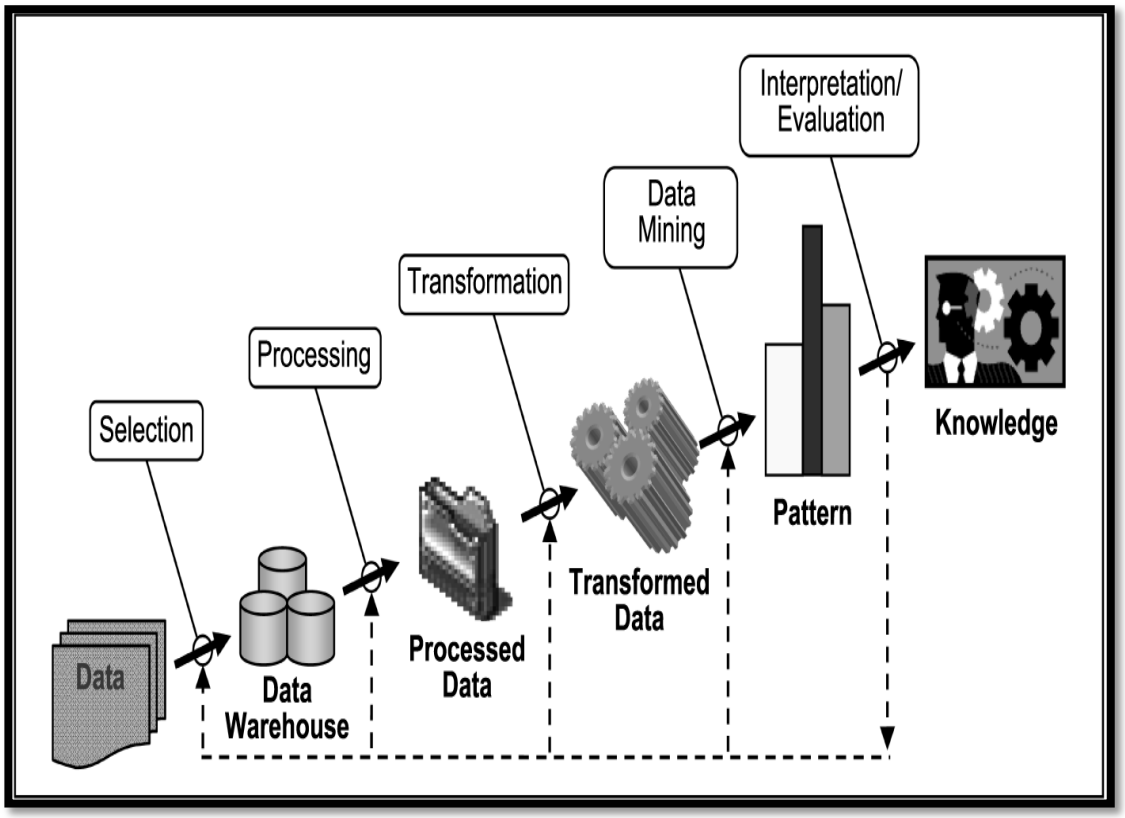


Figure 2.1: Stages of the KDD Process [25].

The KDD Disciplines

Corresponding to the variety of data formats, KDD research can be divided into different "disciplines", i.e. data mining, text mining, graph mining, image mining, web mining, and music mining [25].

- **Data Mining:** this discipline encompasses generic techniques, generally described in terms of data set like data, although often adaptable to other forms of data [9]. Some of the work described in this thesis makes use of the techniques espoused by this filed therefore we will further present it in Section 2.2.
- **Text Mining:** this KDD discipline deals with various forms of electronic textual data [26]. As text mining is central to the theme of this thesis a detailed review is provided in Section 2.4.
- **Graph Mining:** this discipline specializes on mining data represented in the form of graphs. Graph mining may be categorized into transaction graph mining, which searches for patterns in sets of graphs, or single graph mining, which looks for patterns within a single large graph [67].
- **Image Mining:** electronic images such as satellite images, medical images, and digital photographs are the data to be manipulated in this research discipline. ***Hsu et. al.*** in [26] classify into two types: (i) image mining that involves domain-specific applications; and (ii) image mining that involves general applications. Research topics include: satellite image (remote sensing) mining, medical image mining, image classification, image clustering, and image comparison [39].
- **Web Mining:** which concentrates on detecting hidden knowledge from web like data. Three common types of web like data can be identified: web page contents, web hyperlink structures, and users' usage data (server logs). As a consequence, research areas in web mining can be grouped into three divisions/sections: content mining, structure mining, and usage mining. Web content mining is closely related to text mining since web pages usually contain a significant amount of text [14].
- **Music Mining:** electronic music such as MIDI, PCM and MP3 are the data required by this research discipline. One research aspect is music genre classification the automated assignment of "unseen" digital music records into pre-defined musical genres [12].

In this thesis, only data mining and text mining will be used and hence further considered.

2.2 Data Mining (DM)

Data mining (DM) is an essential step in the process of knowledge discovery. It is the process of extracting knowledge hidden from large volumes of raw data. The knowledge must be new, not obvious, and one must be able to use it [34].

DM is "a multidisciplinary field, drawing work from areas including: database technology, machine learning, statistics, pattern recognition, neural networks, knowledge-based system, artificial intelligence, high performance computing, and data visualization" [9]. DM techniques have been widely applied in bioinformatics, geography, marketing and sales studies, e-commerce, and financial studies [20] [34].

The aim of data mining is to learn a model for the data. Data is the input information to be mined or visualized. Different types of data sets possible, e.g. flat file data (attribute, value), and spatial data (locations, value). Data mining techniques vary with the type of a data set. The majority of data mining techniques work on flat file data. A flat file data collection of instances of something and each instance is described using a finite set of attributes [6] [20].

DM functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive mining and predictive mining. Descriptive mining tasks characterize the general properties of the data in the database such as association rule and clustering. Predictive mining tasks perform inference on the current data in order to make predictions such as classification, prediction and outlier analysis [6] [34].

- **Association rules:** are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. Association analysis is commonly used for market basket analysis [34] [69].
- **Classification:** is the organization of data in given classes. Also, known as supervised classification, the classification uses given class labels to order the

objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects [25] [69]. Data Classification is central to the theme of this thesis and is, therefore, further discussed in Section 2.3.

- **Prediction:** has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. Prediction is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is, however, most often referred to the forecast of missing numerical values, or increase/ decrease trends in time-related data. The major idea is to use a large number of past values to consider probable future values [63].
- **Clustering:** is a division of data into groups of similar objects. It is similar to the classification. However, unlike classification, in clusters, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification because the classification is not dictated by giving class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity) [34] [63].
- **Outlier analysis:** outliers are data elements that cannot be grouped in a given class or cluster. Also known as exceptions or surprises, they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable [9].

Since opinion mining can be seen as a classification problem where we can classify opinion as non-spam or spam, in this thesis, only the data classification (more detailed about the data classification will discuss in Section 2.3) will be considered.

2.3 Data Classification (DC)

Data Classification/Categorization (DC) is a classic data mining task. The classification is a supervised learning task that estimates the correct classes of objects [9]. In general, there is a two-step process for DC. In the first step; a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or "learning from" a training set made up of database tuples and their associated class labels. In the second step, the model is used for classification; the predictive accuracy of the classifier is estimated using the training set to measure the accuracy of the classifier. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. The associated class label of each test tuple is compared with the learned classifier's class prediction for that tuple. If the accuracy of the classifier is considered acceptable, the classifier can be used to classify future data tuples for which the class label is not known [20] [34].

2.4 Text Mining (TM)

Text mining (TM), sometimes alternately referred to as text data mining, - an increasingly important field of research in KDD - applies data mining techniques to textual data collections, and "aims at disclosing the concealed information by means of methods which on the one hand are able to cope with the large number of words and structures in natural language and on the other hand allow to handle vagueness, uncertainty and fuzziness" [26] [38].

TM usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output [25].

Typical textual data include natural language speeches (e.g. dialogues, argumentations), text files (e.g. magazine articles, academic papers), and web documents (e.g. web news, e-mails). In text mining, a given textual data collection is commonly refined in document based fashion. A document based (i.e. a set of electronic documents) usually consists of thousands of documents [26].

For mining large document based it is necessary to pre-process the text documents and store the information in a data structure, which is more appropriate for further processing than a plain text file. Even though, meanwhile several methods exist that try to exploit also the syntactic structure and semantics of text, most TM approaches are based on the idea that a text document can be represented by a set of words, i.e. a text document is described based on the set of words contained in it [26] [38].

TM can be specified in a number of sub tasks - components of a larger text-mining effort - typically include: text classification (more detailed about text classification will discuss in Section 2.5), text clustering, document summarization, and opinion mining [34] [53].

2.5 Text Classification (TC)

Text Classification/Categorization (TC) is the task of assigning one or more predefined categories of natural language text documents, based on their contents. This task that falls at the crossroads of Information Retrieval (IR) and Machine Learning (ML), has witnessed a booming interest in the last ten years from researchers and developers alike [26] [38].

TC can provide conceptual views of document collections and has important applications in the real world. For example, news stories are typically organized by subject categories (topics) or geographical codes; academic papers are often classified by technical domains and sub-domains; patient reports in healthcare organizations are often indexed from multiple aspects, sorting of files into folder hierarchies, topic identifications, dynamic task-based interests, automatic meta-data organization, text filtering and documents organization for databases and web pages [20] [26] [34]. Another widespread application of text categorization is spam filtering, where email messages are classified into the two categories spam and non-spam [26] [34].

2.6 Classifiers

In this section we describe major kinds of classification algorithms which are used in our research: Naïve Bayes (NB), ID3, K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM).

2.6.1 Naïve Bayes (NB) Classifier

Bayesian classifiers are statistical classifiers based on Bayes theorem. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class [34].

A NB classifier assumes that the effect of an attribute value of a given class is independent of the values of the other attributes. This assumption is called class conditional independence. The NB classifier, works as follows [34] [63]:

- Let **D** be training set of tuples and their associated class labels. As usual, each tuple is represented by a n-dimensional attribute vector, **X = (x1, x_2,….x_n)**, n measurements made on the tuple from **n** attribute, respectively, **A_1,A_2,…, A_n.**
- Assume that there are m classes, **C_1, C_2,…, C_m**. Given a tuple, **X**, the classifier will predict that **X** belongs to the class having the highest probability, conditioned on **X**. That is, the NB classifier predicts that tuple **X** belongs to the class **C_i** if and only if

$$P(C_i|X) > P(C_j|X) \quad \text{for } 1 \leq j \leq m, j \neq i \text{ ..... 2.1}$$

Thus we maximize **$P(C_i|X)$**. The class **C_i** for which **$P(C_i|X)$** is the maximized is called the maximum posteriori hypothesis. By Bayes' theorem (Equation 2.2),

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{p(X)} \text{.....2.2}$$

- As **P(X)** is constant for all classes, only **$P(X|C_i)$ $P(C_i)$** needs maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equal.
- Based on the assumption is that attributes are conditionally independent (no dependence relation between attributes), **$P(X|C_i)$** using Equation 2.3.

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \text{.....2.3}$$

Equation 2.3 reduces the computation cost, only counts the class distribution. If $\mathbf{A_k}$ is categorical, $\mathbf{P(X_k|C_i)}$ is the number of tuples in $\mathbf{C_i}$ having value $\mathbf{x_k}$ for $\mathbf{A_k}$ divided by $\mathbf{|C_i, D|}$ (number of tuples of C_i in D). And if $\mathbf{A_k}$ is continuous-valued, $\mathbf{P(x_k|C_i)}$ is usually computed based on a Gaussian distribution with a mean μ and standard deviation **σ** and $\mathbf{P(X_k|C_i)}$ is

$$P(X|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \text{.....2.4}$$

$$g(x_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}} \text{.....2.5}$$

Where $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\sigma^2}$ is the variance. If an attribute value doesn't occur with every class value, the probability will be zero, and a posteriori probability will also be zero.

NB classifier is fast, accurate, simple, and easy to implement, thus chosen to be one of the classifiers in this case. It is based on a simplistic assumption in real life and is only valid to multiply probabilities when the events are independent. Despite its naïve nature, NB classifier actually works well on actual data sets [34]. It is chosen to be used in this thesis.

2.6.2 ID3 Classifier

ID3 is a decision tree classifier, where "ID" stands for "Interactive Dichotomizer" and "3" stand for "version 3" is a rooted tree containing nodes and edges. Each internal node is a test node and corresponds to an attribute. The edges going out of a node correspond to the possible values of that attribute. The ID3 algorithm works as follows. The tree is constructed top-down in a recursive fashion. At the root, each attribute is tested to determine how well it alone classifies the samples. The "best" attribute is then chosen and the samples are partitioned according to this attribute. The ID3 algorithm is then recursively called for each child of this node, using the corresponding subset of data [14] [20]. The ID3 algorithm, works as shows in Figure 2.2, given a set of

non-categorical attributes **C1, C2, .., Cn**, the categorical attributes **C**, and a training set **T** of records.

```
function ID3 (R: a set of non-categorical attributes,
              C: the categorical attribute,
              S: a training set) returns a decision tree;
begin
     If S is empty, return a single node with value Failure;
     If S consists of records all with the same value for
        the categorical attribute,
        return a single node with that value;
     If R is empty, then return a single node with as value
        the most frequent of the values of the categorical attribute
        that are found in records of S; [note that then there
        will be errors, that is, records that will be improperly
        classified];
     Let D be the attribute with largest Gain(D,S)
        among attributes in R;
     Let {dj| j=1,2, .., m} be the values of attribute D;
     Let {Sj| j=1,2, .., m} be the subsets of S consisting
        respectively of records with value dj for attribute D;
     Return a tree with root labeled D and arcs labeled
        d1, d2, .., dm going respectively to the trees

          ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), .., ID3(R-{D}, C, Sm);
end ID3;
```

Figure 2.2: The ID3 Algorithm [20].

ID3 trees are a classic way to represent information from a machine learning algorithm, and offer a fast and powerful way to express structures in data. ID3 trees require little data preparation and are easy to understand and interpret. It uses a white box model, which contrasts to a black box model such as artificial neural network, meaning that the explanation for the condition is easily explained by Boolean logic [14] [20]. It is also chosen to be used in this thesis.

2.6.3 K-Nearest Neighbor (K-NN) Classifier

K-Nearest Neighbor (K-NN) algorithm is one of the supervised learning algorithms that have been used in many applications in the field of data mining, statistical pattern recognition and many others. It follows a method for classifying objects based on closest training examples in the feature space [9].

An object is classified by a majority of its neighbors. **K** is always a positive integer. The neighbors are selected from a set of objects for which the correct classification is known. The K-NN algorithm is as follows [9] [14]:

1. Determine the parameter **K** i.e., the number of nearest neighbors beforehand.
2. The distance between the query-instance and all the training samples is calculated using Euclidean distance. Euclidean distance between two points, **X = (x1, x_2,….x_n)** and **Y = (y1, y_2,….y_n)** is:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \text{.....2.6}$$

3. Distances for all the training samples are sorted and nearest neighbor based on the K-th minimum distance is determined.
4. Since the K-NN is supervised learning, get all the categories of your training data for the sorted value which fall under **K**.
5. The predicted value is measured by using the majority of nearest neighbors.

K-NN works well even when there are some missing data. K-NN is good at specified which predictions have low confidence. It has some strong consistent results. As the amount of data approaches infinity, the algorithm is guaranteed to yield an error rate no worse than twice the Bayes error rate (the minimum achievable error rate given the distribution of the data) [16].

2.6.4 Support Vector Machine (SVM) Classifier

A support vector machine (SVM) is a set of related supervised learning methods used for classification and regression. In simple words, given a set of training examples, each marked as belonging to one of two categories, the SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. Intuitively, SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on [18] [28] [34].

More formally, SVM constructs a hyperplane or a set of hyperplanes in a high dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier [18] [28] [34].

Currently, SVM is widely used in object detection and recognition, content-based image retrieval, text recognition, biometrics, speech recognition, speaker identification, benchmarking time-series prediction tests. Using SVM in text classification is proposed by [47], and subsequently used in [19] [37].

Equation 2.7 is dot product formula and used for the output of linear SVM, where **x** is a feature vector of classification documents composed of words, **w** is the weight of corresponding **x**, and **b** is a bias parameter determined by training process.

$$y = w.X - b \quad \text{..... 2.7}$$

The following summarizes SVM steps:

1. Map the data to a predetermined very high-dimensional space via a kernel function.
2. Find the hyperplane that maximizes the margin between the two classes.
3. If data are not separable find the hyperplane that maximizes the margin and minimizes the (a weighted average of the) misclassifications.

SVM can be used for both linear and nonlinear data. It uses a nonlinear mapping to transform the original training data into a higher dimension. With the new dimension, it searches for the linear optimal separating hyperplane (i.e., "decision boundary"). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. SVM finds this hyperplane using support vectors ("essential" training tuples) and margins (defined by the support vectors). Figure 2.3 shows support vectors and how margins are maximized [18] [28] [34].

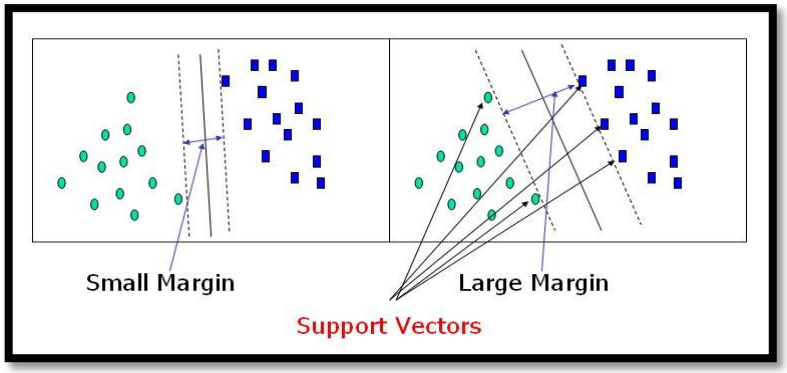


Figure 2.3: Support Vectors [18].

SVM is effective for high dimensional data because the complexity of trained classifier is characterized by the number of support vectors rather than the dimensionality of the data, the support vectors are the essential or critical training examples, they lie closest to the decision boundary, if all other training examples are removed and the training is repeated, the same separating hyperplane would be found. The number of support vectors found can be used to compute an (upper) bound on the expected error rate of the SVM classifier, which is independent of the data dimensionality. Thus, an SVM with a small number of support vectors can have good generalization, even when the dimensionality of the data is high [11] [18] [28] [34].

SVM classifiers have been considered state-of-the-art for sentiment classification by many research papers [9] [47] [59]. Practical real-world modelers have found that SVM have performed well when other classifiers did poorly. SVM classifiers have been widely used in text classification tasks with unbalanced training. It is also chosen to be used in this thesis.

2.7 Imbalance Class Distribution Problem

The classification techniques usually assume a balanced class distribution (i.e. their data in the class are equally distributed). Usually, a classifier performs well when the classification technique is applied to a data set evenly distributed among different classes. But many real applications face the imbalanced class distribution problem. In this situation, the classification task imposes difficulties when the classes present in the training data are imbalanced [8].

The imbalanced class distribution problem occurs when one class is represented by a large number of examples (majority class) while the other is represented by only a few (minority class). In this case, a classifier usually tends to predict that samples have the majority class and completely ignore the minority class. This is known as the class imbalance problem [13] [31]. Figure 2.4 illustrates the idea of the class imbalance problem where a minority class is represented by only 1% of the training data and 99% for majority class.

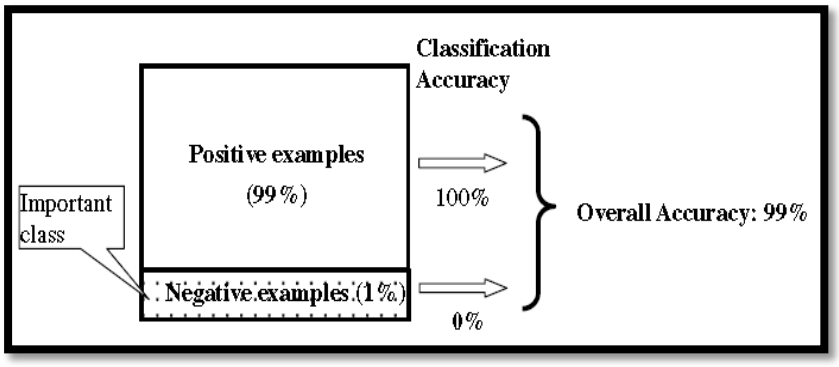


Figure 2.4: The Illustrated of Class Imbalance Problems [13].

Unfortunately, this problem is very pervasive in many domains. For example, with text classification tasks whose training sets typically contain many fewer documents of interest to the reader than on irrelevant topics. Other domains suffering from class imbalances include target detection, fault detection, or credit card fraud detection problems, disease diagnosis, bioinformatics, oil-spill detection and many other areas, which contain much fewer instances of the event of interest than of irrelevant events [25].

Class imbalanced presents several difficulties in learning, including imbalances in class distribution, lack of data and concept complexity. There are many techniques to handle imbalanced in class distribution, sampling technique are famous once. Sampling technique is applied to balance class distribution [21].

Sampling Techniques

A popular way to deal with the class imbalance problem is sampling. Sampling methods modify the distributions of the majority and minority class in the training data set to obtain a more balanced number of instances in each class [8]. To minimize class imbalance in training data, there are two basic methods, under-sampling and over-sampling.

- **Under-Sampling:** it removes data from the original data set by randomly select a set of majority class examples and then remove this sample [31]. Hence, an under-sample approach is aimed to decrease the skewed distribution of majority class and minority class by lowering the size of majority class [68]. Under-Sampling is suitable for large application where the number of majority samples is very large and lessening the training instances reduces the training time and storage [8] [40].

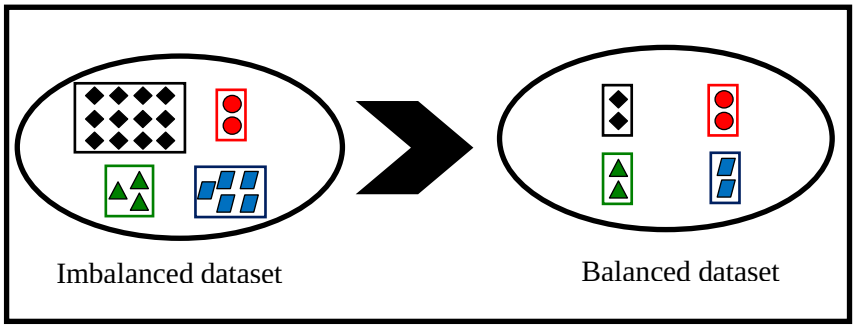


Figure 2.5: The Distribution of Samples Before and After Applying Under-Sample Approach [13].

Figure 2.5 illustrates the distribution of samples in a data set before and after applying under-sample approach. For example, from the Figure 2.5 we find the circle is represented minority class which has two instances. So, for this reason we take randomly only two instances from other shapes which are represented majority classes in this case. The drawback of this technique is that there does not exist a control to remove patterns of the majority class, thus it can discard data potentially important for the classification process [32], which degrade classifier performance.

- **Over-Sampling:** it is a method to add a set of sampled from the minority class by randomly select minority class examples and then replicating the selected examples and adding them to the data set [31]. It is different from under-sample approach so there is no information is lost, all instances are employed. However, the major problem of this technique leads to a higher computational cost [21] [40].

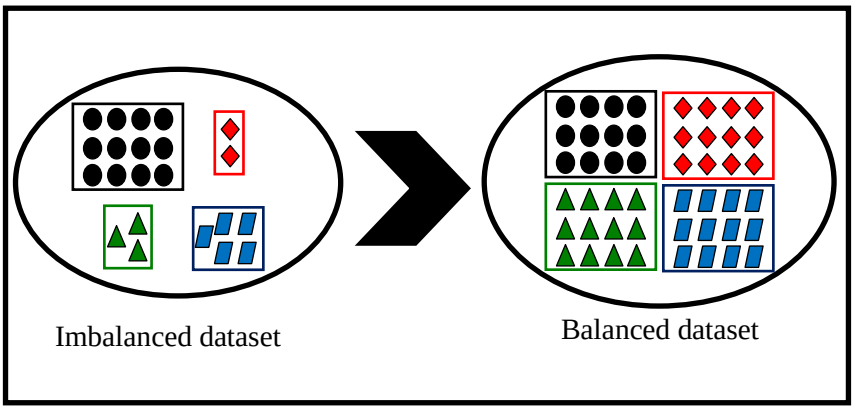


Figure 2.6: The Distribution of Samples Before and After Applying Over-Sample Approach [10].

Figure 2.6 illustrates the distribution of samples in a data set before and after applying over-sample approach. For example, from the Figure 2.6 we find the circle is represented majority class which has twelve instances. So, for this reason we replicate instances from other shapes which are represented minority classes until they reach to twelve instances approximately in this case.

2.8 Opinion Mining

As discussed in Section 1.1, opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [14].

Opinion mining discovers opinionated knowledge at different levels such as at word, sentence or document level.

- **Word Level:** opinion mining at word level is a task of determining positive or negative sentiment of certain word in certain contexts or domain. Words that encode a desirable state (e.g., "ممتاز" (excellent)) have a positive orientation, while words that represent an undesirable state have a negative orientation (e.g.,"سيئ" (bad)). To apply opinion mining, researchers have compiled sets of words and phrases for adjectives, adverbs, verbs, and nouns. Such lists are collectively called the opinion lexicon [65].
- **Sentence Level:** opinion mining at sentence level classifies each sentence as expressing a positive or a negative opinion. It is an action that can be associated

with two tasks. Initial work is to identify whether the sentence is subjective (opinionated) or objective. The second task is to classify a subjective sentence and determine if it is positive, negative or neutral. Two approaches can be used to my opinions at sentence level. Either based on word level or using sentence structure [1].

- **Document Level:** document opinion analysis is about classifying the overall sentiments expressed by the authors in the entire document text. The task is to determine whether a document is positive, negative or neutral about a certain object [57]. Document level polarity categorization attempts to classify sentiments in movie reviews, news articles, or web forum postings.

For mining large document collections it is necessary to pre-process the text documents and store the information in a data structure, which is more appropriate for further processing than a plain text file. Most text mining approaches are based on the idea that a text document can be represented by a set of words. In order to obtain all words that are used in a given text, a *tokenization* process is required, i.e. a text document is split into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters with single white spaces. This tokenized representation is then used for further processing. The set of different words obtained by merging all text documents of a collection is called the dictionary of a document collection [38].

In order to reduce the size of the dictionary and thus the dimensionality of the description of documents within the collection, the set of words describing the documents can be reduced by *filtering* and *lemmatization* or *stemming* methods [38] [52].

- **Filtering Methods:** remove words from the dictionary and thus from the documents. A standard filtering method is stopword filtering. The idea of stopword filtering is to remove words that bear little or no content information, like articles, conjunctions, prepositions, etc. [38].
- **Lemmatization Methods:** try to map verb forms to the infinite tense and nouns to the singular form. However, in order to achieve this, the word form has to be known, i.e. the part of speech of every word in the text document has to be assigned. Since this tagging process is usually quite time-consuming and still error-prone, in practice frequently stemming methods are applied [38].

- **Stemming Methods**: try to build the basic forms of words. A stem is a natural group of words with equal (or very similar) meaning. The stemming algorithm is a computational process that gathers all words that share the same stem and has some semantic relation [60]. The main objective of the stemming process is to remove all possible affixes and thus reduce the word to its stem. After the stemming process, every word is represented by its stem [38] [52]. Stemming is needed in many applications such as natural language processing, compression of data, and information retrieval systems. Many stemmers have been developed for English and other European languages. These stemmers mostly deal with the removal of suffixes as this is sufficient for most information retrieval purposes. Some of the most widely known stemmers for English are Lovins [57] and Porter [61] stemming algorithms. Most Arabic language stemming approaches fall into three classes: root based stemming, light stemming and statistical stemming [57] [60]. Root-Based stemmers use morphological analysis to extract the root of a given Arabic word. Light stemming refers to the process of stripping off a small set of prefixes and/or suffixes without trying to deal with infixes or recognize patterns and find roots. Statistical stemmers attempt to group words variances using clustering techniques [57].

Opinion mining faced some of the challenges which may provide more lively interest to the researchers to further carry out in this area. They are: opinion spam detection, dealing of colloquial terms, and feature extraction. In this research, we discuss an opinion spam detection problem.

2.9 Opinion Spam Detection

Opinions of e-commerce websites are increasingly used by individuals and organizations for making purchase decisions and for marketing and product design. Positive opinions often mean profits and fames for businesses and individuals, which, unfortunately, give strong incentives for people to game the system by posting spam opinions or reviews to promote or to discredit some target products, services, organizations, individuals, and even ideas without disclosing their true intentions, or the person or organization that they are secretly working for. Such individuals are called opinion spammers and their activities are called opinion spamming [5] [55].

Opinion spamming about e-commerce issues can even be frightening as they can warp opinions and highly negative impact about a particular product. It is safe to say that as opinions on e-commerce websites are increasingly used in practice, opinion spamming will become more and more rampant and also sophisticated, which presents a major challenge for their detection. However, they must be detected in order to ensure that the e-commerce websites continues to be a trusted source of public opinions, rather than being full of fake opinions, lies, and deceptions [56].

Spam detection in general has been studied in many fields. Web spam and email spam are the two most widely studied types of spam [45]. Opinion spam is, however, very different. There are two main types of Web spam, i.e., link spam and content spam. Link spam is spam on hyperlinks, which hardly exist in reviews. Although advertising links are common in other forms of social media, they are relatively easy to detect. Content spam adds popular (but irrelevant) words on target Web pages in order to fool search engines to make them relevant to many search queries, but this hardly occurs in opinion postings [45] [72]. Email spam refers to unsolicited advertisements, which are also rare in online opinions [45] [71].

The ultimate goal of this research is to develop an approach that detects spam in Arabic opinion reviews. Detecting spam in Arabic opinion reviews can be formulated as a classification problem with two classes, spam and non-spam.

2.9.1 Type of Spam and Spamming

Three types of spam reviews were identified in [44] [56]:

1. **Type 1 (untruthful opinions):** these are spam reviews that are written not based on the reviewers' genuine experiences of using the products or services, but are written with hidden motives. They often contain undeserving positive opinions about some target entities (products or services) in order to promote the entities and/or unjust or false negative opinions about some other entities in order to damage their reputations.
2. **Type 2 (reviews about brands only):** these reviews do not comment on the specific products or services that they are supposed to review, but only comment on the brands or the manufacturers of the products. Although they

may be genuine, they are considered as spam as they are not targeted at the specific products and are often biased.

3. **Type 3 (non-reviews):** these are not reviews, which have two main sub-types: (1) advertisements and (2) other irrelevant reviews containing no opinions (e.g., questions, answers, and random text).

Spam reviews may be written by many types of people, e.g., friends and family, company employees, competitors, businesses that provide spam review writing services, and even genuine customers (some businesses give discounts and even full refund to some of their customers on the condition that the customers write positive reviews for them) [17] [44] [56].

In general, a spammer may work individually, or knowingly or unknowingly work as a member of a group (these activities are often highly secretive) [17] [56].

1. **Individual spammers:** in this case, a spammer, who does not work with anyone else, writes reviews. The spammer may register at the review site as a single user, or as many fake users using different user-IDs. He/she can also register at multiple review sites and write spam reviews.
2. **Group spammers:** there are two main sub-cases [4] [5].
    - A group of spammers (persons) works in collusion to promote a target entity and/or to damage the reputation of another. The individual spammers in the group may or may not know each other.
    - A single person registers multiple user-IDs and spam using these user-IDs. These multiple user-IDs behave just like a group in collusion.

Group spamming is highly damaging due to the sheer number of members in a group, it can take total control of the sentiment on a product and completely mislead potential customers, especially at the beginning of a product launch [4] [5] [17] [56].

2.9.2 Types of Data, Features and Detection

Three main types of data have been used for review spam detection [45] [56]:

1. **Review Content:** the actual text content of each review. From the content, we can extract linguistic features such as word and POS n-grams and other syntactic and semantic clues for deceptions and lies. However, linguistic

features may not be enough because one can fairly easily craft a spam review that is just like a genuine one.

2. **Meta-data about each Reviewer:** the data such as the star rating, user-ID of the reviewer, the time when the review was posted, the host IP address, MAC address of the reviewer's computer, and the geo-location of the reviewer.
3. **Product Information:** information about the entity being reviewed, e.g., the product description and sales volume/rank.

These types of data have been used to produce many spam features. One can also classify the data into public data and site private data. By public data, we mean the data displayed on the review pages of the hosting site, e.g., the review content, the reviewer's user-ID and the time when the review was posted. By private data, we mean the data that the site collects, but is not displayed on their review pages for public viewing, e.g., the IP address and MAC address from the reviewer's computer, and the cookie information [56].

2.10 Summary

This chapter gave an overview of basic theoretical foundation about Knowledge Discovery in Databases (KDD), Data Mining (DM), and Data Classification (DC). Then, it introduced Text Mining (TM), and Text Classification (TC). In addition, it described major kinds of classification algorithms, which are used in our research: Naïve Bayes (NB), ID3, K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM). Then, it discussed the imbalance class distribution problem and major existing techniques related to this problem. Finally, it discussed opinion mining and opinion spam detection. The next chapter will review the related work that was done for detecting spam in non-Arabic opinion reviews.

Chapter 3

Related Works

In this chapter, different related works are studied and investigated. The related works are introduced and analyzed for detecting spam in non-Arabic opinion reviews because in Arabic language, there is still no published work in this area. Parts of the related works can be a basis for solving the thesis problem, but no one can present a complete solution. The related works presented addressed the area of detecting spam in non-Arabic opinion reviews.

In the next section, we focus on detecting individual spam reviews and reviewers, in Section 3.2, we discuss the detection of spammer groups and in Section 3.3, we discuss the detecting spam in Chinese opinion mining.

3.1 Detecting Individual Review Spam

Several detecting individual review spam algorithms have been proposed and evaluated in various ways.

A preliminary study was reported by ***Jindal et. al.*** in [44]. A more in-depth investigation was given by ***Jindal et. al.*** in [45]. Their study based on 5.8 million reviews, and 2.14 million reviewers crawled from amazon.com. They discovered that spam activities are widespread. For example, they found a large number of duplicates and near-duplicate reviews written by the same reviewers on different products or by different reviewers (possibly different user-IDs of the same persons) on the same products or different products. They employed three sets of features about reviews, reviewers, and products. They also identified three categories of spams: spam reviews (also called untruthful opinions), reviews on brand only, and non-reviews. Based on the features and the training data, they used a logistic regression model to detect spam reviews. The method is evaluated using the area under the curve. In general, the area under the curve does not exceed 78%, which are considering not high. Also, they did not apply more than one classifier for classification, and their approach relies heavily on text similarity; therefore, it is only good for certain types of spamming activities, i.e., duplicated review spamming.

Huang et. al. in [42] was attempting to identify fake reviews. In their case, a manually labeled fake review corpus was built from Epinions reviews. In Epinion after a review is posted, users can evaluate the review by giving it a helpfulness score. They can also write comments about the reviews. The authors manually labeled a set of fake and non-fake reviews by reading the reviews and the comments. For learning, several types of

features were proposed, which are similar to those in [45] with some additions, e.g., subjective and objectivity features, positive and negative features, reviewer's profile, authority score computed using PageRank. They tried several supervised methods, including SVM, Logistic Regression, and NB to detect spam reviews. In general, their experiments showed that NB achieved promising results comparing with other supervised methods. Also, the F-Score term does not exceed 58.30%, which are considered low.

Ott et. al. in [59], in their case, fake hotel reviews were built from 20 most popular Chicago hotels. They tried several classification approaches, which have been used in related tasks such as genre identification, psycholinguistic deception detection, and text classification. All these tasks have some existing features proposed by researchers. Their experiments showed that combined classifier with both n-gram and psychological deception features achieves best results nearly 90% cross-validated accuracy.

Jindal et. al. in [46], in their case, fake reviews gathered from Amazon. They identified several unusual reviewer behavior models based on different review patterns that suggest spamming. Each model assigns a numeric spamming behavior score to a reviewer by measuring the extent to which the reviewer practices spamming behavior of the type. All the scores are then combined to produce the final spam score. Thus, this method focuses on finding spammers rather than fake reviews. The spamming behavior models are:

- ***Targeting products:*** to game an online review system, it is hypothesized that a spammer will direct most of his efforts on promoting or victimizing a few products, which are collectively called the targeted products. They are expected to monitor the targeted products closely and mitigate the ratings by writing spam reviews when the time is appropriate.
- ***Targeting groups:*** the spam behavior model defines the pattern of spammers manipulating ratings of a set of products sharing some attribute(s) within a short span of time.
- ***General rating deviation:*** a genuine reviewer is expected to give rating to other raters of the same product. As spammers attempt to promote or demote some products, their rating typically deviates a great deal from those of other reviewers.

- ***Early rating deviation:*** early deviation captures the behavior of a spammer contributing a spam review soon after product launch. Such reviews are likely to attract attention from other reviewers, allowing spammers to affect the views of subsequent reviewers.

Based on the labeled spammers which labeled by the evaluators, they used Linear Regression model. Their results showed that proposed ranking and supervised methods are effective in discovering spammers and outperform other baseline method based on helpfulness votes alone.

Cunningham et. al. in [15], in their case, fake hotel reviews were built from Irish hotels on TripAdvisor. They proposed an unsupervised method to detect spam reviews based on a distortion algorithm based on the ranking of products. Their idea is that spam reviews may distort product ranking more than randomly chosen reviews. Their results showed that the distortion algorithm is effective, but restricted to the situation where the ranking of products is available, which cannot be applied to general review data sets.

Algur et. al. in [2], in their case, fake reviews were built from digital camera product. They proposed a technique called shingle technique, which uses descriptive features of reviews for detecting spam reviews from non-spam ones. Their proposed technique involves the following steps:

- ***Review Data Store:*** where review regions in web sites are detected by review region extractor and then individual reviews are extracted using the review extractor.
- ***Shingling Technique:*** in this step, firstly, pre-processing is done on reviews in order to remove some extra characters such as punctuations. Features of previously extracted reviews were then extracted. By creating all possible combinations with "***W***" numbers of these extracted features (***W*** is determined by users), shingles of size "***W***" are created for each review. Finally, similarity of each pair of reviews is calculated. Reviews are then classified as spam and non-spam based on the returned value.

Their results showed that shingling technique obtained accuracy values of 83.54%.

3.2 Detecting Group Review Spam

The previous works focus on catching individual spammers; however, it is not applicable to the general case of discovering spammers group. In this section, we discuss detecting group review spam approach.

Glance et. al. in [29], in their case, fake reviews gathered from Amazon reviewers and their reviews. They proposed a method to detect spammer groups. Their method first uses a frequent item set mining method to find a set of candidate groups. Then, they use several behavioral models derived from the collusion phenomenon among fake reviewers and relation models (called GSRank) based on the relationships among groups. Their results showed that ranking is effective, and they reflect people's perceptions of spam and non-spam, and achieve 80% Kappa scores. Also, their method focused on catching spammer groups, and is not applicable to the general case of discovering individual spammers.

3.3 Detecting Spam in Chinese Opinion Mining

The above-mentioned works focus on detection spam reviews written in the English. In this section, we discuss detection written in Chinese product reviews.

Liu et. al. in [50], in their case, approximately 600 reviews (300 normal reviews and 300 spam reviews) gained from four online shopping websites, namely Amazon, IT 168, Jingdog online shop, and Zhongguancun online. The product type involves phone of three different brands of Nokia, Samsung, and Motorola. They proposed technique used four important features based on questions and hyperlinks are used to characterize reviews, then they detect them by Logistic Regression model. The proposed technique involves the following steps: firstly, 2-gram model is proposed to characterize reviews with the consideration of the word order, and then the Katz smoothing method is applied to smoothing the model. Finally, KL divergence is applied to detect the untruthful reviews. Their experiments showed the effectiveness of detecting the spam in the field of Chinese product reviews, and achieved the best results nearly 87% accuracy values. Their method, however, unable on finding new spam reviews timely, and continuing to expand date collections and perfect distinguishing effect.

3.4 Summary

From the previous works, we can conclude that few works proposed detecting spam in non-Arabic opinion reviews. This means that opinion spam detection is still in its early stage. In Arabic language, there is still no published work in this area. We preferred to work at the detecting spam in Arabic opinion reviews because it is a new area and try to solve one of the challenging problems in opinion mining [27]. The next chapter will present an approach for detecting spam in Arabic opinion reviews.

Chapter 4

An Approach for Detecting Spam in Arabic Opinion Reviews

In this chapter, we explain our proposed Spam Detection in Arabic Opinion Reviews (SDAOR) approach which we followed in this research. This chapter organized into eight sections. Section 4.1, presents general view of our proposed approach. Section 4.2, will give a description of the collecting various data sets for designing experimental data. Section 4.3, performs integrated data sets from multiple sources into a coherent form to get better input data for data mining techniques. Section 4.4, perform identification of the spam review label. Section 4.5, presents preprocessing steps. Section 4.6, presents processing stage. Section 4.7, applies the approach by using data mining method. Section 4.8, evaluates the approach using accuracy and F-measure.

To implement and evaluate this approach, various steps have to be performed. The main required steps are shown in Figure 4.1 and stated below:

- **Data Acquisition:** from online Arabic economic websites, including tripadvisor.com.eg, booking.com, and agoda.ae with different characteristics and sizes by crawls.
- **Data Integration:** from multiple sources which are: TripAdvisor dataset, Booking dataset and Agoda dataset into a coherent form which is: TBA dataset to get better input data for data mining techniques.
- **Spam Identification Labeling:** different types of the spam will be identified after integration of TBA dataset, and manually labeled each record with spam and non-spam.
- **Preprocessing:** we apply a number of preprocessing techniques to deal with noisy, missing, inconsistent data, Arabic stopword removal, and light stemming.
- **Processing Stage:** the processing stage will be implemented based on the following steps:
    - Data mining classification experiments.
    - Text mining classification experiments.
    - Data-Text mining classification experiments.

  Then we applied each previous step by using classification algorithms: Naïve Bayes (NB), ID3, K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) as multi classification.

- **Evaluate the Approach:** to evaluate the classification performance of our approach, we use accuracy and F-measure.
- **Comparing Phase:** we apply two comparisons:
    - Compare performance before using our proposed approach and after using it.
    - Compare performance between our proposed approach and other published work, which has been used for detecting spam in non-Arabic opinion reviews.

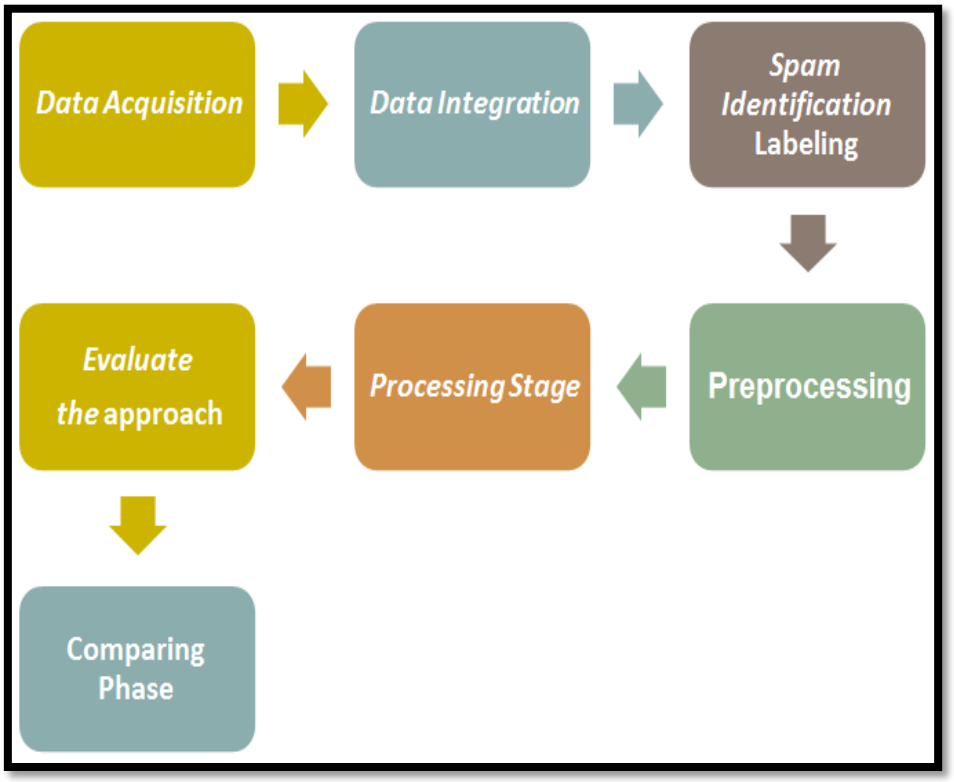


Figure 4.1: Methodology Steps [25].

4.1 Spam Detection in Arabic Opinion Reviews (SDAOR) Approach

The main objective of this research is to develop a SDAOR approach which, is to detect spam review in Arabic opinion reviews that can be valid for the economic domain in an efficient way with high accuracy and F-measure. To achieve this, we used a combination of review content feature, meta-data about each reviewer features, and hotel information features in one mining classification experiment that means combining methods from data mining and text mining. In addition, we try to overcome the drawbacks

of the class imbalance problem by using an over-sample approach (more about an over-sample approach in Section 2.7).

4.2 Data Acquisition

Since there are no publicly available Arabic opinion data set to test the effectiveness of our classifiers and to evaluate the effectiveness of the proposed approach, we have built an in-house data set of spam reviews and reviewers using human collected from online Arabic economic websites, including tripadvisor.com.eg, booking.com, and agoda.ae with different characteristics and sizes by crawls. We randomly collected and selected from among any of the records available from June 2007 to October 2012. Table 4.1 presents, general information about these three data sets.

Table 4.1: General Information about Data Sets.

Data sets	# instance	# Attribute
TripAdvisor	718	29
Booking	1408	17
Agoda	722	15

4.2.1 TripAdvisor Dataset

TripAdvisor offers over 100,000 hotels in 30 countries. The TripAdvisor database contains over 75 million reviews and opinions from travelers. Its data set that was taken manually from the tripadvisor.com.eg website, which consists of 718 records chosen randomly from among any of the records available in the its database and 29 attributes that available in its database. Table 4.2 presents, the attributes and their description of the TripAdvisor dataset.

Table 4.2: TripAdvisor Dataset Description.

Attribute	Description
reviewerUseName	The user name of the reviewer.
reviewerAge	The age of the reviewer.
reviewerGender	The gender of the reviewer.
reviewerState	The state of the reviewer.
reviewerType	The type of the reviewer.
aboutReviewer	The information about the reviewer.
reviewerTravelStyle	The travel style of the reviewer.
whenTraveling,Reviewer	When the reviewer travels?
reviewerUsuallyTravelFor	The reviewer usually travels forward.
reviewerTravelWith	The reviewer travel with.
reviewerReviewsNumber	The number reviews of the reviewer.
hotelName	The name of the hotel.
hotelState	The state of the hotel.
hotelPrice	The price of the hotel.
hotelReviewsNumber	The number reviews of the hotel.
hotelRate	The rate of the hotel.
reviewBody	The body of the review.
reviewRate	The reviewer rate of the review.
valuRates	The reviewer rate of the hotel value.
roomsRate	The reviewer rate of the hotel rooms.
locationRate	The reviewer rate of the hotel location.
cleanlinessRate	The reviewer rate of the hotel cleanliness.
sleepQualityRate	The reviewer rate of the hotel sleeps quality.
serviceRate	The reviewer rate of the hotel service.
reviewDate	The date of the review.
reviewsBy	The review source.
stayingDate	The reviewer staying the date.
tripType	The type of the trip.
isHelpful	The review is helpful.

4.2.2 Booking Dataset

Booking offers over 247,432 hotels in 177 countries. The Booking database contains reviews and opinions in 41 languages from travelers. Its data set that was taken

manually from the booking.com website, which consists of 1408 records chosen randomly from among any of the records available in the its database and 17 attributes that available in its database. Table 4.3 presents, the attributes and their description of the Booking dataset.

Table 4.3: Booking Dataset Description.

Attribute	Description
reviewerName	The name of the reviewer.
reviewerLocation	The location of the reviewer.
hotelName	The name of the hotel.
hotelLocation	The location of the hotel.
hotelReviewsNumber	The number reviews of the hotel.
cleanlinessHotelRate	The rate of the hotel cleanliness.
comfortHotelRate	The rate of the hotel comfort.
locationHotelRate	The rate of the hotel location.
serviceHotelRate	The rate of the hotel service.
staffHotelRate	The rate of the hotel staff.
valueHotelRate	The rate of the hotel value.
reviewBody	The body of the review.
hotelRate	The rate of the hotel.
reviewRate	The reviewer rate of the review.
reviewDate	The date of the review.
journeyType	The type of the journey.
isHelpful	The review is helpful.

4.2.3 Agoda Dataset

Agoda offers over 200,000 hotels around the world. The Agoda database contains over 3 million reviews and opinions from travelers. Its data set that was taken manually from the agoda.ae website, which consists of 722 records chosen randomly from among any of the records available in its database and 15 attributes that available in its database. Table 4.4 presents, the attributes and their description of the Agoda dataset.

Table 4.4: Agoda Dataset Description.

Attribute	Description
reviewer	The user name of the reviewer.
hotelName	The name of the hotel.
hotelLocation	The location of the hotel.
hotelReviewsNumber	The number reviews of the hotel.
hotelRate	The rate of the hotel.
review	The review.
rate	The reviewer rate of the review.
valueRate	The reviewer rate of the hotel value.
roomsRate	The reviewer rate of the hotel rooms.
locationRate	The reviewer rate of the hotel location.
cleanlinessRate	The reviewer rate of the hotel cleanliness.
sleepQualityRate	The reviewer rate of the hotel sleeps quality.
serviceRate	The reviewer rate of the hotel service.
reviewDate	The date of the review.
tripType	The type of the trip.

4.3 Data Integration

In order to increase data, mining projects require data from more than one data source [34] [63]. Our integrated data from multiple sources which are: TripAdvisor dataset, Booking dataset and Agoda dataset into a coherent form which is: TBA dataset to get better input data for data mining techniques. The TBA dataset consists of 2848 records and 35 attributes (because there duplicate attributes between attributes in the three data sets). Table 4.5 presents, the attributes and their description of the TBA dataset.

Table 4.5: TBA Dataset Description.

Attribute	Description	Selected
reviewerUseName	The user name of the reviewer.	
reviewerAge	The age of the reviewer.	√
reviewerGender	The gender of the reviewer.	√
reviewerLocation	The location of the reviewer.	√
reviewerType	The type of the reviewer.	
aboutReviewer	The information about the reviewer.	
reviewerTravelStyle	The travel style of the reviewer.	
whenTraveling,Reviewer	When the reviewer travels?	
reviewerUsuallyTravelFor	The reviewer usually travels forward.	
reviewerTravelWith	The reviewer travel with.	
reviewerReviewsNumber	The number reviews of the reviewer.	√
hotelName	The name of the hotel.	√
hotelLocation	The location of the hotel.	√
hotelPrice	The price of the hotel.	
hotelReviewsNumber	The number reviews of the hotel.	√
cleanlinessHotelRate	The rate of the hotel cleanliness.	√
comfortHotelRate	The rate of the hotel comfort.	√
locationHotelRate	The rate of the hotel location.	√
serviceHotelRate	The rate of the hotel service.	√
staffHotelRate	The rate of the hotel staff.	√
valueHotelRate	The rate of the hotel value.	√
hotelRate	The rate of the hotel.	√
reviewBody	The body of the review.	
reviewRate	The reviewer rate of the review.	√
valueRate	The reviewer rate of the hotel value.	√
roomsRate	The reviewer rate of the hotel rooms.	√
locationRate	The reviewer rate of the hotel location.	√
cleanlinessRate	The reviewer rate of the hotel cleanliness.	√
sleepQualityRate	The reviewer rate of the hotel sleeps quality.	√
serviceRate	The reviewer rate of the hotel service.	√
reviewDate	The date of the review.	√
reviewsBy	The review source.	
stayingDate	The reviewer staying the date.	
tripType	The type of the trip.	√
isHelpful	The review is helpful.	√
Class	The class label	√

The following steps are performed as part of the integration of the three data sets:

- The same attribute in the three data sets may have different names. So for efficient later integration, simplified data description and understanding of data mining results, we unified these attributes to a unified attribute name in TBA dataset. Table 4.6 presents, the attributes that have different names in the three data sets and its unification.

Table 4.6: Attributes Before and After Unification.

TripAdvisor	Booking	Agoda	Unification
reviewerUseName	reviewerName	reviewer	**reviewerUseName**
reviewerState	reviewerLocation		**reviewerLocation**
hotelState	hotelLocation	hotelLocation	**hotelLocation**
reviewBody	reviewBody	review	**reviewBody**
reviewRate	reviewRate	rate	**reviewRate**
tripType	journeyType	tripType	**tripType**

- The values in *reviewerLocation* and *hotelLocation* attributes in the three data sets contain values of these attributes might be different (e.g., "مصر " (Egypt)), (e.g., "مصر, الإسكندرية" (Egypt, Alexandria)), (e.g., "مصر, الإسكندرية، برج العرب" (Egypt, Alexandria, Burj Al Arab)). So for efficient later integration, simplified data description and understanding of data mining results, we unified the values for these attributes to the country. For example, we formatted all values for these attributes into a country value; (e.g., "مصر"(Egypt)).
- The *reviewDate* attribute in the three data sets contain dates in many formats, e.g. "Sep 24, 2011", "9/24/11", "24.09.11". So for efficient later integration, simplified data description and understanding of data mining results, we transformed this attribute to a standard value. For example, we formatted all dates into a standard value; e.g. "24 September 2011".
- The *tripType* attribute in the three data sets contain values of attributes might be different. So for efficient later integrating, we unified these attributes to a unified attribute name. For example, we grouped all trip types into five categorical segments; ("مسافر منفرد"(solo traveler), "السفر كزوجين" (traveling as a couple), "السفر مع العائلة"(traveling with family), " السفر مع مجموعة أصدقاء" (traveling with friends), "السفر في عمل "(traveling on business)).

- None selected attributes of TBA dataset that presented in Table 4.5. We will discuss it in Subsection 4.5.1.

4.4 Spam Identification Labeling

There was no labeled data set for opinion spam before this project. To evaluate our method, we need to label each record in TBA dataset by identifying *Class* attribute by spam review and non-spam review.

There are three main types of information in the TBA dataset: the content of the review, the reviewer who wrote the review, and the hotel being reviewed. We manually label a *Class* attribute for each record in the TBA dataset. We have done this step after integration of TBA dataset because we found that spam reviewers often post on more than one website. The following is performed as part to the identification of the spam review:

- Reviews about brands only.
- Non-reviews.
- Irrelevant review.
- General review.
- Hotels that are 100% positive review. We think every hotel has some drawbacks, even if they are small.
- The contradiction in the *reviewBody* attributes for the same reviews.
- Contradiction between attributes, e.g. "*reviewRate* vs. *reviewBody*".
- Duplication and near-duplicate (not an exact copy) of the features between the two records.
- Duplicate and near-duplicate (not an exact copy) reviews. We are taking into our account the following types of duplicates (the duplicates include near-duplicates):
    - Duplicate from the same *reviewerUseName* on the same *hotelName*.
    - Duplicates from different *reviewerUseNames* on the same *hotelName*.
    - Duplicates from the same *reviewerUseName* on different *hotelNames*.
    - Duplicates from different *reviewerUseNames* on different *hotelNames*.

The *Class* attribute for each record in the TBA dataset may be labeled with the spam class based on under one or more of the previous steps of the identification of the spam record.

4.5 Preprocessing

Today's real-world databases are highly susceptible to noise, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results. Preprocessing is a necessary step for serious, effective, real-world data mining. There are a number of preprocessing techniques such as: cleaning, transformations, reduction, tokenization, Arabic stopword removal, and light stemming [63].

4.5.1 Data Preprocessing

In order to detect spam in Arabic opinion reviews by applying data mining method, the data should first be preprocessed to get better input data for data mining techniques [34] [63]. In the data preprocessing step, we did some preprocessing of the TBA dataset before loading the data set to the data mining software, irrelevant attributes should be removed because they add noise to the data and affect model accuracy. The attributes marked as selected as seen in Table 4.5 are processed via the RapidMiner environment [24] to apply the data mining methods on them. The attributes such as the *reviewerUseNamer, aboutReviewer, reviewerTravelStyle, reviewerTravelWith, WhenTraveling,Reviewer, reviewerUsuallyTravelFor, ReviewsBy,* or *stayingDate* are not selected to be part of the mining process; this is because they do not provide any knowledge for the data set processing, and they present personal information of the reviewers. In addition, they have very large variances or duplicate information, which makes them irrelevant for data mining. The *reviewBody* attribute is not selected to be part of the data mining process; this is because it; we use in the text mining process.

The following is performed as part of the preprocessing of the data set:

- The TBA dataset contains 4984 missing values in various attributes from 2848 records and because the size of the data set is not too large (the rule of thumb: number of records 5,000 or more desired [48]), we used *Replace Missing Values* operator to fill missing values using the average value

(because it is the smart way to fill missing values [48]) of the existing values in the column. After this procedure, the missing values will fill and no record is ignored.

- The *reviewerAge* attribute in the TBA dataset contains a large number of continuous values (ages). So for efficient later processing, simplified data description and understanding of data mining results, we discretized this attribute to categorical one. For example, we grouped all ages into seven categorical segments; 13-17, 18-24, 25-34, 35-49, 50-64, 65+ and "أفضل عدم الإجابة" (prefer not to answer).

After applying the previous preprocessing methods, we try to analyze the data visually and figure out the distribution of values, specifically the class of reviewers. Figure 4.2 depicts the distribution of type reviewers according to their class; it is apparent from the Figure 4.2 that the spam reviewers present about 13.34% (379 records) of the TBA dataset that means the class imbalance problem occur (more detailed about the class imbalance problem is discussed in Section 2.7).

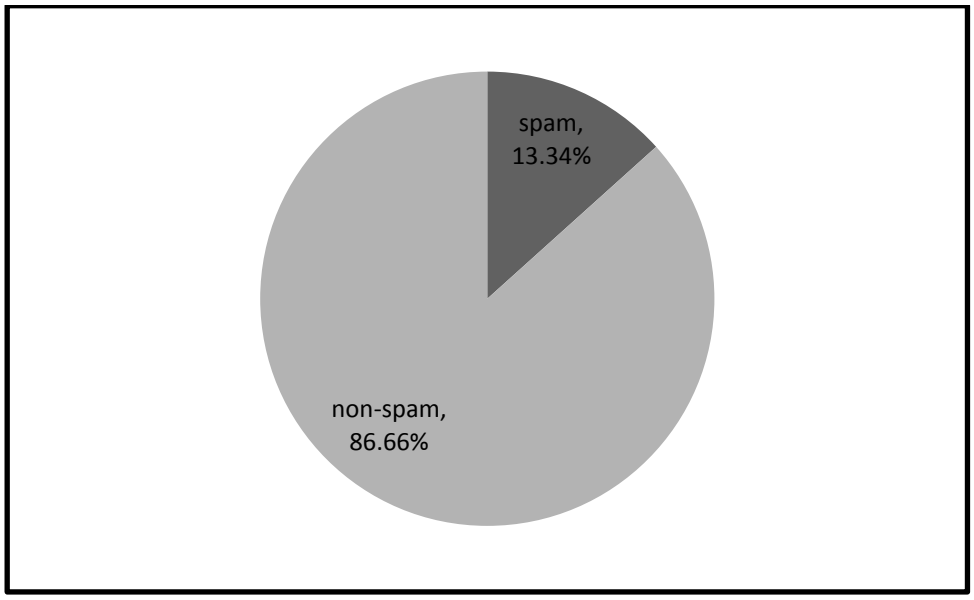


Figure 4.2: The Distribution of Type Reviewer According to their Class.

The imbalanced class distribution problem occurs when one class is represented by a large number of examples (non-spam class) while the other is represented by only a few (spam classes). In this case, a classifier usually tends to predict that samples have the majority class and completely ignore the minority class [21]. So we must try to overcome

the drawbacks of the class imbalance problem by using sampling technique, which discussed in Section 2.7.

4.5.2 Text Preprocessing

In order to detect spam in Arabic opinion reviews by applying text mining methods, we have used an in-house Arabic TBA (ATBA) corpus collected from *reviewBody* attribute of TBA dataset. In this ATBA corpus, each *reviewBody* was saved in a separate file. The ATBA corpus contains 3854 distinct tokens in 2848 documents that vary in length. These documents fall into two classification categories that vary in the number of documents. Table 4.7 shows the number of documents in each category.

Table 4.7: The Number of Documents in Each Category.

Category	Total Number of Documents for Each Category
Spam	379
Non-spam	2469

In the preprocessing step, documents are transformed into a representation suitable for applying the learning algorithms. The main required Arabic documents process steps are shown in Figure 4.3 and stated below:

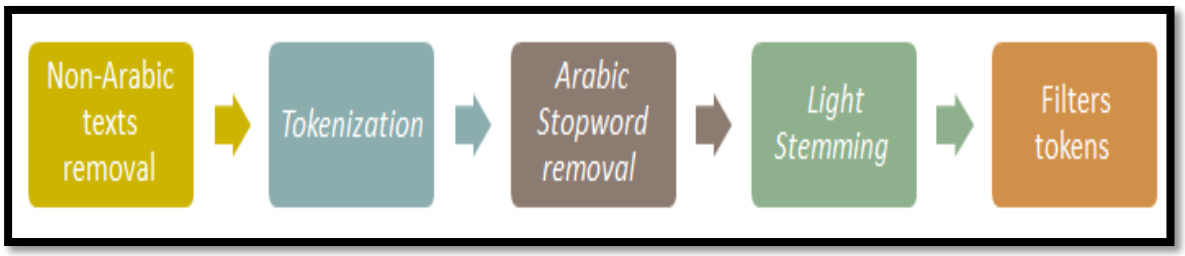


Figure 4.3: Structuring Text Data Process.

- All the non-Arabic texts were removed.
- *Tokenization* consists of separating strings by word boundaries (such as spaces in Western languages).
- *Arabic Stopword removal* deletes tokens that are frequent, but generally not content-bearing.
- *Light Stemming* reduces Arabic words to their light stems by removing frequently used prefixes and suffixes in Arabic words. Light Stemming

chosen because it allows remarkably good information retrieval without providing correct morphological analyses (more detailed about stemming methods is discussed in Section 2.8) [57].

- *Filter token* based on their default length (i.e. the minimal number of characters that a token must contain is equal to 4 characters and the maximal number of characters that a token must contain to be considered is equal to 25 characters) in order not to exclude important words. Thus, the tokens less than 4 characters and greater than 25 characters are cut out, while the token between 4 characters and 25 characters remains within the data set (more detailed about filtering methods is discussed in Section 2.8).

4.5.3 Data-Text Preprocessing

In order to detect spam in reviews by applying machine learning techniques, we have used an in-house ATBA Hotels (ATBAH) dataset collected from selected features from TBA dataset as mentioned in Table 4.5 and extracted features from ATBA corpus.

Since ATBA corpus contains 3854 distinct tokens, which are too large, we used *Remove Correlated Attributes* operator to remove correlated tokens and *Remove Useless Attributes* operator to remove all useless tokens from an ATBA corpus to obtain the best Arabic text tokens for ATBA corpus. Table 4.8 lists the best 102 Arabic text tokens for ATBA corpus, these tokens are selected manually by the highest frequency among other tokens. The listed tokens are sorted from the best to worst according to their frequency.

Now ATBAH dataset collected from selected features from TBA dataset as mentioned in Table 4.5 and the best 102 Arabic text tokens in the ATBA corpus as mentioned in Table 4.8.

Table 4.8: Best 102 Arabic Text Token for ATBA Corpus.

اخر (another)	اثاث (furniture)	اتمن (wish)	احد (single)
اجمل (beautiful)	احترام (respect)	اجر (wage)	ابتسام (smile)
اختيار (choose)	اجاز (allow)	اجد (find)	اثناء (during)
ادب (polite)	اجراء (procedures)	اخلاق (ethics)	اتصال (contact)
احمد (thank)	ابدا (never)	استمتع (enjoy)	احتياج (lack)
احسن (best)	احب (like)	اجانب (foreigners)	اخص (specifically)
ادفع (pay)	اجراءت (procedures)	اخذ (take)	اجواء (ambiences)
احراج (embarrass)	ائتمان (credit)	ابراج (towers)	اجهز (prepare)
اتقدم (extend)	اجمالا (overall)	ابواب (doors)	احسست (felt)
ابلاغ (informing)	احجز (book)	اخذنا (took)	اختار (choose)
احتجت (protested)	احتساب (calculation)	احببت (liked)	اخبرت (told)
ارتفاع (high)	استرخاء (relaxation)	ابسط (simplest)	اصحاب (owners)
اتوقع (expect)	استجمام (recreation)	ابعد (farthest)	اداء (performance)
ابداع (creativity)	اتساع (breadth)	ابشع (ugliest)	الفندق (hotel)
اتصلت (contacted)	ابهر (impress)	ازدحام (congestion	اتخاذ (adoption)
احضار (bring)	اثاث (furniture)	ادعو (invite)	اتيكيت (etiquette)
اتردد (hesitate)	ابتعاد (away)	اجازة (holiday)	اتناول (take)
ازعاج (disturbance)	استغرق (took)	استقبال (reception)	اسعار (prices)
اسلوب (style)	اسوء (worst)	اشكر (thank)	امتاز (excel)
امضيت (spent)	بطء (slowness)	بعد (next)	بقشيش (tip)
منتجع (resort)	منظر (view)	موقع (location)	نت (internet)
نشعر (feel)	نواد (clubs)	هدوء (calm)	واسع (wide)
وجود (presence)	يحجز (book)	يستحق (merit)	يشمل (includes)
يفتقد (misses)	يفتقر (lacks)	يقدم (offers)	يمتاز (advantage)
يمتلك (owns)	يناسب (fits)	ينصح (advised)	ينظف (cleans)
يوجد (exist)	بود (like)		

4.6 Processing Stage

In this section, we present our strategy which we followed to achieve our goal, which tries to develop an approach to detect spam in Arabic opinion reviews that can be valid for the economic domain in an efficient way with high accuracy and F-measure. To do that, we implemented the following steps:

- Data mining classification experiments.
- Text mining classification experiments.
- Data-Text mining classification experiments.

4.6.1 Data Mining Classification Experiments

In our experiments, the following steps are performed as part on the data classification:

- We start by classifying instances before doing any resampling to the TBA dataset to test the classification accuracy.
- We are classifying instances after applying an under-sample approach (more detailed about an under-sample approach is discussed in Section 2.7) to the TBA dataset to test the classification accuracy.
- We are classifying instances after applying an over-sample approach (more detailed about an over-sample approach is discussed in Section 2.7) to the TBA dataset to test the classification accuracy.

4.6.2 Text Mining Classification Experiments

We are classifying instances after applying an over-sample approach to the ATBA corpus to test the classification accuracy.

4.6.3 Data-Text Mining Classification Experiments

We are classifying instances after applying an over-sample approach on combining methods from data mining and text mining (our approach) in one mining classification experiment for gaining many spam features. Also, we try to overcome the drawbacks of the class imbalance problem by using an over-sample approach. Therefore, we are classifying instances after applying an over-sample approach to the ATBAH dataset to test the classification accuracy.

4.7 Apply the SDAOR Approach

This section describes the major kinds of classification algorithms, which are used in our research: Naïve Bayes (NB), ID3, K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) which are provided by RapidMiner environment [24]. In the following

subsections, we present these classification algorithms and their settings which are used during experiment results.

4.7.1 Naïve Bayes (NB)

We use NB (as discussed in Subsection 2.6.1) in our research, which is a technique for estimating probabilities of individual variable values, given a class, from training data and then allow the use of these probabilities to classify new entities [34] [48]. Figure 4.4 illustrates the settings of NB. We use Laplace correction to prevent high influence of zero probabilities [48].

Figure 4.4: Setting of NB.

4.7.2 ID3

We use ID3 (as discussed in Subsection 2.6.2) in our research, which is a technique for constructing the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node [34] [63]. ID3 run with the gain ratio for the criterion term, a minimal size of 10 for a node to allow a split, a minimal size of 2 for all leaves, and a minimal gain of 0.1 to produce a split to pick a good attribute for the root of the tree and give us tree with the greatest predictive accuracy.

4.7.3 K-Nearest Neighbor (K-NN)

We use K-NN (as discussed in Subsection 2.6.3) in our research, which a supervised learning algorithm where the result of the new instance query is classified based on majority of K-Nearest Neighbor category. The purpose of this algorithm is to classify a new object based on attributes and training samples [34] [35] [64]. We start by setting k=1 in the parameter setting and try a series of increasing k's with (K=1, 3, 5, 7, 9) and take the

highest accuracy, also set the measure types appropriately (in this case we have numeric predictors and a nominal label, hence we choose mixed measures).

4.7.4 Support Vector Machine (SVM)

We use SVM (as discussed in Subsection 2.6.4) in our research, which is a supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basis of this algorithm is to take a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier [6] [21]. Figure 4.5 illustrates the settings of Support Vector Machine.

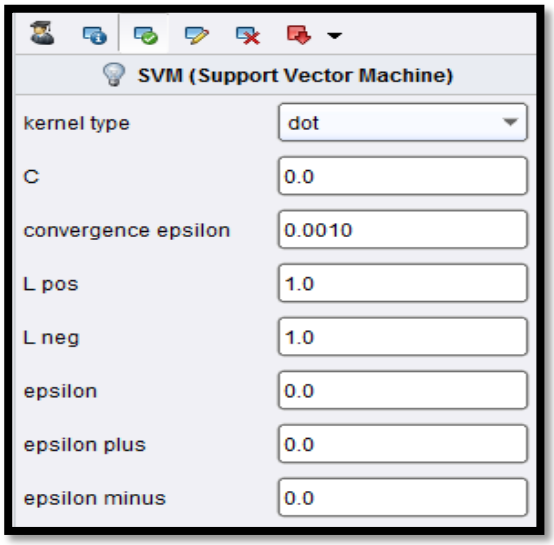


Figure 4.5: Setting of SVM.

4.8 Evaluate the Approach

Evaluation metrics play an important role to evaluate classification performance. Accuracy measure is the most commonly for these purposes. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier [34] [70]. However, for classification of imbalanced data, accuracy is unsuitable metric since the minority class has a very little impact on the accuracy as compared to that of the majority class [21] [58]. For example, in a problem where a minority class is represented by only 1% of the training data and 99% for majority class, a simple strategy can be one that predicts the majority class label for every example. It can achieve a high accuracy of 99% may mean nothing to some application where the learning concern is the

identification of the minority cases. Therefore, other metrics have been proposed to evaluate classifier performance for imbalanced datasets. Three important measures are commonly used, precision, recall and F-measure [59]. The precision is the ratio of the number of positive examples correctly recognized, and the total number of examples (both positive and negative) recognized [70]. The recall is the ratio of the number of positive examples correctly recognized and the number of all positive examples. F-measure is defined as the harmonic mean of recall and precision [48]. A high F-measure value signifies a high value for both recall and precision. It is evaluated when the learning objective is to achieve a performance between the identification rate (recall) and the identification accuracy (precision) of a specific class [63]. F-measure which is shown in Equation 4.1.

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{precision}}{\text{Recall+ precision}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad 4.1$$

In our experiments, we use F-measure and compare it with accuracy to evaluate the performances of the compared classifier for the data set. Also; for evaluation purpose, we use k-fold cross-validation method provided by RapidMiner environment [24]. K-fold cross-validation works by using part of the data to train the model, and the rest of the data set to test the accuracy of the trained model. In k-fold cross-validation the data is first partitioned into k equally (or nearly equally) sized segments or folds. In this case, we will divide the data set into k = 10 parts, then train and test for each part. For example, if we have 100 data records, we will train the model with the ninety records chosen randomly by shuffled sampling type, and then test the accuracy of the trained model with the ten records.

4.9 Summary

This chapter describes the methodology used for our research. It presents our processing strategy which we followed to achieve our goal with more detail. In addition, it explained the classification algorithms which are used during experiment results. The next chapter will be discussing the results of our experiments using our approach and the described methodology.

Chapter 5

Experimental Results and Evaluation

In this chapter, we present and analyze the experimented results. Different machine learning classifiers used for our experiments named, Naïve Bayes (NB), ID3, K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) which are provided by RapidMiner environment [24] [34] [66]. These selected classification methods achieve the best accuracy among other classification methods in our data sets to classify the instances. We explained the machine environment, and the tools used in our research. In addition, we present the evaluation measurements for a classification model during sets of experiments by using the equation of accuracy, and F-measure. Furthermore, we present most appropriate attributes play an important role in detecting spam in Arabic opinion reviews. Additionally, we set that the ratio of the number of majority of class samples to the number of minority class samples in the training data to be 1:1.

We apply a set of experiments; in Section 5.2, we classified instances based on data mining experiments without resampling approach on the TBA dataset, and then classified instances based on data mining experiments with an under-sample approach on the TBA dataset, and finally classified instances based on data mining experiments with an over-sample approach on the TBA dataset. In Section 5.3; we classified instances based on text mining experiments with an over-sample approach to the ATBA corpus. In Section 5.4, we classified instances based on data-text mining experiments with an over-sample approach on the ATBAH dataset. Finally; we discussed the results of all our experiments and compared our results with some other published work on the field of detecting spam in non-Arabic opinion reviews.

5.1 Experiments Setup

In this section, a description about the experimental environment, tools used in experiments, measures of performance evaluation of classifiers and SDAOR has been provided.

5.1.1 Experimental Environment and Tools

We applied experiments on a machine with properties that is Intel (R) Core (TM) 2 Quad CPU Q8300 @ 2.50 GHz, 4.00 GB RAM, 320 GB hard disk drive and Windows 7 operating system installed. To carry out our thesis (including the experimentation), special tools and programs were used:

1. **RapidMiner application program:** used to build our approach, and conduct experiments practical and extracting the required results.
2. **Microsoft Excel:** used to organize and store datasets in tables, do some simple preprocessing and analyze the results.

5.1.2 Measurements for Experiments

The measures of evaluating the performance of classification are a confusion matrix. Also, to perform the comparisons of the tested algorithms, through the performance of each classifier was evaluated using the accuracy and F-measure which are illustrated in Section 4.8. Based on the Equation 4.1 in Section 4.8, we extract our experiment results in the next sections.

5.2 Data Mining Classification Experiments

We apply set of experiments; in the Subsection 5.2.1, we classified instances without making any resampling approach in the TBA dataset. Finally, in the Subsection 5.2.2; we classified instances after applying the resampling approach on the TBA dataset.

5.2.1 Data Mining Classification Experiments without Resampling Approach

We start with classifying instances before making any resampling approach in the TBA dataset to test the classification accuracy. Table 5.1 illustrates the accuracy before did any change in the TBA dataset. We note that accuracy range from 92% to 94%, which is considered a good result.

Table 5.1: Accuracy for TBA Dataset in Data Mining Classification Experiments without Resampling Approach.

Classifier	Naïve Bayes	ID3	K-NN with K=3
Accuracy	94.15%	92.74%	92.86%

However; we cannot depend on accuracy metric as a measure for classification for imbalanced data as mentioned in Chapter 2. Therefore, we compute an F-measure of whole classes to evaluate classification performance. Table 5.2 shows an F-measure for the TBA dataset. We note that in general F-measure is much less than accuracy that is meant the TBA dataset to have an imbalanced problem. For instance, the accuracy of a NB classifier is 94.15% while the F-measure is 39.44%. So, the accuracy measure cannot detect the

imbalanced problem and cannot give us the actual classification performance, especially when the data set has an imbalanced class distribution problem.

Table 5.2: F-measure for TBA Dataset in Data Mining Classification Experiments without Resampling Approach.

Classifier	Naïve Bayes	ID3	K-NN with K=3
F-measure	39.44%	27.90%	22.79%

5.2.2 Data Mining Classification Experiments with Resampling Approach

In order to tackle the class imbalance problem, the data should first be sampled to get better input data for data mining techniques. Sampling methods (more detailed about the sampling techniques is discussed in Section 2.7) modify the distributions of the majority and minority class for the training data set to obtain a more balanced number of instances in each class [58]. To minimize class imbalance in training data, there are two basic methods, an under-sampling and an over-sampling.

5.2.2.1 Data Mining Classification Experiments with Under-Sample Approach

In this experiment, we classify instances after applying an under-sample approach (more detailed about an under-sample approach is discussed in Section 2.7) to the TBA dataset to test the classification accuracy. Applying an under-sample approach which, is supposed to reduce the number of samples from the majority class. Hence, an under-sample approach is aimed to decrease the skewed distribution of majority class and minority class by lowering the size of majority class. In this approach, first we look at the minority class which has a smaller number of instances, and then takes the same number of instances of other majority classes. In order to do this, we use *Sample* operator provided by RapidMiner environment [24]. This operator performs a random sampling from each majority class. Finally, we obtain the new data set with a balance number of instances in each class (The numbers of records are decreased to 758 records). Table 5.3 illustrates the accuracy of TBA dataset. We note that accuracy range from 67% to 71%, which is considered the low result.

Table 5.3: Accuracy for TBA Dataset in Data Mining Classification Experiments with Under-Sampling Approach.

Classifier	Naïve Bayes	ID3	K-NN with K=3
Accuracy	68.48%	67.11%	71.05%

Table 5.4 presents the F-measure after applying an under-sample approach on TBA dataset. Although it has less accuracy, we find under-sample approach creates improvement on an F-measure comparing with the results obtain from classification experiments without resampling approach, which is considered the low result. The reason is that an under-sampling may remove some instances, which are important to the classification process.

Table 5.4: F-measure for TBA Dataset in Data Mining Classification Experiments with Under-Sampling Approach.

Classifier	Naïve Bayes	ID3	K-NN with K=3
F-measure	70.10%	70.59%	69.45%

5.2.2.2 Data Mining Classification Experiments with Over-Sample Approach

In this experiment, we classify instances after applying an over-sample approach (more detailed about an over-sample approach is discussed in Section 2.7) to the TBA dataset to test the classification accuracy. Applying an over-sample approach which, is duplicating the sample of the minority class and adding them to the data set. It is different than an under-sample approach, so there is no information is lost; all instances are employed. In this approach, first we look at the majority class which has the greater number of instances, and then we replicate samples from other minority classes until reach to the same or a close number of instances in majority class. Finally, we obtain the new data set with a balance number of instances for each class (The numbers of records are increased to 4937 records). Table 5.5 illustrates the accuracy of TBA dataset. We note that accuracy range from 87% to 97%, which is considered a good result comparing with the results obtain from classification experiments without resampling approach and classification experiments without an under-sample approach to TBA dataset.

Table 5.5: Accuracy for TBA Dataset in Data Mining Classification Experiments with Over-Sampling Approach.

Classifier	Naïve Bayes	ID3	K-NN with K=3
Accuracy	87.01%	97.43%	94.94%

Table 5.6 presents the F-measure after applying an over-sample approach on TBA dataset. Although it has less accuracy, we find an over-sample approach creates significant improvement on an F-measure comparing with the results obtain from classification

experiments without resampling approach and classification experiments with an under-sample approach to TBA dataset. For example, the F-measure of a NB is 39.44% in the classification experiments without resampling approach; the F-measure of a NB is 70.10% in the classification experiments with an under-sample approach, and the F-measure of a NB is 87.98% in the classification experiments with an over-sample approach.

Table 5.6: F-measure for TBA Dataset in Data Mining Classification Experiments with Over-Sampling Approach.

Classifier	Naïve Bayes	ID3	K-NN with K=3
F-measure	87.98%	97.50%	95.18%

5.3 Text Mining Classification Experiments with Over-Sample Approach

In this experiment, we classify instances after applying an over-sample approach (only using an over-sample approach because we found an over-sample approach is better than an under-sample approach, because it is different from an under-sample approach, so there is no information is lost; all instances are employed) on the ATBA corpus to test the classification accuracy. Whereas, data in ATBA corpus are too large, ID3 algorithms not finished their computation in a long time. So, we use Support Vector Machine algorithm instead of ID3 algorithm. Table 5.7 illustrates accuracy for ATBA corpus. We note that accuracy range from 93% to 98%, which is considered a good result comparing with the results obtain from data mining classification experiments on TBA dataset.

Table 5.7: Accuracy for ATBA Corpus in Text Mining Classification Experiments with Over-Sampling Approach.

Classifier	Naïve Bayes	Support Vector	K-NN with K=3
Accuracy	98.07%	93.52%	97. 37%

Table 5.8 presents the F-measure after applying an over-sample approach on the ATBA corpus. We note that the F-measure range from 90% to 97%, which is considered a good result. Also, we find text mining classification experiments with an over-sampling approach creates significant improvement on an F-measure comparing with the results obtain from data mining classification experiments with an over-sampling approach. For example, the F-measure of a NB is 87.98% in the data mining classification experiments with an over-sample approach, and the F-measure of a NB is 90.15% in the text mining classification experiments with an over-sampling approach.

Table 5.8: F-measure for ATBA Corpus in Text Mining Classification Experiments with Over-Sampling Approach.

Classifier	Naïve Bayes	Support Vector	K-NN with K=3
F-measure	90.15%	93.81%	97.51%

5.4 Data-Text Mining Classification Experiments with Over-Sample Approach

In this experiment, we classify instances after applying an over-sample approach on combining methods from data mining and text mining (our approach). This means we used a combination of the review content features; meta-data about each reviewer feature, and hotel information features in one mining classification experiment. In addition, we try to overcome the drawbacks of the class imbalance problem by using an over-sample approach. Therefore, we classify instances after applying an over-sample approach (only using an over-sample approach because we found an over-sample approach is better than an under-sample approach because it is different from an under-sample approach, so there is no information is lost; all instances are employed) on the ATBAH dataset to test the classification accuracy. Table 5.9 illustrates accuracy for the ATBAH dataset. We note that accuracy range from 93% to 99%, which is considered a good result comparing with the results obtain from data mining classification experiments on the TBA dataset and text mining classification experiments with an over-sample approach on the ATBA corpus.

Table 5.9: Accuracy for ATBAH Dataset in Data-Text Mining Classification Experiments with Over-Sample Approach.

Classifier	Naïve Bayes	Support Vector	K-NN with K=3
Accuracy	99.20%	93.81%	97.97 %

Table 5.10 presents the F-measure after applying an over-sample approach on the ATBAH dataset. We note that F-measure range from 93% to 99%, which is considered a good result. Also, we find Data-Text mining classification experiments with an over-sample approach to the ATBAH dataset created significant improvement on an F-measure of all selected classifiers comparing with the results obtain from data mining classification experiments with an over-sample approach to TBA dataset and text mining classification experiments with an over-sample approach on the ATBA corpus in most cases. For example, the F-measure of a NB improved from 87 to 99. For example, the

F-measure of a NB is 87.98% in the data mining classification experiments with an over-sample approach; the F-measure of a NB is 90.15% in the text mining classification experiments with an over-sample approach, and the F-measure of a NB is 99.59% of the data-text mining classification experiments with an over-sample approach.

Table 5.10: F-measure for ATBAH Dataset in Data-Text Mining Classification Experiments with Over-Sample Approach.

Classifier	Naïve Bayes	Support Vector	K-NN with K=3
F-measure	99.59%	93.87%	97.94%

5.5 Optimal Attributes

We chose the most appropriate attributes play an important role in F-measure results. For this reason, we try to find the optimal attributes by starts with an empty attribute set and iteratively add the attribute whose inclusion improves the F-measure the most.

From our experiments, we found that the selected attribute for TBA dataset in Table 4.5 and the best 102 Arabic text feature for ATBA corpus in Table 4.8 are optimal attributes that influence the category to the target class (spam and non-spam) and achieved the best F-measure. Figure 5.1 depicts a part of the tree that resulted from applying the ID3 classification algorithm for the class of the reviewer as a target class.

```
hotelName  = Holiday Villa Madinah
|   reviewRate  = 10
|   |   reviewDate  = 21 September 2011: spam {Non-spam=0, spam=7}
|   |   reviewDate  = 28 July 2011: spam {Non-spam=0, spam=6}
|   |   reviewDate  = 6 January 2012: Non-spam {Non-spam=1, spam=0}
|   reviewRate  = 5.3: Non-spam {Non-spam=1, spam=0}
|   reviewRate  = 6.7: Non-spam {Non-spam=2, spam=0}
|   reviewRate  = 8.7: Non-spam {Non-spam=1, spam=0}
|   reviewRate  = 9: Non-spam {Non-spam=1, spam=0}

hotelName  = فندق راديسون بلو، القاهرة هليوبوليس
|   reviewerAge  = 18-24: Non-spam {Non-spam=1, spam=0}
|   reviewerAge  = 25-34
|   |   reviewDate  = 23 August 2012: spam {Non-spam=0, spam=7}
|   |   reviewDate  = 26 February 2012: spam {Non-spam=0, spam=7}
|   |   reviewDate  = 29 April 2012: Non-spam {Non-spam=1, spam=0}
|   |   reviewDate  = 9 May 2012: spam {Non-spam=0, spam=6}
hotelName  = برج ساعة مكة الملكي، فندق فيرمونت
|   reviewerLocation  = الإمارات العربية المتحدة: Non-spam {Non-spam=1, spam=0}
|   reviewerLocation  = المكسيك: Non-spam {Non-spam=1, spam=0}
|   reviewerLocation  = المملكة العربية السعودية
|   |   reviewRate  = 2: spam {Non-spam=0, spam=7}
|   |   reviewRate  = 5.7: Non-spam {Non-spam=1, spam=0}
|   |   reviewRate  = 7: spam {Non-spam=0, spam=7}
|   |   reviewRate  = 9.3: Non-spam {Non-spam=1, spam=0}
|   reviewerLocation  = قطر: Non-spam {Non-spam=1, spam=0}

hotelName  = Red Carpet Resort
|   tripType  = السفر كزوجين
|   |   isHelpful  = 1: Non-spam {Non-spam=2, spam=0}
|   |   isHelpful  = 8: spam {Non-spam=0, spam=7}
|   tripType  = السفر مع العائلة: Non-spam {Non-spam=14, spam=0}
|   tripType  = مجموعة أصدقاء
|   |   reviewRate  = 10: Non-spam {Non-spam=2, spam=0}
|   |   reviewRate  = 2.9: spam {Non-spam=0, spam=7}
|   |   reviewRate  = 5: Non-spam {Non-spam=1, spam=0}
|   tripType  = مسافر منفرد: Non-spam {Non-spam=1, spam=0}

hotelName  = كانترى كلوب ڤسنيلا دى مارى جولف
|   isHelpful  = 1
|   |   reviewDate  = 1 September 2012: spam {Non-spam=0, spam=7}
|   |   reviewDate  = 12 July 2012: Non-spam {Non-spam=1, spam=0}
|   |   reviewDate  = 14 August 2012: Non-spam {Non-spam=1, spam=0}
|   |   reviewDate  = 2 September 2012: spam {Non-spam=0, spam=6}
|   |   reviewDate  = 25 July 2012: spam {Non-spam=0, spam=7}
|   |   reviewDate  = 26 August 2012: Non-spam {Non-spam=1, spam=0}
|   |   reviewDate  = 29 August 2012: spam {Non-spam=0, spam=7}
|   |   reviewDate  = 30 August 2012: Non-spam {Non-spam=1, spam=0}
|   |   reviewDate  = 9 September 2012: spam {Non-spam=0, spam=7}
|   isHelpful  = 2: Non-spam {Non-spam=1, spam=0}
```

Figure 5.1: The Resulted ID3.

To interpret the rules in the ID3, as an example; the first branch of the tree says that, if the hotelName = Holiday Villa Madinah, reviewRate = 10 and reviewDate = 21 September 2011, the specialty of the reviewer can be predicted as "spam" and so on for the rest of the tree.

5.6 Discussion

Form our data sets, we noted that there are different types of spam available in Arabic opinion reviews, such as: reviews about brands only, non-reviews, irrelevant review, general review, hotels that are 100% positive review, e.g. "hotelRate = 10", contradiction between attributes, e.g. "contradiction between reviewBody and reviewRate", duplication and near-duplicate of the features, e.g. "duplication and near-duplicate of reviewerUseName for the same hotelName", and duplication and near-duplicate of reviews, e.g. "duplication and near-duplicate of reviewBody for the same hotelName". Table 5.11, presents a part of the opinion spam and its type that presents in our data set.

Table 5.11: Some of Opinion Spam and its Type.

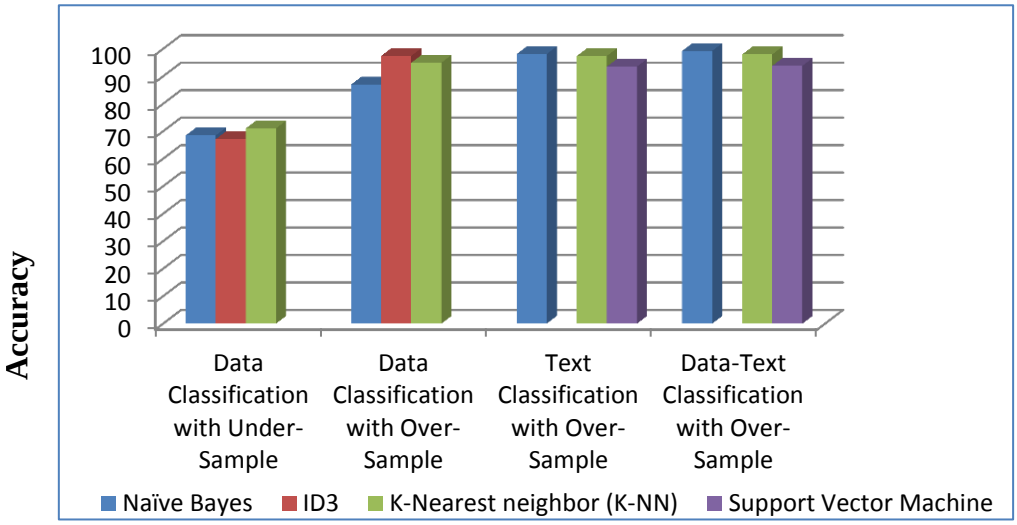
Opinion Spam	
الموقع: تتسم المنشأة السياحية فندق اطلس المطار بموقعها المميز المتصل بالمطار في مدينة النواصر (الدار البيضاء)، وتضم معالم الجذب السياحي القريبة تكنوبارك الدار البيضاء وجامعة الحسن الثاني . **ميزات الفندق:** يوجد في فندق اطلس المطار حمام سباحة مفتوح ومركز اللياقة البدنية والعناية الصحية وغرفة بخار.تتوفر خدمة الاتصال اللاسلكي بالإنترنت مجانا في الأماكن العامة. تشتمل تسهيلات رجال الأعمال في هذه المنشأة السياحية من فئة 3 نجوم على مركز لرجال الأعمال وقاعة اجتماعات وخدمات سكرتارية .تحتوي هذه المنشأة السياحية على 2 مطاعم إلى جانب كافيتيريا وبار إلى جانب حمام السباحة .يتم تقديم إفطار مجاني.تشتمل المرافق الإضافية على خدمات التدليك والعناية وساونا ومكتب لخدمات الاستقبال والإرشاد . **غرف النزلاء:** تتمتع غرف النزلاء بمناظر تطل على حمام السباحة أو الحديقة .تشتمل غرف النزلاء مكيفة الهواء البالغ عددها 189 في فندق اطلس المطار على بارات مصغرة و خزائن.تم تجهيز أجهزة التلفزيون بقنوات فضائية .توفر كل تجهيزات الإقامة مكاتب وجرائد مجانية وهاتف للاتصال المباشر .تشتمل المرافق الإضافية على نوافذ يمكن فتحها ومستلزمات مجانية للعناية الشخصية .بالإضافة إلى ذلك، فإن المرافق المتاحة حسب الطلب تشتمل على ثلاجات وجلسات تدليك في الغرف ومكالمات إيقاظ .يتم تقديم خدمة تنظيف الغرف يوميًا.	
Spam Type	Reviews about brands only.
Opinion Spam	
آخر مرة زرت فندق اطلس المطار في فبراير 2012م.	
Spam Type	Non-reviews.
Opinion Spam	
فندق محبب وخدمة ودودة	
Spam Type	General review.
Opinion Spam	
قمت بالغطس مرة واحدة مع غواصي دهب، كما أن مدربي الغطس كانوا ممتازين حقا ومختصين. غواصو دهب بالتأكيد هم	
Spam Type	Irrelevant review.

Also, we noted that group of reviewers who works together written spam reviews to promote or demote a set of target hotels . Many sets of unusual behaviors such as: writing reviews together in a short time window, writing reviews right after the hotel lunch, and group content similarity.

From all our experiments, we can say our approach achieved the best classification accuracy of a minority class. Table 5.12 and Figure 5.2 show the accuracy of all experimental results.

Table 5.12: Accuracy for All our Experiments.

Classifier	NB	ID3	SVM	K-NN with K=3
Data Classification with Under-Sample	68.48%	67.11%		71.05%
Data Classification with Over-Sample	87.01%	97.43%		94.94%
Text Classification with Over-Sample	98.07%		93.52%	97.37%
Data-Text Classification with Over-Sample	99.20%		93.81%	97.97%



Experiment Name

Figure 5.2: Accuracy for All our Experiments.

We can summarize accuracy results for all our experiments results as is in NB, the highest accuracy result (99.20%) was in our approach (data-text mining classification experiments with over-sample approach). In SVM, the highest accuracy result (93.81%) was on our approach. In K-NN, the highest accuracy result (97.97%) was on our approach.

Also, we note using our approach to perform significant improvement and the best F-measure results. Table 5.13 and Figure 5.3 show the F-measure for all experimental results.

Table 5.13: F-measure for All our Experiments.

Classifier	NB	ID3	SVM	K-NN with K=3
Data Classification with Under-Sample	70.10%	70.59%		69.45%
Data Classification with Over-Sample	87.98%	97.50%		95.18%
Text Classification with Over-Sample	90.15%		93.81%	97.51%
Data-Text Classification with Over-Sample	99.59%		93.87%	97.94%

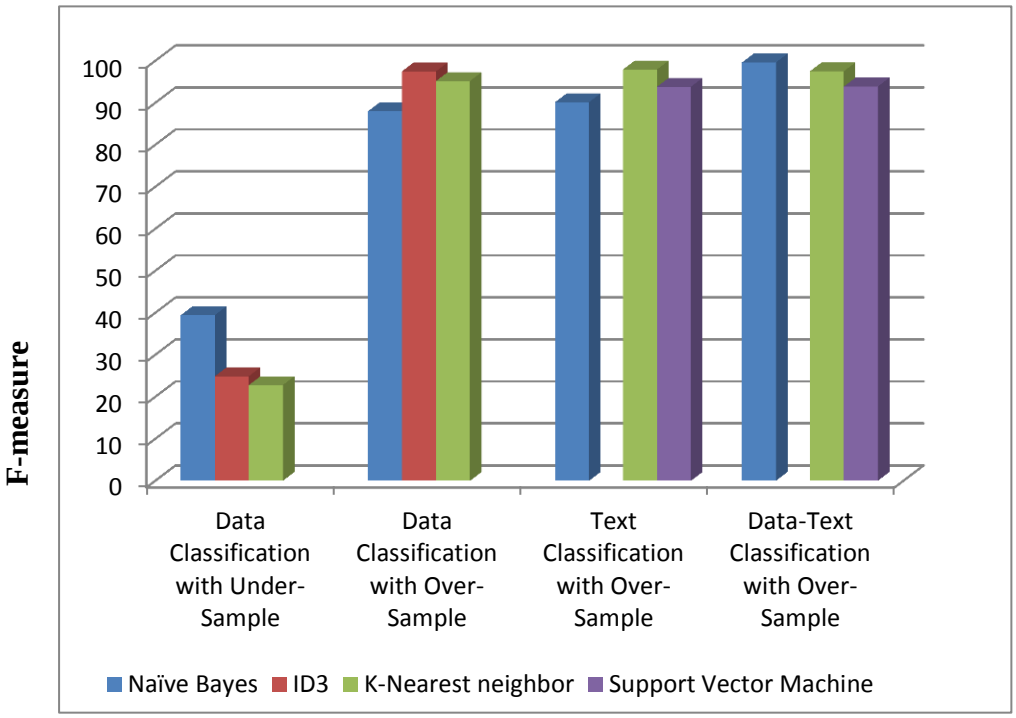


Figure 5.3: F-measure for All our Experiments.

We can summarize F-measure results for all our experiments results as is in NB, the highest F-measure result (99.59%) was in our approach (data-text mining classification experiments with over-sample approach). In SVM, the highest F-measure result (93.87%)

was on our approach. In K-NN, the highest F-measure result (97.94%) was on our approach.

We can note the great difference in improvement before and after applying our approach in accuracy and F-measure. For example, in the NB, the accuracy is 87.01% and the F-measure is 87.98% in data mining classification experiments with an over-sample approach, after we apply text mining classification experiments with an over-sample approach, we obtain 98.07% for accuracy and 90.15% for F-measure, and after we apply our approach, we obtain 99.20% for accuracy and 99.59% for F-measure.

Also, we note that the resulting rules achieve the conditions necessary for the patterns as follows:

- **Valid:** the discovered rules are valid with respect to a certain level, and a NB classifier in our approach has an accuracy of 99.59% .
- **Novel:** the discovered rules are previously unknown or obvious, for example, in the ID3 model, if the hotelName equal Holiday Villa Madinah, reviewRate = 10 and reviewDate = 21 September 2011, the specialty of the reviewer can be predicted as "spam".
- **Useful:** the discovered rules provide information useful to the business for knowing the class of the reviewers (spam and non-spam).
- **Understandable:** the discovered rules are understandable and facilitate a better understanding of the underlying data.

We find an under-sample approach is a good solution for imbalanced data distribution, but an over-sample approach is better than an under-sample approach because it is different from an under-sample approach, so there is no information is lost; all instances are employed.

From all the above, experimental results confirm our findings, which are saying the combining methods from data mining and text mining (our approach) achieved the best classification accuracy of minority class for detecting spam in Arabic opinion reviews problem. Because the combination of review content feature, meta-data about each reviewer features, and hotel information features in one mining classification experiment gain many spam features.

To confirm our experimental results, Table 5.14, compares our work with some other published work on the field of detecting spam in non-Arabic opinion reviews.

Table 5.14: Comparison Between our Research and Some Other Research Related to Detecting Spam in Non-Arabic Opinion Reviews.

Study	Method	Result
Our research	Naive Bayes, SVM Classification, ID3, K-NN.	Naive Bayes achieved promising results comparing with other supervised methods. Naive Bayes obtains 99.20% accuracy and 99.59% F-measure.
Algur et. al. [2]	Shingling Technique.	Shingling technique obtains 83.54% accuracy.
Huang et. al. [9]	SVM Classification, Logistic Regression, Naive Bayes.	Naive Bayes achieved promising results comparing with other supervised methods. The F-measure does not exceed 58.30%
Glance et. al. [29]	GSRank method.	GSRank method achieved 80% Kappa scores.
Jindal et. al. [45]	Logistic Regression.	Logistic regression did not exceed 78% area under the curve (AUC).
Liu et. al. [50]	Logistic Regression.	Logistic regression achieved the best results nearly 87% accuracy values.
Ott et. al. [59]	SVM Classification.	SVM achieved the best results nearly 90% cross-validated accuracy.

Although we did not use the same language and data, in general we can conclude that SDAOR achieved the best results for accuracy and F-measure, and general we expect the approach to be used for a wide range of applications.

5.7 Summary

This chapter presents and analyzes the experimented results. In addition, it explained the machine environment, and the tools used in our research. Also, it presented the evaluation measurements for a classification model during sets of experiments. Furthermore, it presented most appropriate attributes play an important role in detecting spam in Arabic opinion reviews.

Chapter 6

Conclusion and Future Work

This chapter draws a conclusion, which includes its results, discussion and comparing our approach with others, then gives some suggestions for future work.

6.1 Conclusion

In this thesis, we proposed a novel technique for detecting spam in Arabic opinion reviews. The proposed approach is called SDAOR and is based on combining methods from data mining and text mining. The proposed technique structure and components were presented and explained. It involves the following steps for detecting spam in Arabic opinion reviews: data acquisition, data integration, spam identification labeling, preprocessing, processing stage and finally evaluates the approach.

For our experiments, we built an in-house a labeled data set of spam reviews and reviewers using human collected from online Arabic economic websites, including tripadvisor.com.eg, booking.com, and agoda.ae with different characteristics and sizes by crawls. Then, we integrated data from multiple sources into a coherent form to get better input data for mining techniques. Then, we dealt with noisy, missing, inconsistent data, Arabic stopword removal, and light stemming by some of the preprocessing techniques. In an integrated data set, we find an imbalanced class distribution problem occurs. To tackle this problem, we perform some of the sampling techniques, namely an under-sample approach and an over-sample approach. In addition, we find an under-sample approach is a good solution for imbalanced data distribution, but an over-sample approach is better than an under-sample approach, because it is different than an under-sample approach, so there is no information is lost; all instances are employed. Finally, we implemented our strategy which we followed to achieve our goal, which is trying to develop an approach to detect spam in Arabic opinion reviews that can be valid for the economic domain with high accuracy and F-measure.

For evaluation purposes, we use k-fold cross-validation method provided by RapidMiner environment [24]. In addition, we set that the ratio of the number of majority of class samples to the number of minority class samples in the training data to be 1:1. Experimental results show our approach perform significant improvement on F-measure results in some case F-measure improved up to 99.59%.

6.2 Future Work

Possible directions for future work include:

- Evaluation of the effectiveness of the proposed methodology based on a larger data set.
- Evaluation of the methods proposed in this work to opinions coming from other domains.
- Using more types of data (private data), such as: the host IP address, MAC address of the reviewer's computer, and the geo-location of the reviewer to detect more sophisticated spam Arabic opinion reviews.
- Developing a new method by using collaborative methods classification and clustering in conjunction with one another to detect spam Arabic opinion reviews.
- Generalizing our approach to other kinds of user-generated content, e.g., Internet forums, discussion groups and blogs.
- Using multiple languages, such as: Arabic, English, and European in the same review.
- Generalizing our approach to other kinds of spams, e.g., web spam, and email spam.

References

[1] AbouAssi, R., Challita, E., Farra, N., and Hajj, H. (2010) *Sentence Level and Document Level Sentiment Mining for Arabic Texts*. 2010 IEEE International Conference on Data Mining Workshops.

[2] Algur, S., Hiremath, E., Patil, A., and Shivashankar, S. (2010) *Spam Detection of Customer Reviews from Web Pages*. In Proceedings of the 2nd International Conference on IT and Business Intelligence held in IMT Nagpur.

[3] Anand, S.S., Bell, D.A., and Hughes, J.G. (1995) *Evidence Based Discovery of Knowledge in Databases*. IEEE Colloquium on Knowledge Discovery in Databases, Digest No: 1995/021 (A): 9/1 – 9/5, London, UK.

[4] Arjun, M., Glance, N. and Liu, B. (2012) *Spotting Fake Reviewer Groups in Consumer Reviews*. In Proceedings of International World Web Conference (WWW-2012).

[5] Arjun, M. and Liu, B. (2012) *Modeling Review Comments*. In Proceedings of the 50th Annual Meeting of Association for Computational Linguistics (ACL-2012) (Accepted for Publication).

[6] Berry, M.J.A., and Linoff, G. (1997) *Data Mining Techniques for Marketing, Sales, and Customer Support*. John Wiley and Sons, Inc., USA. (ISBN 0-471-17980-9).

[7] Black, W., Elkateb, S., Farwell , D., Fellbaum, C., Pease, A. and Vossen, P. (2006) *Arabic WordNet and the Challenges of Arabic*. Proceedings of the International Conference of Arabic NLP/MT, London.

[8] Bouzerdoum, A., Nguyen, G. and Phung, S. (2009) *Learning Pattern Classification Tasks with Imbalanced Data Sets*. In Proceedings Yin (Eds.), Pattern Recognition, Pages 193-208, Vukovar, Croatia.

[9] Bramer, M. (2007) *Principles of Data Mining*. Springer-Verlag, London. (ISBN 1-84628-765-0).

[10] Bu, J., Chen, Ch. Liu, B., and Qiu, G. (2009) *Expanding Domain Sentiment Lexicon through Double Propagation*. In IJCAI, Pages 1199–1204, USA.

[11] Chang, K., Fan, R., Hsieh, C., Lin, C., and Wang, X. (2008) *LIBLINEAR - A Library for Large Linear Classification*. (Access Online). Available: www.csie.ntu.edu.tw/~cjlin/liblinear.

[12] Chen, A.L.P., Liu, H., Lin, R. and Wu, H. (2004) *Music Classification Using Significant Repeating Patterns*. At the 9th International Conference on Database Systems for Advanced Applications (DASFAA-2004, Springer-Verlag), Pages 506 – 518, Jeju Island, Korea. (LNCS 2973, ISBN 3-540-21047-4).

[13] Chen, L., Chen, M., Hsu, C., and Zeng, W. (2008) *An Information Granulation Based Data Mining Approach for Classifying Imbalanced Data*. Information Sciences, Vol. 78, No. 16, Pages 3214-3227.

[14] Cios, K.J., Pedrycz, W., and Swiniarski, R.W. (1998) *Data Mining Methods for Knowledge Discovery*. Kluwer Academic Publishers, Norwell, MA, USA. (ISBN 0-7923-8252-8).

[15] Cunningham, P., Greene, D., Smyth, B. and Wu, G. (2010) *Distortion as a Validation Criterion in the Identification of Suspicious Reviews*. In Proceedings of 1st Workshop on Social Media Analytics (SOMA 10), Washington, DC, USA. Tech. Rep.

[16] David, O. and Francesco, M. (2010) *Research Challenge on Opinion Mining and Sentiment Analysis*. The CROSSROAD Roadmap on ICT for Governance and Policy Modeling.

[17] Dray, M. P. G. and Harb, A. (2008) *Web Opinion Mining: How to Extract Opinions from Blogs?*. Proceedings of the 5th International Conference on Soft Computing as Trans Disciplinary Science and Technology. ACM New York, NY, USA.

[18] Duda, R., Hart, P. and Stork D. (2001) *Pattern Classification*. 2nd Edition, Wiley Interscience.

[19] Dumais, S. T., Heckerman, D., Platt, J. and Sahami M. (1998) *Inductive Learning Algorithms and Representations for Text Categorization*. In Proceedings of ACM-CIKM98, Pages 148-155.

[20] Dunham, M. (2003) *Data Mining: Introductory and Advanced Topics*. 1st Edition, Pearson Education.

[21] Estabrooks, T., Japkowicz, Jo and N. (2004) *A Multiple Resampling Method for Learning from Imbalanced Data Set*. Computational Intelligence.

[22] El-Halees, A. (2011) *Mining Opinions in User-Generated Contents to Improve Course Evaluation* . In J.M. Zain et al. (Eds.): The 1nd International Conference on Software Engineering and Computer Systems, Part II, CCIS 180, Pages 107–115.

[23] El-Halees, A. (2011) *Arabic Opinion Mining Using Combined Classification Approach*. In Proceedings of the International Arab Conference on Information Technology (ACIT'2011), Naif Arab University for Security Science (NAUSS), Riyadh, Saudi Arabia.

[24] Euler, T., Klinkenberg, R., Mierswa, I., Scholz, M. and Wurst, M. (2006) *YALE: Rapid Prototyping for Complex Data Mining Tasks*. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06).

[25] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996) *The KDD Process for Extracting Useful Knowledge from Volumes of Data*. Communications of the. ACM, No. 39, Vol. 11, Pages 27-34.

[26] Feldman, R., and Sanger, J. (2007) *The Text Mining Handbook: Advanced Approaches In Analyzing Unstructured Data*, Cambridge University Press.

[27] Francesco, M. and David, O. (2010) *Research Challenge on Opinion Mining and Sentiment Analysis*. In Proceedings of ICT for Governance and Policy Modeling.

[28] Frank, E., and Witten, I. (2005*) Data Mining: Practical Machine Learning Tools and Techniques*. 2nd Edition, Morgan Kaufmann, San Francisco.

[29] Frawley, W.J., Matheus, C.J., and Piatetsky-Shapiro, G. (1991) *Knowledge Discovery in Databases: An Overview*. In: Piatetsky-Shapiro, G., and Frawley, W.J. (Eds.): Knowledge Discovery in Databases, AAAI/MIT Press, Pages 1 – 27. (ISBN 0262660709).

[30] Furuse, O., Hiroshima, N., Kataoka, R. and Yamada, S. (2007) *Opinion Sentence Search Engine on Open-Domain Blog*. In IJCAI, Pages 2760–2765, USA.

[31] Garcia, E., and He, H. (2009) *Learning from Imbalanced Data*. IEEE Transactions on Knowledge and Data Engineering 21 (9), Pages 1263 –1284.

[32] García, V., Sánchez, J.S., Alejo, R., and Sotoca, M. (2007) *The Class Imbalance Problem in Pattern Classification and Learning*, Pages 283-291, Spain.

[33] Glance, N., Jindal, N., Liu, B., Mukherjee, A. and Wang, J (2011). *Detecting Group Review Spam*. In Proceedings of WWW. 2011. (Poster paper).

[34] Han, J., and Kamber, M. (2006) *Data Mining: Concepts and Techniques*. 2nd Edition. Morgan Kaufmann Publishers, San Francisco, USA. (ISBN 1-55860-901-6).

[35] Han, H. and Mao, B. (2010) *Fuzzy-Rough K-Nearest Neighbor Algorithm for Imbalanced Data Sets Learning*, In Proceedings of FSKD 2010, Pages 1286-1290.

[36] Harb, A., and Plantié, M. (2008) *Web Opinion Mining: How to Extract Opinions from Blogs?*. Proceedings of the 5th International Conference on Soft Computing as Trans Disciplinary Science and Technology. ACM New York, NY, USA.

[37] Haruno, M., and Taira, H. (1999) *Feature Selection in SVM Text Categorization*. In Proceedings of the 16th National Conference on Artificial Intelligence, Pages 480 - 486.

[38] Hotho, A., Nürnberger, A., and Paaß, G. (2005) *A Brief Survey of Text Mining*. LDV Forum – GLDV Journal for Computational Linguistics and Language Technology 20 (1), Pages 19 – 62. (ISSN 0175-1336).

[39] Hsu, W., Lee, M.L., and Zhang, J. (2002) *Image Mining: Trends and Developments*. Journal of Intelligent Information Systems 19 (1), Pages 7 – 23.

[40] Hu, S., Liang, Y., Ma, L. and He, Y. (2009). *MSMOTE: Improving Classification Performance When Training Data is Imbalanced*. In Proceedings of 2nd International Workshop on Computer Science and Engineering, Vol.2, Pages 13-17.

[41] Huang, M., Li, F. and Zhu, X. (2010) *Sentiment Analysis with Global Topics and Local Dependency*. In AAAI.

[42] Huang, M., Li, F., Yang, Y. and Zhu, X. (2011) *Learning to Identify Review Spam*. In Proceedings of the 22nd International Joint Conference on Artificial Intelligence (ACL-2011).

[43] Huberman, B. and Wu, F. (2010) *Opinion Formation under Costly Expression.* ACM Trans. Intell. Syst. Technol.

[44] Jindal, N., and Liu, B. (2007). *Review Spam Detection*. In Proceedings of WWW (Poster paper).

[45] Jindal, N., and Liu, B. (2008). *Opinion Spam and Analysis*. In Proceedings of the Conference on Web Search and Data Mining (WSDM-2008).

[46] Jindal, N., Liu, B., Lim, E., Lauw, H. and Nguyen, V. A. (2010) *Detecting Product Review Spammers Using Rating Behavior*. In Proceedings of the Conference on Information and Knowledge Management (CIKM-2010).

[47] Joachims, T. (1998) *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. In Proceedings of ECML-98, 10th European Conference on Machine Learning.

[48] Kaur, G. and Singh, L (2011). *Data Mining: An Overview*, IJCST, Vol. 2, Issue 2, Pages 336-339.

[49] Lalitrojwong, P. and Somprasertsri, G. (2010) *Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization*. Journal of Universal Computer Science, (16), Pages 938-955.

[50] Liu, L. and Wang, Y. (2012) *A Method for Sorting Out the Spam from Chinese Product Reviews*. In Proceedings of the Conference on Consumer Electronics, Communications and Networks (CECNet).

[51] Lee, L. and Pang, B. (2008) *Opinion Mining and Sentiment Analysis*. Foundations and Trends in Information Retrieval. 2 (1), Pages 1-35.

[52] Lee, W.S., Liu, B., Li, X. and Yu, P.S. (2004) *Text Classification by Labeling Words*. In proceedings of the Nineteenth National Conference on Artificial Intelligence, 16th Conference on Innovative Applications of Artificial Intelligence (AAAI/IAAI-04, AAAI/MIT Press), Pages 425 – 430,San Jose, CA, USA. (ISBN 0-262-51183-5).

[53] Liu , B. (2006) *Searching Opinions in User-Generated Contents*. Invited talk at the Sixth Annual Emerging Information Technology Conference (EITC-06), Dallas, Texas.

[54] Liu, B. (2007) *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer-Verlag Berlin Heidelberg. (ISBN 3-540-37881-2).

[55] Liu, B. (2011) *Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data*, Springer Series Data-Centric Systems and Applications, 2nd Edition.

[56] Liu, B. (2012) *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies.

[57] Lovins, J. (1968) *Development of a Stemming Algorithm*. Mechanical Translation and Computational Linguistics, Vol. 11, No. 1&2, Pages 22-31.

[58] Nguyen, G., Hoang., Bouzerdoum, A. and Phung, S. (2009) *Learning Pattern Classification Tasks with Imbalanced Data Sets*. In Proceedings of Yin (Eds.), Pattern Recognition, Pages 193-208. Vukovar, Croatia.

[59] Ott, M., Choi, Y., Cardie, C. and Hancock, J.T. (2011) *Finding Deceptive Opinion Spam by any Stretch of The Imagination*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies – Vol. 1, Page 309, USA.

[60] Paice, Ch. (1996) *Method for Evaluation of Stemming Algorithms Based on Error Counting*. Journal of the American Society for Information Science, Vol. 47, No. 8, Pages 632 – 649.

[61] Porter, M. (1980) *an Algorithm for Suffix Stripping*. Program, Vol. 14, No. 3, Pages 130-137.

[62] Ryding, K. (2005) *A Reference Grammar of Modern Standard Arabic*. From http://bilder.buecher.de/zusatz/14/14749/14749960_vorw_1.pdf 2005.

[63] Sun, Y., Kamel, M.S., Wong, A.K.C. and Wang, Y. (2007) *Cost-Sensitive Boosting for Classification of Imbalanced Data*. Pattern Recognition, Vol. 40, No. 12, Pages 3358-3378.

[64] Tan, P., Steinbach, M. and Kumar, V. (2006). *Introduction to Data Mining*. Addison-Wesley, Reading, MA.

[65] Turney, P.D. (2002) *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, Pennsylvania, USA.

[66] Wang, S. (2008) *Class Imbalance Learning*. Thesis Proposal.

[67] Washio, T., Kok, J.N. and Raedt, L.C. (2005) *Advances in Mining Graphs, Trees And Sequences*. IOS press. (ISBN 1586035282).

[68] Yen, S. and Lee, Y. (2009) *Cluster-based Under-Sampling Approaches for Imbalanced Data Distributions*. Expert Systems with Applications, Pages 5718-5727.

[69] Zaiane, O. (1999) *Introduction Survey to Data Mining*. CMPUT690 Principles of Knowledge Discovery in Databases, University of Alberta.

[70] Zhang, J. and Mani, I. (2003). *KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction*. In Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Data Sets.

[71] http://en.wikipedia.org/wiki/Email_spam (Access Online).

[72] http://en.wikipedia.org/wiki/Web_spam (Access Online).