نموذج رقم (1)

إقـــــرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Detecting Subjectivity in Staff Performance Appraisals using Text Mining

(Teachers' Appraislas of Palestinian Government Case Study)

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وإن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل لنيل درجة أو لقب علمي أو بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

DECLARATION

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted elsewhere for any other degree or qualification

Student's name: اسم الطالب: أماني علي عبد الله عابد

Signature: التوقيع: Amani Abed

Date: التاريخ: 2015/12/19

Islamic University – Gaza

Deanery of Higher Studies

Faculty of Information Technology

Detecting Subjectivity in Staff Performance Appraisals Using Text Mining

(Teachers' Appraisals of Palestinian Government Case Study)

Submitted by
Amani A. Abed

Supervised by
Prof. Alaa M. Al Halees

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

November, 2015

بسم الله الرحمن الرحيم

الجامعة الإسلامية – غزة
The Islamic University - Gaza

مكتب نائب الرئيس للبحث العلمي والدراسات العليا هاتف داخلي: 1150

الرقم. ج.س.غ/35/..... Ref

التاريخ 2015/11/16م..... Date

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة شئون البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحثة/ **أماني علي عبدالله عابد** لنيل درجة الماجستير في **كلية *تكنولوجيا المعلومات*** **برنامج تكنولوجيا المعلومات** وموضوعها:

اكتشاف عدم الموضوعية في تقييمات أداء الموظفين بإستخدام التنقيب عن النصوص

Detecting Subjectivity in Staff Performance Appraisals Using Text Mining

وبعد المناقشة العلنية التي تمت اليوم الثلاثاء 05 صفر 1437هـ، الموافق2015/11/17م الساعة العاشرة صباحاً بمبنى اللحيدان، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

أ.د. علاء مصطفى الهليس	مشرفاً و رئيساً	.....
د. راوية فوزي عوض الله	مناقشاً داخلياً	Rawia Awad Allah
د. أحمد يحيى محمود	مناقشاً خارجياً	.....

وبعد المداولة أوصت اللجنة بمنح الباحثة درجة الماجستير في **كلية *تكنولوجيا المعلومات*/ برنامج تكنولوجيا المعلومات.**

واللجنة إذ تمنحها هذه الدرجة فإنها توصيها بتقوى الله ولزوم طاعته وأن تسخر علمها في خدمة دينها ووطنها.

والله ولي التوفيق،،،

نائب الرئيس لشئون البحث العلمي والدراسات العليا

أ.د. عبدالرؤوف علي المناعمة

P.O. Box 108, Rimal, Gaza, Palestine fax: +970 (8) 286 0800 فاكس Tel: +970 (8) 286 0700 هاتف ص.ب. 108 الرمال، غزة، فلسطين
public@iugaza.edu.ps www.iugaza.edu.ps

Acknowledgements

(وَمَا رَمَيْتَ إِذْ رَمَيْتَ وَلَكِنَّ اللَّهَ رَمَى وَلِيُبْلِيَ الْمُؤْمِنِينَ مِنْهُ بَلاءً حَسَناً إِنَّ اللَّهَ سَمِيعٌ عَلِيمٌ) [الأنفال: 17]

First and foremost, I thank ***Allah (subhana wa taala)*** for endowing me with health, patience, and knowledge to complete this work and made me lucky to be supervised by such a professor as ***Prof. Alaa Alhalees.*** It gives me pleasure to thank my supervisor ***Prof. Alaa Alhalees*** for all he taught me, without his help, sage advice, insightful criticism, and continuous follow-up; this research would never have been. I also would like to take this opportunity to express my deepest gratitude to the academic staff of information technology program at the Islamic University-Gaza.

Also, I am so grateful to the domain experts***: Eng. Iyad Abu Safia, Eng. Osama Younis*** and ***Eng. Osama Qassem*** whom helped me very much in all the phases of my research from idea formation to judgment of results. I would like to thank them for their time and encouragement.

I wish to express my considerable gratitude to many people who, in one way or another, have helped with the process of doing this research. Very special thanks to ***my mother*** and ***my father*** for all they do for me, without their pray, patience and encouragement I can't do this work. Thanks to ***my brothers*** and ***sisters, my friends, managers*** and ***coworkers*** whom I consider as brothers and sisters.

Abstract

As human resources are the resources that carry out many important activities in an organization, Human Resource Management (HRM) should catch up with the latest developments to manage these resources efficiently. Staff appraising is one of the most important roles of HRM. Accurate appraising systems promise organizations of a plethora of benefits. Right managerial decisions and staff's perception of fairness are some of these benefits. Non-subjective appraising is such a characteristic of accurate appraising systems. However, almost already applied processes for ensuring non-subjectivity in staff appraisals are manual, infeasible, hard and time consuming. For large organizations with large number of staff such as the Palestinian government, it become more difficult.

A considerable effort has been directed to detecting subjectivity in opinion reviews. However, to the best of our knowledge, there is no previous work that detect subjectivity in staff appraisals. Our contribution in this work is to use text mining methods in finding context and domain driven clues of subjectivity in staff appraisals.

The objective of this work is to propose a text mining based approach that supports HRM in detecting subjectivity in staff performance appraisals. The approach detects three clues of subjectivity in reviews, where each clue represents a level of subjectivity. First level, textual reviews that are irrelevant to the domain of staff appraising. Second level, duplication and near duplication in reviews. Third level, textual reviews that do not provide significance meaning; nearly a duplication of items in the non-textual part of appraisal.

For proving our approach, we applied it on the teachers' staff appraisals of the Palestinian government. According to our experiments, we found that the approach is effective regarding our evaluations; where we used expert opinion, precision, recall, accuracy and F-measure. In the first level, we reached the F-measure of 88%, in the second level, we used expert staff's opinion, where they decided the percent of duplication to be 85% and in the third level, we achieved the best average F-measure of 84%.

Keywords: *Staff Appraisal, Subjectivity Detection, Opinion Mining, Text Mining, Human Resource Management.*

عنوان البحث:

اكتشاف عدم الموضوعية في تقييمات أداء الموظفين باستخدام التنقيب عن النصوص

(تقييمات المعلمين في الحكومة الفلسطينية كدراسة حالة)

ملخص

إن الموارد البشرية هي تلك الموارد التي تؤدي العديد من الأنشطة في أي مؤسسة، لذا فإنه لمن المهم لإدارة الموارد البشرية أن تواكب اخر التطورات لإدارة هذه الموارد بشكل كفؤ. وإن تقييم الموظفين هو واحد من أهم الوظائف في إدارة الموارد البشرية. كما أن أنظمة التقييمات المعدة بشكل سليم تعد المؤسسات بالعديد من الفوائد. من أهمها القرارات الإدارية الصائبة و شعور الموظفين بالعدل. وتُعدّ التقييمات الموضوعية كسمة من سمات التقييمات السليمة. ومع ذلك فإن الطرق المتبعة للتأكد من موضوعية التقييمات غالبا ما تكون يدوية، غير مجدية، صعبة و تستغرق الكثير من الوقت. ولمؤسسة كبيرة بها عدد كبير من الموظفين كالحكومة الفلسطينية فإن العملية تصبح أكثر صعوبة.

لقد وجهت جهود معتبرة لكشف عدم الموضوعية في تحليل النصوص. ومع ذلك على أفضل علمنا، فإنه ليس هناك أعمال سابقة لكشف عدم الموضوعية في تقييمات الموظفين. وإن مساهمتنا في هذا العمل هي استخدام تحليل النصوص في إيجاد دلائل مشتقة من السياق والمجال لعدم الموضوعية في تقييمات الموظفين.

الهدف من هذا البحث هو اقتراح آلية مبنية على تحليل النصوص لتدعم إدارة الموارد البشرية في كشف عدم الموضوعية في تقييمات الموظفين. تكتشف الآلية ثلاثة دلائل من دلائل عدم الموضوعية، حيث إن كل دليل يمثل مستوى من عدم الموضوعية. المستوى الأول: التقييمات النصية التي ليس لها علاقة بمجال تقييم الموظفين، المستوى الثاني: هو التكرار الكلي أو الجزئي للتقييمات النصية، المستوى الثالث: هو التقييمات النصية التي لا تعطي معنى ذو مغزى أي أنها تقريبا تكرار لبنود التقييم غير النصية.

كإثبات لكفاءة الآلية المقترحة قمنا بتطبيقها على تقييمات المعلمين في الحكومة الفلسطينية، وبناءً على التجارب التي قمنا بها فقد وجدنا أن هذه الآلية فعالة على حسب المقاييس التي استخدمناها، ففي المرحلة الأولى من كشف عدم الموضوعية وصلنا الى دقة 88%، وفي المرحلة الثانية حيث استخدمنا رأي الخبراء في المجال فقد اخترنا نسبة التكرار لتكون 85%، وفي المرحلة الثالثة وصلنا لأفضل متوسط دقة 84%.

الكلمات الأساسية: تقييمات الموظفين، كشف عدم الموضوعية، تحليل الآراء، تحليل النصوص، إدارة الموارد البشرية.

Table of Contents:

Abstract……III

Chapter One Introduction……....1

1.1 The case study……2

1.2 Staff appraising problem……3

1.3 Definition of subjectivity……...…3

1.4 Clues of subjectivity……..3

1.5 Text Mining……...4

1.6 Opinion Mining……..4

1.7 Definition of Subjectivity……..5

1.8 Research problem……..5

1.9 Research objective……..……5

1.9.1 Main objective……5

1.9.2 Specific objectives…….……5

1.10 Importance of the work……...……6

1.11 Scope and limitation……6

1.12 Methodology……7

1.13 Thesis organization……..9

Chapter Two Theoretical Foundation……10

2.1 Human resources management and staff appraisals……11

2.2 Knowledge Discovery in Database…….……11

2.3 Text Mining……13

2.4 Unsupervised Vs. Supervised Learning……..……..14

2.4.1 Unsupervised learning…….……14

2.4.2 Supervised learning…….……14

2.5 Text Categorization…….…..……15

2.6 Classifiers…….……15

2.6.1 Support Vector Machine ……..……15

2.6.2 K-Nearest Neighbor……..……17

2.6.3 Naïve Bayes……..……17

2.7 Opinion Mining……..…….19
2.8 Subjectivity and objectivity…….19
2.9 Subjectivity detection methods……20
2.10 Subjectivity and Sentiment Classification…….21
2.11 Summary……21
Chapter Three State of the Art……...22
3.1 Data mining for human resource management……23
3.2 Text mining for human resource management……24
3.3 Opinion Mining……26
3.3.1 Applications of opinion mining……26
3.3.2 Opinion mining for Appraisal systems……,…27
3.4 Subjectivity Detection……,.28
3.5 Summary……,30
Chapter Four Proposed Approach……..,32
4.1 Understand the business domain……..33
4.2 Define the Appraising process……,33
4.3 Data acquisition……..,.36
4.4 Data set understanding……36
4.5 Identify clues of subjectivity……...36
4.6 Text preprocessing……...37
4.7 Apply mining processes……...39
4.7.1 Feature extraction for generating objective wordlist……39
4.7.2 Similarity Measurement……39
4.7.3 Machine Learning processes (classification) ……...40
4.7.3.1 Support Vector Machine……40
4.7.3.2 Naïve Bayse……..……40
4.7.3.3 K-Nearest Neighbor……...41
4.7.4 Subjectivity Detection……..41
4.8 Evaluation……43
4.8.1 Domain expert judgment……..43
4.8.2 Accuracy measurements……...43

4.8.3 10-Fold cross validation……44
4.9 Summary……..45
Chapter Five Experiments and Results……46
5.1 Experiments settings……47
5.2 Rapidminer……..47
5.3 The Data Set……47
5.4 First level of subjectivity……47
5.4.1 Data preprocessing……48
5.4.2 Feature extraction for generating objective wordlist……49
5.4.3 Subjectivity Detection……..52
5.4.4 Evaluation and results……..53
5.5 Second level of subjectivity……54
5.5.1 Preprocessing data……54
5.5.2 Similarity Measurement……55
5.5.3 Subjectivity Detection……...55
5.5.4 Evaluation and results……...57
5.6 Third level of subjectivity……58
5.6.1 Data preprocessing……59
5.6.2 Machine Learning processes (classification) ……...59
5.6.2.1 Support Vector Machine……59
5.6.2.2 Naïve Bayse ……..59
5.6.2.3 K-Nearest Neighbor……...60
5.6.3 Subjectivity Detection……..60
5.6.4 Evaluation……60
5.7 Summary……..62
Chapter Six Conclusion and Future Works……...63
6.1 Conclusion……...64
6.2 Future Works……...64
References……66
Appendix A Objective Wordlist……74

List of tables:

Table 5.1: Number of tokens after each step of preprocessing……49
Table 5.2: Splits of data into input data and testing data……..51
Table 5.3: measurements of accuracy with different values for threshold……54
Table 5.4: number of employees for each reviewer for second level of subjectivity detection…54
Table 5.5: the results of classification for first fold validation…….61
Table 5.6: average of 10 fold classification……..61
Table A.1: objective word list for first level of subjectivity…….75

List of figures:

Figure 1.1: Appraisal analyzer steps……7
Figure 2.1 Overview of the steps constituting the KDD process……...12
Figure 2.2: Support Vectors……...16
Figure 4.1 staff appraising process according to domain experts' opinion……...34
Figure 4.2: Appraisal analyzer steps……..35
Figure 4.3: representation of two documents in 2-D space……...39
Figure 5.1: Preprocessing steps on Rapidminer……48
Figure 5.2: Sub processes of Process document from data process……...50
Figure 5.3: Processing Documents Process…….....50
Figure 5.4: examples of real reviews with their relevance percent……53
Figure 5.5: Similarity Measurement Process……55
Figure 5.6: Examples of similar real answers with their similarity percent……..57
Figure 5.7 Applying SVM for classification……59
Figure 5.8 Applying NB for classification……59
Figure 5.9 Applying KNN for classification……60
Figure 5.10 applying classification model for data……60

List of abbreviations:

HRM	Human Resource Management.
TM	Text Mining.
OM	Opinion Mining.
KDD	Knowledge Discovery in Databases.
DM	Data Mining.
VSM	Vector Space Model.
TF-IDF	Term Frequency Inverse Document Frequency.
TC	Text Categorization.
SVM	Support Vector Machine.
KNN	K-Nearest Neighbor.
NB	Naïve Bayes.
HRMS	Human Resource Management Systems
HP-Subj	High-Precision Subjectivity Classifiers

Chapter One

Introduction

As its role to create, implement, oversee policies governing employees' behavior and the behavior of the organization toward its employees, Human Resource Management (HRM) should catch up with the latest development of IT in order to play its' role efficiently. It should provide insights for investments on business human capital. As a one intelligent solution is using data mining techniques [1].

Performance evaluation is one of the most crucial issues of HRM, it is a systematic way of reviewing and assessing the performance of an employee during a given period of time [2]. And accordingly, top management makes a lot of decisions upon these appraisals. However, the problem is how does the top management get the overall picture that enables them to make decisions? Moreover, what monitoring procedures are they doing to monitor these appraisals? Is it reliable? Our concern in this research is to propose a text mining based approach that support monitoring staff appraisals by detecting subjectivity.

1.1 The case study:

Our case study is on teachers' appraisals in the Palestinian government. Palestinian government has a large number of ministries and organizations with a large number of employees working in many fields. These organizations and ministries are following the same policies in staff appraising. These policies are set and monitored by the General Personnel Council. There is more than 70 staff appraising template in the government, each for a field. All these templates have the same structure, which consists of two parts: one called a basic form, which is a list of weighted items to be given marks from zero to 100%. These items measures staff commitment, performance, technical skills and supervising skills. The other part, called additional form, consists of textual items that includes opinion of the reviewer, manager who is in charge of evaluating his employees, about staff weakness points if his mark is less than 65% and strength points, contributions and accomplishments if his mark is greater than 85%. The first part, basic form, is dynamic for each field, while the second part, additional form, is constant for all fields, and it is required to be answered if employees' mark is greater than 85% or less than 65%.

In our work, we used the additional form of teachers' appraisals for teachers with marks greater than 85%, so that we could have a large data set. We took the appraisals from the General Personnel Council and took their permission to apply our approach on these appraisals.

The already exist methods for monitoring staff performance appraisals and detecting subjectivity concentrates only on the basic form and they have nothing to do with the additional form. In this research, we worked on the additional form.

1.2 Staff appraising problem:

Despite the importance of performance appraising for organizations, it is not on the top of the list of "favorite things to do" for managers [3]. Appraising systems are experiencing many problems such as managers are conducting performance appraisals carelessly, not being trained to conduct effectively, and in some cases they conduct it with some bias against certain groups of people on non-job-related grounds. In this work, we used the term subjectivity in staff performance appraisals to address these problems.

1.3 Definition of Subjectivity:

According to literary theory, subjectivity is a term for linguistic expression of private states [4]. Quirk et al. [5] gave the general term private state, for referring to the mental and emotional states of the writer or speaker. Private states is defined as something that is not open to objective observation or verification [6]. This definition of subjectivity complies with the staff appraising problem we discussed in 1.2. We detect subjectivity in appraisals by finding one of the three clues that we identify in section 1.4.

1.4 Clues of subjectivity:

According to domain experts' opinion, an appraisal is considered to be subjective, if it contains only one or more of the clues of subjectivity. Each clue represents a level of subjectivity.

The clues are as follows:

- **Irrelevance:** this clue represents the lowest level of subjectivity, where reviewers' answers are irrelevant to the domain of teachers' appraisals.
- **Duplication:** is another level of subjectivity, where reviewer is duplicating or near duplicating the same answer to different employees.
- **Insignificance:** a higher level of subjectivity where reviewers' answers are meaningless to the question.

1.5 Text Mining (TM):

Text mining is the process of analyzing large quantities of natural language text and detects lexical or linguistic usage pattern in an attempt to extract probably useful information [7].

Natural language text may represent the majority of information available to a particular research or data mining project [8]. One of the very common applications of text mining is analyzing open ended survey or appraisal responses where respondents are permitted to express their opinions without constraining them to particular dimensions or particular response format [8]. To teach computers how to analyze and understand natural language, text mining technologies like information extraction, summarization, categorization, classification and clustering are used [9]. In our work, we used text mining technologies to detect subjectivity in staff performance appraisals. Subjectivity detection is a one type of opinion mining; which is a subfield of text mining. Yet, we had our own clues of subjectivity that is derived from domain and context. These clues, as we had discussed in section 1.3, are irrelevance, duplication and insignificance. The definition of insignificant reviews is similar to the definition of subjectivity detection of opinion mining, which will be discussed in section 1.6.

1.6 Opinion Mining (OM):

Opinion mining is an interdisciplinary field that combines natural language processing and text mining. It is basically people's opinion study, study of emotions and appraisals in the direction of any social issue, people or entity [10]. Unlike text categorization of text mining, opinion mining have relatively few classes (e.g., "positive" or "negative") that generalize across many domains and users [11] . Despite the little number of classes in opinion mining, it is not a simplified task of text categorization, as the complexity of the natural language processing is inherited to this field. There are two different types of text classification in opinion mining: subjectivity detection and polarity detection. In subjectivity detection the task is to determine whether a given text represents an opinion or a fact, or more precisely whether given information is factual or nonfactual, whereas the aim of polarity detection is to find whether the opinion expressed in a text is positive or negative [12].

1.7 Definition of Subjectivity:

An objective sentence presents some factual information about the world, while a subjective sentence expresses some personal feelings, views, or beliefs. An example of objective sentence is "iPhone is an Apple product." An example of subjective sentence is "I like iPhone." [13]
Subjective remarks come in a variety of forms, including opinions, rants, allegations, accusations, suspicions, and speculations [14].

Our definition of the third level of subjectivity detection "insignificant reviews" is similar to this definition, as we detect opinionated answers where the ideal answer is to mention staff's accomplishments (factual information).

1.8 Research problem:

Already applied manual processes for monitoring objectivity of staff appraisals is inaccurate, unfair, infeasible, hard and time consuming, and for large organizations with large number of staff such as the Palestinian government it becomes more difficult. To the best of our knowledge, there is no previous work that detect subjectivity in staff appraisals.

1.9 Research objective:

1.9.1 Main objective:

Our main objective is to develop an approach that supports staff appraisals systems in monitoring the objectivity of the appraising process. The proposed approach is based on text mining techniques. We applied our approach on teachers' appraisals of the Palestinian government.

1.9.2 Specific objectives:

- Study the already applied processes for monitoring staff appraisal systems.
- Choose and specify the more suitable data set.
- Mine the reviews from the additional form of the appraisals in order to find clues of subjectivity.
- Classify appraisals as objective or subjective in three levels.

- Evaluate the effectiveness and accuracy of the proposed approach using metrics such as recall, precision, f- measure and accuracy.

1.10 Importance of the work:

On the one hand, the assumption appears to be that an effectively designed, implemented, and administered performance evaluation system can provide the organization, the manager, and the employee with a plethora of benefits [2]. On the other hand, the perception of unfairness can negatively influence employee loyalty and role-related behaviors. However, in spite of the attention and resources applied to the practice, dissatisfaction with the process still abounds and systems are often viewed by employees as inaccurate and unfair [2].

Our work contributes very much in enhancing the monitoring process of staff appraising by detecting review's subjectivity, as it will overcome the limitations of the manual monitoring process, which is tedious, hard, and time consuming as well as, it is almost impossible to provide management with the overall picture of the appraisals' objectivity.
In addition, it will support managers in decision making, e.g. employees who had been appraised subjectively, will be given higher concern if they appeal their appraisals.

1.11 Scope and limitation

- This research proposes an approach to detect subjectivity in staff appraisals.
- We used the appraisals of teachers in the Palestinian government for the years 2012 and 2013 to evaluate the proposed approach.
- Another issue is the definition of subjectivity; in our work, an appraisal is considered to be subjective if only it contains one or more of these clues:
  - Reviews are irrelevant to the domain.
  - Review is a duplication or near duplication of other reviews of the same reviewer.
  - Reviews do not provide significant meaning to the question in the textual part (additional form).
- Also, the approach considers only the Arabic texts, it doesn’t support other languages.

1.12 Methodology:

In this work, we understood the business domain from the domain expert staff, analyzed how the appraising process is accomplished and how it should be accomplished.

Our work, is proposing an approach for the appraisal analyzer, the process that analyzes appraisals, so that we could detect subjective appraisals according to it.

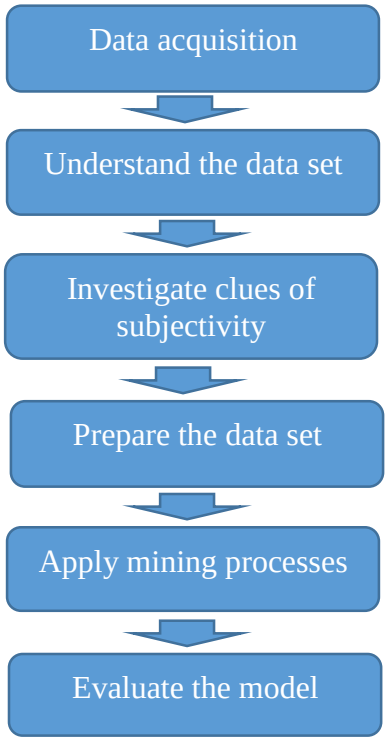

Figure 1.1: Appraisal analyzer steps

The steps of the appraisal analyzer approach, as illustrated in figure 1.1, are as follows:

- **Data acquisition:**

  As we mentioned in section 1.10, we took the appraisals of teachers in the Palestinian government for the years of 2012 and 2013, consisting of around 4400 records.

- **Data set understanding:**

  We worked with the domain expert staff, to understand the data and domain and analyze the problems in appraisals; how do reviewers answer appraisal questions and how should they answer.

- **Investigate clues of subjectivity:**

  From our understanding of the data set, we came up with three clues of subjectivity, where each clue represents a level of subjectivity:

  - Irrelevance to the domain; that is, teachers' appraisals. For example: a manager repeat the question or types words that irrelevant to the domain such as: **"اذكر الاعمال البارزة للموظف التي أدت الى تجاوز معدلات الأداء"**
  - Duplication or near duplication of the same review to many employees.
  - Insignificant meaning answers that are nearly duplication of items in the basic form. Such as: "**المعلمة نشيطة و مجتهدة وتحضر باكرا**".

- **Prepare the data set:** In this step, we prepared the data set, answers of reviewers, for applying the mining algorithms. We used methods for text preprocessing such as tokenization, stemming, removing stop words and term weighting, as well as labeling the data manually as subjective or objective.

- **Apply mining processes:**

  This process is the core part of our approach, where we used different mining processes for each level. For the first level, we used feature extraction using unigrams and bigrams for generating an objective wordlist, in order to compare reviews with this wordlist. For the second level, we used similarity measurement in order to detect duplicated and near duplicated reviews. For the third level, we used classification, in order to detect reviews with insignificant meaning.

- **Evaluation**

  We evaluated our approach by using the measurements of: expert staff judgment, accuracy, precision, recall and f measure.

1.13 Thesis organization:

Thesis research is composed of six chapters. Chapter 2 presents the theoretical information of the research. Chapter 3 introduces and discusses some related work. Chapter 4 presents our proposed approach. Chapter 5 presents the experimental results and the evaluation of the approach. Finally, Chapter 6 summarizes the work and outlines possible further extensions to the current work.

Chapter Two

Theoretical Foundation

This chapter presents the fundamental concepts, which represents the basis for understanding of the thesis work. First, we discussed human resource management and staff appraisals, how Knowledge Discovery in Databases (KDD) supports making decisions according to staff appraising. Then we discussed the steps of KDD, followed by, text mining, unsupervised and supervised learning, text categorization, opinion mining, subjectivity and objectivity, subjectivity detection methods, subjectivity and sentiment classification.

2.1 Human resources management and staff appraisals:

It is as true for the Government as for any other organization that human resources are the resources that carry out many important activities in the organization. Successful implementation of performance management will not only help managers to promote the management level and efficiency, but also achieve the organization's strategic objectives.

Human Resource Management is a long-established task within the Government's Management Framework [15]. Through this task the Government meets its obligation to be a good employer; seeks to secure staff commitment; and develops and manages staff to give of their best to help the Government serve the community [15]. Nevertheless, Management guru Peter Drucker famously said: "What gets measured, gets managed" [16]. Performance appraisal is a one powerful measurement tool. It assesses an individual's performance against previously agreed work objectives [15]. However, its purpose is more than identifying individuals' strength and weakness. It enables management to make crucial decisions for achieving the organization's strategic objectives. Therefore, staff appraisal systems should be monitored to ensure right decisions. These decisions are not only for planning strategies for the organization, but also for correcting the process of staff appraising itself. In order to make a decision, managers need knowledge. In case of massive data amounts, issues may occur because of data analysis and necessary knowledge extract. Data is analyzed through an automated process, known as knowledge discovery in data mining techniques [17].

2.2 Knowledge Discovery in Database (KDD):

Knowledge discovery in database is the non-trivial process of identifying valid, potentially useful, and ultimately understandable patterns in data [18].

The process of KDD, as illustrated in figure 2.1, could be summarized as [18]:

- **Understanding the domain of the business:** this includes the relevant prior knowledge and the goals of the application.
- **Producing a dataset:** includes selecting a dataset on which discovery is to be performed.
- **Preparing the data set:** includes data cleaning and preprocessing in which incomplete, noisy and inconsistent data is handled, also the data set is preprocessed and transformed so that data mining algorithms could be performed.
- **Data mining:** data mining algorithms and methods are applied in order to extract data patterns.
- **Pattern evaluation:** to identify the truly interesting patterns representing knowledge based on some interesting measures.
- **Knowledge presentation:** visualization and knowledge representation techniques are used to present the mined knowledge to the user.

As could be shown in figure 2.1, in the KDD process, one typically iterates many times over previous steps and the process is fairly messy with plenty of experimentation [18] .

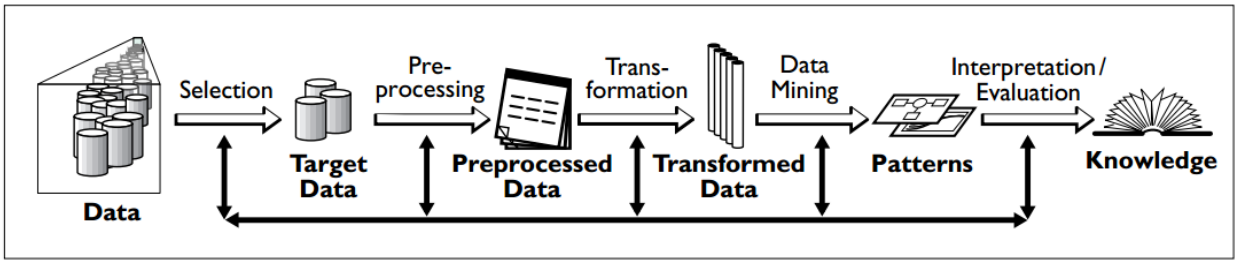

Figure 2.1 Overview of the steps constituting the KDD process [18]

Data mining (DM) is the application of specific algorithms for extracting patterns from data [19]. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporating appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data [19]. Blind application of data mining methods (rightly criticized as "data dredging" in the statistical literature) can be a dangerous activity easily leading to discovery of meaningless patterns [19].

However, many people tends to shorten the term of "knowledge Discovery in Databases" to "Data Mining", as it is the most essential process in KDD, so it is treated as the synonym of KDD.

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 [20].

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

2.3 Text Mining (TM):

Text mining is similar to data mining but it is an extended form of data mining. It leads to discovery of new knowledge from large volume of the existing unstructured data [21]. It is also called, as text data mining and knowledge discovery from word-based databases [21].

Typical text mining tasks include text classification (text categorization), text clustering, concept/entity extraction, topic tracking, information visualization, question answering, document summarization etc. [22].

The process of text mining is the same as that of KDD, we discussed earlier, except that the methods in the preprocessing and algorithms in the mining phase could be different.

Some of the methods in preprocessing are as follows:

- **Tokenization:**

  In this process, a sequence of strings is broken into pieces such as words, phrases, symbols and other elements, called tokens, so that, text mining algorithms could be used. Arabic tokenization is complex due to the rich morphological features of Arabic [23].

- **Stemming**

  In this process affixes (prefixes and suffixes) are removed from features. This process is used to reduce the number of features in the feature space and improve the performance of the classifier when the different forms of features are stemmed into a single feature [24]. It tries to find the basic form of the word.

- **Stop word filtering:**

  The idea of stop word filtering is to remove high frequent words that are commonly used in the language and carry no information (i.e. pronouns, prepositions, conjunctions, etc.), so that we could focus on words that are critical to the domain [25]

- **Vector Space Model (VSM):**

  It is an algebraic model for representing text documents as vectors of identifiers (tokens), in m-dimensional space, where m is the number of words or tokens and the value of each element in the vector is represented by one if the corresponding word exists in the document, and zero if it does not exist. VSMs extract knowledge automatically from a given corpus, thus they require much less labor than other approaches to semantics, such as hand-coded knowledge bases and ontologies [26].

- **Term weight of text documents:**

  Term weighting helps us to locate important terms in a document collection for ranking purposes [27]. **Several term weighting schemes are used such as:**

  - **Boolean model:** which indicates the presence or absence of a word with Booleans one or zero respectively.
  - **Term Frequency:** is the number that term t occurs in the document d.
  - **Term Frequency Inverse Document Frequency (TF-IDF):** is a common weight scheme that is more meaningful, where large weights are assigned to terms that are used frequently in relevant documents but rarely in the whole document collection [28][20].

  In our work, we used term frequency for wordlist generation in the first level, and TF-IDF for classification in the third level.

2.4 Unsupervised Vs. Supervised Learning:

There is two types of learning: supervised and unsupervised; both of them were used in this work.

2.4.1 Unsupervised learning: is a technique operates by trying to find hidden structure in unlabeled data by investigating useful relations among the elements of these data. Clustering technique is an example of unsupervised learning [29].

2.4.2 Supervised learning: is a technique in which a training data (observations or measurements), are accompanied by labels or classes constructing a training set, and then is used for creating a decision function so that new data is classified based on the training set [29].

2.5 Text Categorization (TC):

Text categorization, also known as text classification, seeks classifying documents into predefined topics based on their contents. There can be many possible categories, the definitions of which might be user and application dependent; and for a given task, we might be dealing with as few as two classes (binary classification) or as many as thousands of classes (e.g., classifying documents with respect to a complex taxonomy) [11]. Text categorization can be characterized as a supervised learning problem [29].

2.6 Classifiers:

In the set of experiments for detecting the third level of subjectivity, we compared between three classifiers so that we could chose the most suitable for our domain. The classifiers we used are: Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Naïve Bayes (NB).

2.6.1 Support Vector Machine (SVM):

Support vector machine, was introduced as a class of supervised machine learning techniques. It offers one of the most robust and accurate methods among all well-known algorithms [30]. In a two-class learning task, the aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data [30]

Given a set of N linearly separable points $S = \{ x_i \in R^n \mid i = 1,2,.., N \}$, each point x_i belongs to one of the two classes, labeled as $y_i \in \{ -1,+1\}$. A separating hyper-plane separates S into 2 sides, each side containing points with the same class label only. The separating hyper-plane can be identified by the pair (w, b) that satisfies:

$y = w.x + b$ **……..2.1**

and

$$\left\{ \begin{array}{l} w.x_i + b \geq +1 \quad \text{if } y_i = +1 \\ w.x_i + b \leq -1 \quad \text{if } y_i = -1 \end{array} \right\} \quad \text{…… 2.2}$$

For i = 1,2,…, N and where

$$w.x = \sum_i w_i.x_i \qquad \text{……..……2.3}$$

for vectors **w** and **x**.

The idea of SVM is that; it creates a hyper-plane that separates the data set into two sets with the maximum margin as seen in figure 2.2. The optimal separating hyper-plane that has the maximum margin to both sides is identified by the formula 2.4:

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 \qquad \text{……..…..2.4}$$

Subject to

$$\left\{ \begin{matrix} w.x_i + b \geq +1 & \text{if } y_i = +1 \\ w.x_i + b \leq -1 & \text{if } y_i = -1 \end{matrix} \right\} \text{ for } i = 1, 2, \ldots, N \qquad \text{..……2.5}$$

The reason why SVM insists on finding the maximum margin hyper-planes is that it offers the best generalization ability. It allows not only the best classification performance (e.g., accuracy) on the training data, but also leaves much room for the correct classification of the future data [30]

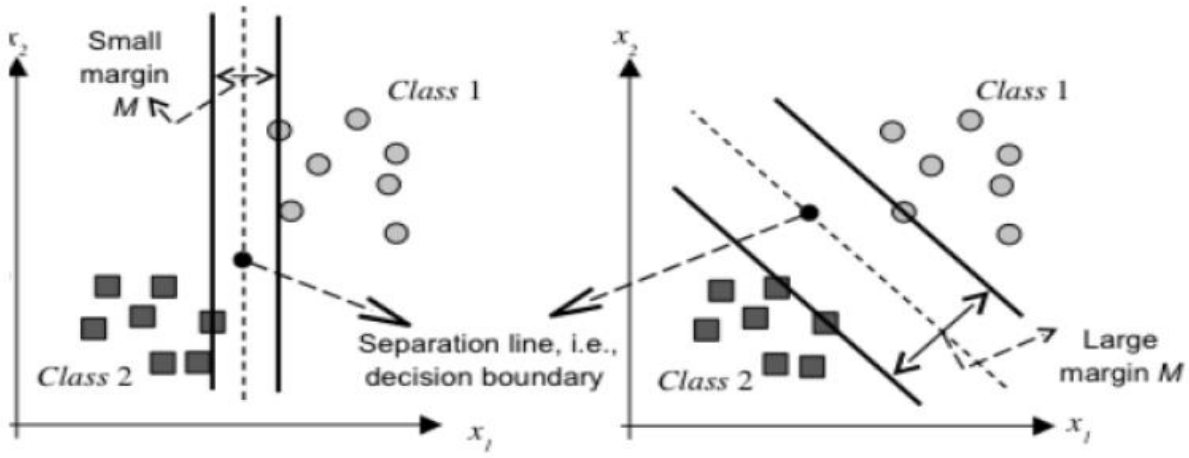

Figure 2.2: Support Vectors [31]

2.6.2 K-Nearest Neighbor (KNN):

K-nearest neighbor finds a group of *k* objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighborhood. To classify an unlabeled object, the distance of this object to the labeled objects is computed, its *k*-nearest neighbors are identified, and the class labels of these nearest neighbors are then used to determine the class label of the object. Once the k-nearest neighbor list is obtained, the test object is classified based on the majority class of its nearest neighbors:

$$\text{Majority Voting: } y' = \underset{v}{\operatorname{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i), \quad \text{.....2.6}$$

where v is a class label, yi is the class label for the i^{th} nearest neighbors, and I (·) is an indicator function that returns the value one if its argument is true and zero otherwise [30].

2.6.3 Naïve Bayes (NB):

This method is important for several reasons. It is very easy to construct, not needing any complicated iterative parameter estimation schemes. This means it may be readily applied to huge data sets. It is easy to interpret, so users unskilled in classifier technology can understand why it is making the classification it makes. And finally, it often does surprisingly well: it may not be the best possible classifier in any particular application, but it can usually be relied on to be robust and to do quite well [30].

The NB classifier, works as follows [20]:

- 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, x2,…xn)**, n measurements made on the tuple from **n** attribute, respectively, **A1,A2,…, An.**
- Assume that there are m classes, **C1, C2,…, Cm**. 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 $\mathbf{C_i}$ if and only if:

$$P(C_i|X) > P(C_j|X) \quad \text{for } 1 \leq j \leq m, j \neq i \quad \text{..... 2.7}$$

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is the maximized, is called the maximum posteriori hypothesis. By Bayes' theorem (Equation 2.8),

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad \text{.....2.8}$$

- 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 that attributes are conditionally independent (no dependence relation between attributes), **$P(X|C_i)$** using Equation 2.9.

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad \text{.....2.9}$$

Equation 2.9 reduces the computation cost, only counts the class distribution.

If A_k is categorical, **$P(X_k|Ci)$** is the number of tuples in C_i having value X_k for A_k divided by **|Ci, D|** (number of tuples of Ci in D).

And if A_k is continuous-valued, **$P(x_k|C_i)$** is usually computed based on a Gaussian distribution with a mean μ and standard deviation **σ** and **$P(X_k|C_i)$** is

$$P(X|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \quad \text{.....2.10}$$

$$g(x_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{.....2.11}$$

Where **μ** 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.

2.7 Opinion Mining (OM):

Opinion mining is a mixture field of natural language processing and text mining. It analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [13]. Unlike text categorization of text mining, opinion mining have relatively few classes (e.g., "positive" or "negative") that generalize across many domains and users [11]. Despite the little number of classes in opinion mining, it is not a simplified task of text categorization, as the complexity of the natural language processing is inherited to this field.

Levels of analysis:

Opinion mining has been investigated at three levels of granularity:

- **Document level:** where the whole document is analyzed and assigned a positive or negative sentiment. This type of analysis assumes that each document is about one entity; it does not deal with comparative opinions [13].
- **Sentence level:** in this level, each sentence is analyzed separately as positive, negative or neutral opinion. Neutral means no opinion [13].
- **Entity and aspect level:** this level is a finer-grained analysis where instead of looking at the whole document or sentences, we look at the opinion itself. Putting in mind that opinion consists of sentiment (positive or negative) and a target (aspect) of the opinion [13].

In our work, we used the document level of analysis where we considered each review as a document. Because we want to assess the way managers are making appraisals; are they appraising their employees subjectively or objectively? We are not interested in what aspects are managers talking about.

2.8 Subjectivity and objectivity:

An objective sentence presents some factual information about the world, while a subjective sentence expresses some personal feelings, views, or beliefs [32]. For example, "this teacher is the head of the health committee in the school" is an objective sentence, while "this teacher is one of the most active teachers in the school" is a subjective sentence. The task of determining whether a sentence is subjective or objective is called subjectivity classification. Although, it is common to

relate subjectivity to opinionated, in some cases objective sentences indicates some positive or negative opinions (e.g. “Every day, this teacher arrives before the working time by half an hour”, indicates a positive sentiment due to this desirable fact). On the other hand, there is some cases that subjective sentences does not have any opinion (e.g. “I think I will go home”). However, we extended the meaning of subjectivity to three levels, driven from context and domain.

2.9 Subjectivity detection methods:

There is two main methods for subjectivity detection:

- **Rule based methods:**

  Rule based methods is a kind of unsupervised learning. These methods rely on subjectivity lexicons, where lexicons are built and data set is compared to these lexicons. A document is detected to be subjective if it contains a predetermined number of words from the compared lexicon. There are widely used lexicons, such as OpinionFinder, Sentiwordnet , and general inquirer [33]. However, the problem with these lexicons is that these lexicons do not support many languages such as Arabic, not to mention that we have our own definition of subjectivity that is driven from domain and context. In such a case, lexicons could be extracted from corpora, which includes subjective and objective documents.

- **Supervised methods**:

  Subjectivity detection could be viewed as a special case of text classification, where we have a little number of classes (subjective or objective). So obviously, the same supervised learning methods that commonly used in text classification could be used in subjectivity detection as well [13].

Both of these methods were used in our work, where we extracted a lexicon (wordlist) from corpora for detecting the first level of subjectivity. However, the wordlist we extracted consists of objective words not subjective words; and reviews that do not contain a threshold of these objective words are considered subjective reviews at this level. In addition, similarity measurements that we used in the second level of subjectivity is considered as an unsupervised learning method [34]. In the third level, we used supervised methods where we classified the reviews into meaningful or not.

2.10 Subjectivity and Sentiment Classification:

Recently, subjectivity classification has been considered as an initial task in sentiment classification; that is before classifying the sentence, detect if it expresses an opinion or not (subjective or objective). Although objective sentences are regarded as expressing no sentiment or opinion, some objective sentences expresses desirable or undesirable facts which in turn indicates positive or negative sentiment. Thus, it is more appropriate for the first step to classify each sentence as opinionated or not opinionated, regardless whether it is subjective or objective.

On the other hand, Early research solved subjectivity classification as a standalone problem, i.e., not for the purpose of sentiment classification [13]. In our research, we also regarded it as a standalone problem aiming to monitoring the objectivity of the staff appraising systems.

2.11 Summary:

In this chapter, we gave an overview of the basic theoretical foundation for this work. We gave an overview of human resource management and staff appraising and how Knowledge Discovery in Databases (KDD) could help in monitoring staff appraisals. Then we introduced the processes of KDD. Also, we introduced data mining as a synonym of KDD, text mining as a kind of data mining, then we elaborated the difference between unsupervised and supervised learning After that, we discussed three algorithms for text categorization, then we discussed opinion mining and its levels of analysis, subjectivity and objectivity, subjectivity detection methods, and finally we discussed the relation between subjectivity and sentiment analysis, and how they could be used separately and together.

Chapter Three

State of the Art

This chapter presents some related works that are relevant to our research. To our knowledge, there is no previous work that addressed our research problem. Therefore, we presents our work as one link in a chain of research in the field of enhancing Human Resource Management Systems (HRMS) by using data mining and text mining. Then, we discuss other works that used opinion mining for different applications. Finally, we investigate how subjectivity detection problem was resolved in a variety of domains, as well as, the used techniques for resolving this problem.

3.1 Data mining for human resource management:

Data mining is widely used in diverse areas such as marketing, finance and education but narrowly in HRM. Yet an increasing number of publications concerning data mining research in HRM gives an impression of a prospering new research field [1].

In the management of manpower resource, the job analysis is the basis and performance assessing is the bridge [35]. The widest related field to our work is enhancing HRMS by using data mining. Many researchers have worked in this field; ***Youzheng et. al.*** in [36] developed a human resource management framework to attract and allocate the talents who are the most suitable to their own organization in a construction company. The framework is based on data mining. They explored the association rules between personnel characteristics and work behaviors, including work performance and retention. These rules could be used to identify effective recruitment channels to access construction talents and design the appropriate screening criteria for selecting the right ones for different job functions.

In fact, data mining techniques attract much attention of leading business intelligence vendors such as Oracle, SAP, SAS, Microsoft and IBM. These vendors incorporate analytics and business intelligence features to their HRMS [37].

For example, Oracle has its Oracle Human Resources Analytics. This product includes intelligent features in diverse areas such as monitoring workforce demographics in line with recruitment and retention objectives. Thus, they could analyze efficiency of the entire recruitment process lifecycle and more importantly understand and prevent the drivers of employee turnover.

In the area of targeted workforce development, the product enables them to gain insight into the movement of top and bottom performers in the organization to engage and develop internal talent. In addition, the product takes into concern discovering the learning demands by analyzing course

enrollments by job, delivery methods, and organizations. Also they have some analysis works concerning leave and absence; they get a comprehensive view into employees' current, planned, and historical absence events, thus, they monitor absence trends as a predictor for employee engagement [37].

SAP also uses data mining in its HRMS for talent management analytics and measurement. They analyze employee skills and qualifications, evaluate the efficiency of the recruiting processes and measure the effectiveness of learning programs. In addition, they use data mining for monitoring the progress of aligning employee goals with corporate goals [37].

In SAS also, there is some analytical works; they crunches data on employees who have quit in the past five years; their skills, profiles, studies, and friendships. Then it finds current employees with similar patterns. Another SAS program pinpoints the workers most likely to suffer accidents [38].

In addition, Microsoft has some works; as it uses data mining for finding patterns of success. For example, they study correlations between thriving workers and the schools and companies they arrived from [38].

Also in IBM, research analysts are charting the skills and experience of the entire workforce. Then, studying technology and economic trends. They're trying to predict the skills IBM will need down the road and whether the needed knowhow should be taught or recruited [38].

The narrower related field; which is staff assessment by using data mining has some concern among researchers; for example: ***Hou et. al.*** in [35] used fuzzy data mining technology to analyze the performance assessment of staff in enterprise, grasped the structure of enterprise staff. Patterns resulting from the analysis are used to instruct enterprise that performance to the staff is examined, contribute to policymaker's carrying on the manpower planning, and then increase enterprise's output, so the method improves enterprise's benefit.

3.2 Text mining for human resource management:

As the fact that more than 80% of the data is in some type of unstructured data [39], and the prediction that the amount of textual information double in every three months [40], text mining techniques are also used in HRMS.

A new line of thought suggests that valuable knowledge required for human resource management lies in emails, chat logs and comments on shared documents; that is, in electronic activity of employees [41]. Cataphora, IBM, SAS and Microsoft were at the forefront of the movement [41]. Cataphora a company developing innovative software technologies for finding patterns and anomalies in digital communications such as emails, documents, IM, phone logs, text messages and social networks. It has some innovative works on human resource management and measuring employee productivity. For example, it has developed a fuzzy search algorithm to detect blocks of text that are reused, such as a technical explanation or a document template, reasoning that the employees who produce them are making a comparatively greater impact on the company by doing work that others deem valuable [41]. In another application, Cataphora investigates the relationships between people and some topics such as human resources-related topics, marketing issues and product development. One way in which these relationships are useful, is for studying the relationships between people and topics. If an executive is central to communications about product development, marketing, and finance, but marginal to those about sales, it's likely that she or he is out of the loop when it comes to the newest sales tactics [41].

Also, IBM had a huge analytics project that uses 410,000 employees, analyzing 20-plus million emails and instant messages those employees write, as well as, 2 million blog and database entries and 10 million pieces of data that come from knowledge sharing and learning activities. In this project, they analyzed employees' electronic data and creates a networked map of who they're connected to and where their expertise lies. So that employees could search for people with expertise on certain subjects and find the shortest "social path" it would take to connect them. In the future, IBM plans developing the software so that it could provide real-time, expertise-based recommendations: automatically suggesting connections while employees work on a particular task, or helping managers assemble compatible project teams [41].

In addition, text mining techniques are often utilized to monitor the state of health of a company by means of the systematic analysis of informal documents [42]. Microsoft examines internal communications to identify so-called "super connectors," who communicate frequently with other employees and share information and ideas and others who appear to hold them up, so-called bottlenecks [38][41]. Conoco also, refined its system for the monitoring of textual sources like e-mails, internal surveys of employees' opinions, declarations of the management, internal and

external chat lines. All representing important means for sounding the evolution of company culture [42]. Likewise, Google is testing an algorithm that uses employee review data, promotions, and pay histories to identify its workers who feel underused, and therefore are most likely to leave the company. Their goal is to get inside people's heads before they even think about leaving, and to work harder to keep them engaged [41].

In addition, Text mining techniques are used to manage HR strategically, mainly with applications aiming at analyzing staff's opinions, monitoring the level of employee satisfaction, as well as reading and storing CVs for the selection of new personnel [42].

In the area of monitoring productivity, researchers at IBM and Massachusetts Institute of Technology, for example, analyzed the electronic data of 2,600 business consultants and compared their communication patterns with their billable hours. They concluded that the average email contact is worth $948 in annual revenue [41].

3.3 Opinion Mining:

As one of its' roles to study peoples' opinions and appraisals, opinion mining is very related to our work.

Opinion mining field is relatively young; as ***Pang et. al.*** stated in [11] that the term opinion mining appeared first in [43], where ***Dave et. al.*** stated that the ideal opinion-mining tool would process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good). However, topics in opinion mining expanded and many applications appeared.

3.3.1 Applications of opinion mining:

Opinion mining is used for a variety of applications; for example: for monitoring products marketing; as ***Zabin et. al.*** stated in [44] companies can respond to the consumer insights they generate through social media monitoring and analysis by modifying their marketing messages, brand positioning, product development, and other activities accordingly.

Spam detection is another application, ***Abu Hammad et. al.*** in [20] proposed an approach for spam detection in Arabic opinion reviews.

Recommendation systems is another one, by classifying people opinions, recommendation system will not recommend items that received a lot of negative feedback [11].

Politics and governments have also some applications: such as allowing the automatic analysis of the opinions that people submit about pending policy or government regulation proposals [11]. In addition, opinions matter a great deal in politics. Some works has focused on understanding what voters are thinking [11].

3.3.2 Opinion mining for Appraisal systems:

Bloom et. al., in [45] suggested appraisal expression extraction as a fundamental task in opinion mining. Therefore, they proposed a system for extracting and disambiguating adjectival appraisal expressions in English. An appraisal expression is a textual unit expressing an evaluative stance towards some target. The task was to find and characterize the evaluative attributes of such elements. These attributes include: an attitude (which takes an evaluative stance about an object), a target (the object of the stance), and a source (the person taking the stance) which may be implied. In their system, appraisal extraction runs in several independent stages. First, the appraisal extractor finds appraisal expressions by finding the chunks of text that express attitudes and targets. Then, it links each attitude group found to a target in the text. Finally, it uses a probabilistic model to determine which attitude type should be assigned when attitude chunks were ambiguous. For chunking they used manual lexicons, for linking they used a hand constructed linkage specifications and for disambiguation they used bayes theorem. They have applied this system to two domains of text: product reviews, and movie reviews. Manual evaluation of the extraction showed the system works well

A more related work to ours, is monitoring staff appraisals by using opinion mining applications. In [2] ***Suriyakumari et. al.*** proposed a Domain Driven Data Mining (D3M) approach for monitoring staff appraisals in virtual organizations by utilizing 360 Degree appraisals' data mining for objective measurement and opinion mining for subjective measurement. The combined results of the two measurements are sent to support vector machine classifier for classification of employees. The monitoring process from their

perspective is accomplished by listening properly to opinions of people who talks in chat rooms, newspapers, social networks, etc. about virtual organizations and its business, every day positively or negatively.

3.4 Subjectivity Detection:

According to subjectivity detection in opinion mining, the related works are as follows:

In [46] ***wang et. al.*** proposed a framework to handle different types of lexical clues for subjectivity such as opinion indicator (e.g.: accuse, claim), polar word (e.g.: beautiful, ugly), named entity or pronoun (e.g.: China, he), opinion object (e.g.: price, appearance), adverb of degree (e.g.: very, more). They first employed the chi-square technique to automatically extract subjective clues from training data. To represent sentence subjectivity, they calculated sentiment density using the extracted subjective clues and thus constructed a set of sentiment density subintervals. Finally, they implemented a Naive Bayesian classifier with sentiment density subintervals as features for subjectivity classification.

In [47] ***Tang et. al.,*** discussed some approaches used to automatically assign one document as objective or subjective such as similarity approach, where information retrieval method is used to acquire the documents that are on the same topic as the sentence in question. Then, calculate its similarity scores with each sentence in those documents and make an average value. If the average of similarity scores of opinionated documents is higher than that of factual document, then the sentence is classified as a subjective sentence else it is objective.

In [48] ***Lu et. al.*** presents a new approach for subjectivity classification. The approach combines sentiment lexicon and machine learning techniques for opinion mining. They exploited three kinds of lexicon clues: the reporting verbs, polar items and adverb clues. The used sentiment lexicons are exploited to detect opinionated sentences, by checking whether the subjectivity clues exists in the sentences, if so, the sentence is classified as subjective (opinionated). They used a tuning algorithm to remove items with low precision computed on the training corpus. And with the machine learning, they constructed a vector for each sentence with unigram as features and their frequencies, and fed them into the algorithm for learning. They compared three learning algorithms: Naïve Bayes classification, maximum entropy classification, and Support Vector Machines (SVM). They concluded that the combination of SVM and lexicon based method

outperforms the baselines and all individual classifiers, achieving the best performance in terms of accuracy and F-measure.

In [49] ***Pang et. al.*** used subjectivity detection to improve polarity classification in movie reviews. Therefore, they employed a subjectivity detector that determines whether each sentence is subjective or not and discard the objective ones. Thus, creates an extract that represents reviews' subjective contents for polarity classification. In this way, they prevent the polarity classifier from considering irrelevant or even potentially misleading text. In addition, they used subjectivity extracts to provide users with a summary of sentiment-oriented content of the document. In their work, they used supervised learning for the subjectivity detector. According to their experiments, the use of subjectivity extracts provided satisfying improvement in polarity as they obtained an improvement from 82.8% to 86.4% for polarity classification by applying a subjectivity classifier in advance.

In [50] ***Riloff et. al.*** used a high-precision subjectivity classifiers (HP-Subj) to automatically identify subjective and objective sentences in unannotated texts. The HP-Subj uses lists of lexical items that have been shown in previous works to be good subjectivity clues. This process allowed them to generate a large set of labeled sentences automatically. The high-precision classifiers label a sentence as subjective or objective when they are confident about the classification, and they leave a sentence unlabeled otherwise. Then they used the (automatically) labeled sentences as training data for applying an extraction pattern learning algorithm to automatically generate patterns representing subjective expressions. The learned patterns can be used to automatically identify more subjective sentences, which grows the training set, and the entire process can then be bootstrapped. Their experimental results showed that this bootstrapping process increases the recall of the high precision subjective sentence classifier with little loss in precision. They also find that the learned extraction patterns capture subtle connotations that are more expressive than the individual words by themselves.

In [51]***Yu et. al.***, proposed an opinion question answering system that separates opinions from facts at both the document and sentence level and determines if the opinions are positive or negative. To separate documents that contain opinions from documents that report facts, they applied Naive Bayes. Instead of manually labeling the data, they used articles from Wall Street Journal that contain metadata, which helped them in labeling these articles automatically. At the sentence level, to avoid the need for obtaining individual sentence annotations for training and

evaluation, they relied instead on the expectation that documents classified as opinion on the whole will tend to have mostly opinion sentences, and conversely documents placed in the factual category will tend to have mostly factual sentences. By exploring the hypothesis that within a given topic, opinion sentences will be more similar to other opinion sentences than to factual sentences, they applied similarity measurements to measure the overall similarity of a sentence to the opinion or fact documents. They first select the documents that are on the same topic as the sentence in question. Then they average its similarities with each sentence in those documents. Then they assign the sentence to the category for which the average is higher. Another approach they used was Naive Bayes classifier, where they trained the classifier using the sentences in opinion and fact documents as the examples of the two categories.

In [52] ***Biyani et. al.,*** proposed a method to identify the type of information a forum thread contains i.e. whether it is subjective or factual. The method is intended to enhance search engines by considering what type of information a searcher wants; if he wants a factual answer or opinionated one. They modeled the task as a binary classification of threads in one of the two classes: Subjective and Non-subjective. They used combinations of words and their parts-of-speech tags as features. The features were generated from different structural units of a thread such as title, initial post, reply posts and their combinations. For feature representation, they used term frequency as the weighting scheme as they empirically found it to be more effective than tf-idf and binary representations. As a classifier, they used a Multinomial Naive Bayes classifier because it performs well on word features.

3.5 Summary:

In this chapter, we presented the widest related field to ours; which is enhancing HRMS by using data mining and text mining. We also, discussed other works that used opinion mining for different applications including appraisals extractions and monitoring staff performance appraisals. Finally, we discussed works that resolved subjectivity detection problem in a variety of domains.

As a conclusion of these works, we found that there is a considerable number of published papers concerning using data mining and text mining for HRM and staff appraisal. Less number of publications by using opinion mining. However, to the best of our knowledge no published paper concerning detecting subjectivity in staff appraisals by using text mining (our topic).

Works that addressed the problem of subjectivity detection are of two types; one that considered it as a standalone problem, and the other that considered it as an initial task in sentiment classification; that is before detecting polarity of a review detect if it is subjective or not. According to techniques used in subjectivity detection, both supervised and unsupervised learning were used. In supervised learning, classifiers such as SVM, KNN, NB were explored. In unsupervised learning, lexicon based methods as well as, similarity measurements methods are used. In our work, we combined supervised and unsupervised learning methods for detecting three levels of subjectivity.

Chapter Four

Proposed Approach

This chapter presents the steps of our approach. It defines the process of staff appraising. Then illustrates in details the core part of the process; appraisal analyzer, which is responsible for detecting subjective appraisals at three levels of subjectivity.

The methodology we followed in this work is as follows:

4.1 Understand the business domain:

As any application of data mining in any domain, understanding the business domain is the corner stone of the application. Blind application of data mining methods can be a dangerous activity easily leading to discovery of meaningless patterns [19].

We collected our data from the General Personnel Council of the Palestinian government, which is the organization that is responsible for setting and monitoring policies for staff performance appraisals for all ministries and organizations of the Palestinian government. Therefore, we worked together with experts in the field, working for this organization.

The experts we used to consult during all the phases of the project are:

- Eng. Iyad Abu Safia (Director of HR policies development department in the General Personnel Council, Palestinian government)
- Eng. Osama Younis (Director of IT department in the General Personnel Council, Palestinian government).
- Eng. Osama Qassem (Assistant Deputy Minister and member of monitoring appraisals committee in Palestinian government).

  Experts explained to us the ideation about how the appraising process is accomplished.

4.2 Define the Appraising process:

The appraising process consists of the following parts and could be summarized by figure 4.1:

- **Appraisal Receiver:** once reviewers fill the appraisal forms for their staff and submit these forms, the appraisals are received and forwarded to the appraisal analyzer.
- **Appraisal analyzer:** where the appraisals are analyzed and understood, and clues of subjectivity are specified.
- **Dispatcher:** where the appraisals are dispatched to monitoring staff, labeled as subjective or not, so that he could change the mark of these appraisals, write his comments and

approve the appraisals and then send them to monitoring manager so that he could approve them.

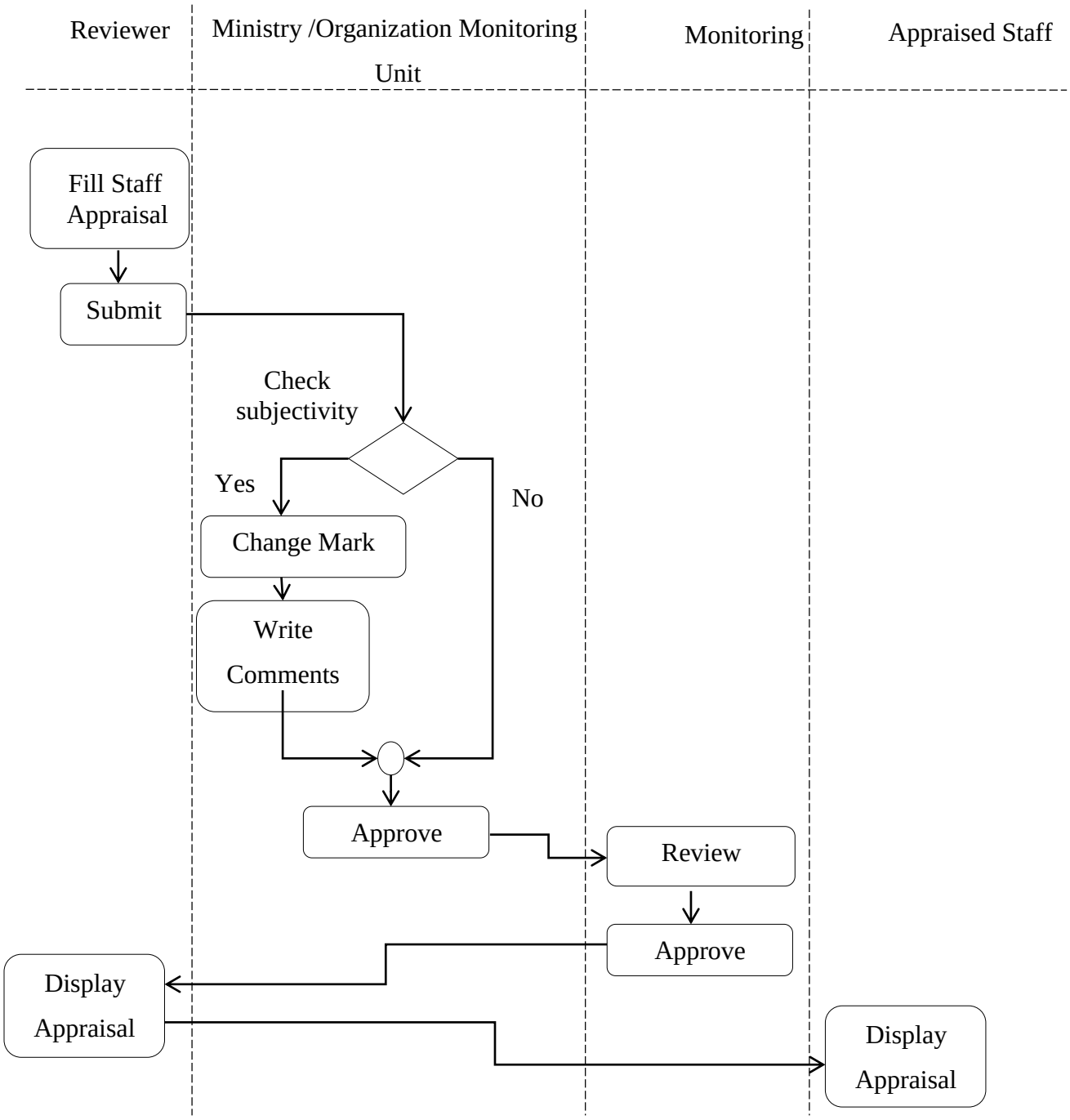

Figure 4.1 staff appraising process according to domain experts' opinion

The Core part of our work, as illustrated in figure 4.2, is to propose an approach for the appraisal analyzer that we could detect subjective appraisals according to it.

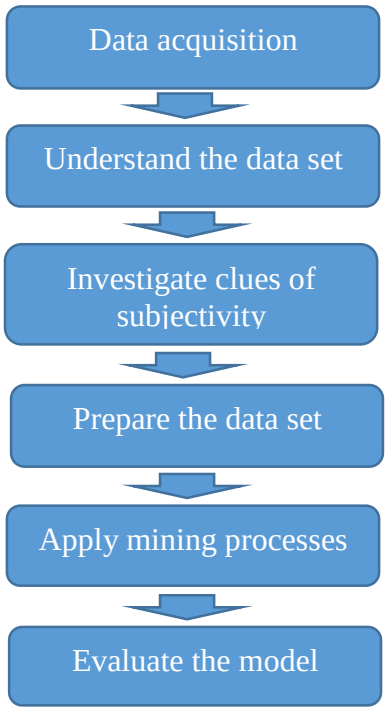

Figure 4.2: Appraisal analyzer steps

Appraisal Analyzer Approach:

The steps of the approach for appraisal analyzer are:

- **Data acquisition:** where we chose the most suitable data set for validating our approach.
- **Understand the data set:** with the help of expert staff, we could understand the procedures of appraising and monitoring.
- **Investigate clues of subjectivity:** from our understanding of the domain and data set, also, with the help of domain experts, we could identify clues of subjectivity.
- **Prepare the data set:** this includes:
  - **Preprocess the text:** where we apply a number of preprocessing techniques such as tokenizing, filtering stop words and stemming.
  - **Labeling the subjectivity of the appraisals:** where we manually labeled reviews as subjective or objective for training and evaluation purposes.

- **Apply mining processes**: such as feature extraction, similarity measurement and classification for building the models for detecting subjective appraisals.
- **Evaluate the model:** we used some measurements such as precision, recall, f-measurement and accuracy as well as, judgment of domain experts.

4.3 Data acquisition:

We chose and specified the more suitable data set for evaluating the approach, which is the appraisals of teachers in the Palestinian government. It is more suitable to choose this data because it is the largest data set we could get. We took the appraisals of the years 2012 and 2013 only for teachers with marks greater than 85%, which consists of 4400 records. We worked on questions from the additional form. As an example, we took the question: "what are the key accomplishments of the employee that made him exceeds performance rates?"

4.4 Data set understanding:

With the help of domain expert staff, we understood the data. As we used the question, "what are the key accomplishments of the employee that made him exceeds performance rates?", They explained to us what is the ideal answer to this question, and what does reviewers actually answer. The ideal answer is a clear and concrete accomplishment. However, actually a number of reviewers answers this question by describing the general behavior of the employee rather than his accomplishments. In addition, some reviewers are giving fake reviews that are irrelevant to the domain of teaching appraisals. Moreover, some reviewers are duplicating the same review to many employees.

4.5 Identify clues of subjectivity:

From our understanding of the data set, we came up with these clues of subjectivity, where each clue represents a level of subjectivity:

- **Irrelevance:** this clue represents the lowest level of subjectivity, where reviewers' answers are irrelevant to the domain of teachers' appraisals, for example, they make fake reviews such as copying the question or writing words that are irrelevant to the domain or duplicating the same words of their review.

- **Duplication:** is another level of subjectivity, where reviewer is duplicating or near duplicating the same answer to different employees.
- **Insignificance:** a higher level of subjectivity where reviewers' answers are meaningless to the question, as we mentioned earlier, they mentions general behaviors and opinions rather than concrete accomplishments.

4.6 Text preprocessing:

In this step, we prepared the data for applying mining methods. This step includes many sub-steps, it starts from tokenizing string into words, then removing stop words, stemming and weighting the terms. Also in this process, we label the data into classes subjective or objective.

- **Tokenizing:**

  Tokenizing is the process of breaking a stream of text up into phrases, words, symbols, or other meaningful elements called tokens [53].

- **Filtering stop words:**

  Where words with little or no content information such as prepositions and conjunctions are removed as these words unlikely help text mining [53].

- **Stemming:**

  This method is used to find out the basic form of a word. For example, the words use, using, user and uses all can be stemmed to the word "USE". The main objective of stemming is to have all the words represented by their stems, by removing all the affixes of the words. The benefit of this process is the reduced number of words and thus the saved memory space and time [54].

  There is two types of stemming:

  - **Root stemming,** where the word is reduced to its origin or root [55].
  - **Light stemming,** where the commonly used affixes in the language are removed without reducing the word to its root [55].

  In our work, we conducted experiments to compare between the two types in each level.

- **Vector Space Model:**

  Despite of its simple data structure without using any explicit semantic information, the vector space model enables very efficient analysis of huge document collections [56]. In vector space model, each document is represented as a vector of the words of

the documents (tokens), in m-dimensional space, where m is the number of words or tokens and the value of each element in the vector is represented by one if the corresponding word exists in the document, and zero if it does not exist [24].

- **Term weight of text documents:**

  To improve performance usually term weighting schemes are used, where the weights reflect the importance of a word in a specific document of the considered collection [57].

  A simple weight scheme is term frequency, where it counts the frequency of the word in the document. The problem of this scheme is that if the word has a large number of frequency in the document and in the whole document collection, then the high weight for this word is with little meaning. Term Frequency Inverse Document Frequency (TF-IDF) is a common weight scheme that is more meaningful, where large weights are assigned to terms that are used frequently in relevant documents but rarely in the whole document collection [57].

  In our approach, we used term frequency in the first level of subjectivity, for the objective wordlist extraction, because we need to extract words that are frequently used in the domain such as:

  "مدرس، معلم، فصل دراسي، منهاج"

  We are not interested in distinguished words; words that are used frequently in a document but rarely in other documents. While in the third level, we used the TF-IDF, because in classification we are interested in distinguished words; frequently used words would not optimize the process.

- **Label the data:**

  Monitoring staff had not ever classified reviews into subjective or objective before, so we need to classify the reviews manually in order to feed them into classifiers, so that classifiers could learn how to classify the new data. This process is essential for the third level of subjectivity, where we classify the answers into either have a significant meaning or not. Also in the first level, we need to classify the data manually for evaluation purposes.

4.7 Apply mining processes:

This is the core step of the process of subjectivity detection; we used feature extraction process for the first level of subjectivity detection, also the process of similarity measurement for the second level, and for the third level, we used processes from machine learning (classification).

4.7.1 Feature extraction for generating objective wordlist:

For the first level of subjectivity, we extracted an objective wordlist from the corpus; that is frequently used words and phrases; unigrams and bigrams. The reason that this works is that when people comment on different aspects of an entity, the vocabulary that they use usually converges [13]. Thus, these frequent words and phrases could be the domain relevant wordlist. This wordlist would be used in the next step for subjectivity detection.

4.7.2 Similarity Measurement:

In this measurement, the features or tokens of documents are represented as vectors in the space, as illustrated in figure 4.3. Typically, the angle between two vectors is used as a measure of divergence between the vectors, and cosine of the angle is used as the numeric similarity, since cosine has the nice property that it is 1.0 for identical vectors and 0.0 for orthogonal vectors [58]. By finding the dot product of the two documents (or reviews in our case), we could find the cosine similarity between the two documents.

Equation 4.1 finds the similarity between two documents x and y.

$$similarity(x,y) = cos(\theta) = \frac{x \cdot y}{||x|| * ||y||} \quad \text{.....4.1}$$

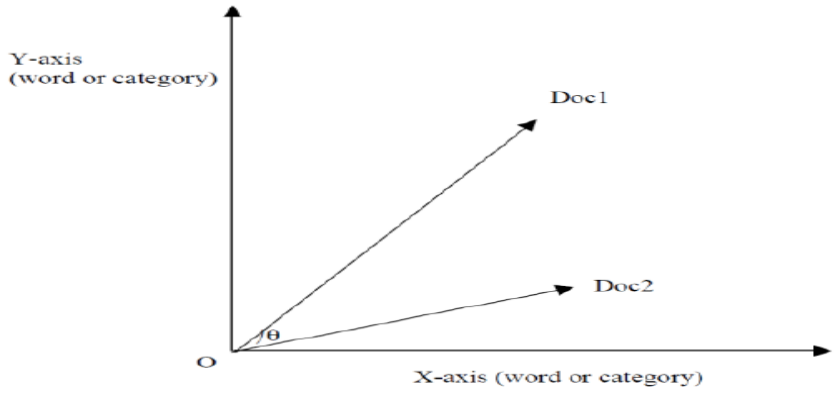

Figure 4.3: representation of two documents in 2-D space

We used similarity measurement for the second level of subjectivity detection, where we detect whether a reviewer is duplicating or near duplicating the same review for his employees. This measurement gives us a percent of the similarity among all reviews.

4.7.3 Machine Learning processes (classification):

For the third level of subjectivity, we used classification methods. Almost all the known techniques for classification such as decision trees, rules, Bayes methods, nearest neighbor classifiers, SVM classifiers, and neural networks have been extended to the case of text data [59]. Recently, a considerable amount of emphasis has been placed on linear classifiers such as neural networks and SVM classifiers, with the latter being particularly suited to the characteristics of text data [59]. In our work, we used Support vector Machine (SVM) and compared its results with two other algorithms: Naïve Bayes (NB) and K Nearest Neighbor (KNN), as these two algorithms gave results close to that of SVM in some other researches in Arabic language [20][24].

We fed these algorithms with the training data, so that these algorithms would build the model that new data could be classified according to it. After that, we compared the results of the three algorithms, so that we could decide which algorithm is the best for our approach.

4.7.3.1 Support Vector Machine (SVM):

Support Vector Machine is a supervised machine learning technique motivated by the statistical learning theory. Based on the structural risk minimization of the statistical learning theory, SVM seeks an optimal separating hyper-plane to divide the training examples into two classes and make decisions based on support vectors, which are selected as the only effective instances in the training set [60]. Intuitively, a good separation is achieved by the hyper-plane 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 [61].

4.7.3.2 Naïve Bayse (NB):

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence

assumptions. A more descriptive term for the underlying probability model would be 'independent feature model'. In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class (i.e. attribute) is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4 inches in diameter. Even if these features depend on each other or upon the existence of the other features, a Naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple [62].

4.7.3.3 K-Nearest Neighbor (KNN):

The K-Nearest Neighbor algorithm is based on learning by analogy, that is, by comparing a given test example with training examples that are similar to it. The training examples are described by n attributes. Each example represents a point in an n-dimensional space. In this way, all of the training examples are stored in an n-dimensional pattern space. When given an unknown example, a k-nearest neighbor algorithm searches the pattern space for the k training examples that are closest to the unknown example. These k training examples are the k "nearest neighbors" of the unknown example [63].

4.7.4 Subjectivity Detection

This step means to detect subjective appraisals in each level.

- **In the first level of subjectivity:**

We check if the answers are relevant to the domain or not. In order to perform this check, we used the objective wordlist that we generated in section 4.7.1. We calculated the number of words in the review that also exists in the wordlist, and divided this number by the number of overall words of the review.

$$\text{Relevance Percent} = \frac{|\{\text{words from review}\} \cap \{\text{words from wordlist}\}|}{|\{\text{Words from review}\}|} \quad \textbf{.....4.2}$$

This percent represents the percent of relevance. Considering relevance as a percent would enable us not only to detect irrelevant reviews, but also to detect duplication in the same review; i.e. if a reviewer is writing a sentence and duplicating it. By experiments, we would decide at which percent we could classify the review as a subjective review.

- **In the Second level of subjectivity:**

We check if the reviewer is duplicating the same reviews to his employees. In this check, we analyzed the reviews of each reviewer in order to see the similarity percent among his reviews. Sometimes reviewers are fully duplicating their reviews to their employees, and sometimes they nearly duplicating the reviews; i.e. changing some words such as the levels or the subjects they teach; for example:

"المعلمة تدرس مادة اللغة العربية للصف السادس الاساسي"

"المعلمة تدرس مادة التربية الفنية للصف الأول الثانوي"

By experiments, as well as, with the help of domain expert staff we would decide at which similarity percent we could classify the review as a subjective one.

- **In the third level of subjectivity:**

We check if the review contains a significant meaning or not. Significant meaning reviews means more than saying, "The teacher is active/good/energetic". The answer should describe a clear accomplishment. Moreover, reviews that nearly duplicate items from the basic form of appraisals are detected such as:

"المعلمة نشيطة و ملتزمة بالأنظمة و القوانين و تمثل قدوة حسنة".

This level of subjectivity detection complies with the definition of subjectivity detection of opinion mining that classifies a review as subjective if it is an opinionated review and objective if it is a factual review. The reason is that the ideal answer of this question should not be an opinion, it should be a description of clear works and accomplishments of the employee; seems to be factual information.

In order to perform this check, we check new data (testing data that was not included in the training process) against the model built in section 4.7.3.

4.8 Evaluation:

We evaluated our work by using domain expert judgment, as well as, measurements of accuracy, precision, recall and F-measure, In addition, we used the 10-fold cross validation technique.

4.8.1 Domain expert judgment:

Incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, ensure that useful knowledge is derived from the data [19]. Intuitively, domain experts are the best ones to provide knowledge, interpretation and judgments for the results. Before we started the experiments, we consulted the domain experts, mentioned in 4.1, about the approach. They encouraged it, and thought it would help very much in monitoring staff appraisals. In addition, during our experiments, they gave us the instructions for labeling the data set into subjective or objective, so that, the measurements of accuracy in the first and third levels would be based on their opinion. Also, in the second level of subjectivity detection, they helped us in specifying the threshold for duplication.

4.8.2 Accuracy measurements:

The measurements we used are:

- **Accuracy:** the accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier [64].

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \quad \text{.....4.3}$$

Where tp is true positive instances, tn is the true negative, fp is the false negative, and fn is the false negative.

But the problem with accuracy is that there are many labeled data sets which have an unbalanced representation among the classes in them, when the imbalance is large, classification accuracy on the smaller class(es) tends to be lower [65]. This problem could be solved by using measurements of precision, recall and f-measure.

- **Precision:** is the fraction of retrieved objects that are relevant [66].

$$\text{Precision} = \frac{tp}{tp + fp} \quad \text{.....4.4}$$

Where tp is the true positive instances, fp is the false positive instances.

For example if our classifier predicts 20 subjective reviews, if only 15 of these reviews are truly classified as subjective and 5 are falsely classified, then the precision for the classifier for the subjective class equals: 15/ (15+5) = 15/20 = 0.75

- **Recall:** is defined as the proportion of relevant objects that are retrieved relative to the total number of relevant objects in the data set [67].

$$\text{Recall} = \frac{tp}{tp + fn} \quad \text{.....4.5}$$

Where tp is the true positive instances and fn is the false negative instances.

For the same example, we mentioned for precision, if there is another 10 subjective reviews that the classifier failed to predict, then the recall equals: 15/ (15+10) = 15/25 = 0.60

- **F-measure:** We could notice that, precision of class C does not tell us anything about the number of class C tuples that the classifier mislabeled. Also, recall of a class C does not tell us how many other tuples were incorrectly labeled as belonging to class C. There tends to be an inverse relationship between precision and recall, where it is possible to increase one at the cost of reducing the other [64].

  Precision and recall scores are typically used together, where precision values are compared for a fixed value of recall, or vice versa. An alternative way to use precision and recall is to combine them into a single measure. This is the approach of the F measure (also known as the F1 score or F-score) [64].

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad \text{.....4.6}$$

4.8.3 10-Fold cross validation:

Using training data to derive a classifier and then estimate the accuracy of the resulting learned model, can result in misleading overoptimistic estimates due to overspecialization

of the learning algorithm to the data. Instead, it is better to measure the classifier's accuracy on a test set consisting of class-labeled tuples that were not used to train the model [64] .

The most basic testing method is called simple validation. To carry this out, we set aside a percentage of the database as a test database, and do not use it in any way in the model building and estimation. This percentage is typically between 5% and 33% [67]. An advanced method is the n-fold cross validation, where the data is randomly split into n mutually exclusive subsets of approximately equal size. An inducer is trained and tested several times. Each time it is tested on one of the n folds and trained using the remaining n−1 folds [67]. In general, stratified 10-fold cross validation is recommended for estimating accuracy (even if computation power allows using more folds) due to its relatively low bias and variance [64].

In our work, we used 10-fold cross validation in the third level of subjectivity detection to evaluate classifiers; we calculated the precision, recall, F-measure and accuracy for each fold, then we took the average of these measurements.

4.9 Summary:

In this chapter we discussed our approach in subjectivity detection, at first we elaborate the overall process of staff appraising, then in detail we described the core part of the process of appraisals analyzer, which detect three levels of subjectivity. We elaborated the details of every level and what methods and algorithms we would use, and then we discussed how we would evaluate our approach.

Chapter Five

Experiments and Results

In this chapter, we describe the conducted experiments to evaluate our approach. We made three sets of experiments; each one detects a level of subjectivity. It also describes the comparison between the methods according to the results, to achieve the best performance. For the evaluation, we used domain experts' judgment, accuracy, precision, recall and F-measure.

5.1 Experiments settings:

The experimental environment used for all experiments was CPU / Intel Pentium i5 processor, Memory of 4 GB RAM, Windows 7. In addition, we used as a software: Rapid miner 5.3 for the mining processes, MS Excel 2013 and Oracle database for analyzing and presenting the results.

5.2 Rapidminer:

RapidMiner is a software platform that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. In 2014, Gartner Research placed RapidMiner in the leader quadrant of its Magic Quadrant for Advanced Analytics. The report described RapidMiner's strengths as a "platform that supports an extensive breadth and depth of functionality, and with that it comes quite close to the market Leaders."[68]

5.3 The Data Set:

For our experiments, we used real data. We used the teachers' appraisals of Palestinian Government for the years 2012 and 2013 for teachers with marks more than 85%, which consist of 4400 records. We applied our experiments on the answers of questions from the additional form. We took as an example, the question: "What are the key accomplishments of the employee that made him exceed performance rates?"

5.4 First level of subjectivity:

This set of experiments aim to detect the lowest level of subjectivity by determining if the textual answers are in the domain or not; which is in our case education and teacher's appraisals. For this purpose, we extracted an objective wordlist, consisting of words from the domain. Our assumption is that if the review contains a threshold of these words, then it could pass this level of subjectivity detection and considered relevant to the domain.

5.4.1 Data preprocessing:

In this step, we prepared our data set for applying mining methods using Rapidminer tool. The process we followed for preprocessing, as illustrated in figure 5.1, consists of tokenizing, where streams of texts are broken into tokens. Then these tokens are passed to a filtering stop words process, where words with little content information are removed such as prepositions and conjunctions. After that, we filtered the resulted tokens by using the process of filter tokens, where we specified tokens less than 3 characters to be removed. Then, we passed the resulted tokens to a stemming process. For stemming, we had two choices; light stemming which removes only the common affixes in Arabic language, and root stemming which returns the word to its root. We compared between the two types according to the results in order to decide which is better to use in this level. After stemming, we used filtering tokens to remove the resulted stems that is less than 3 characters.

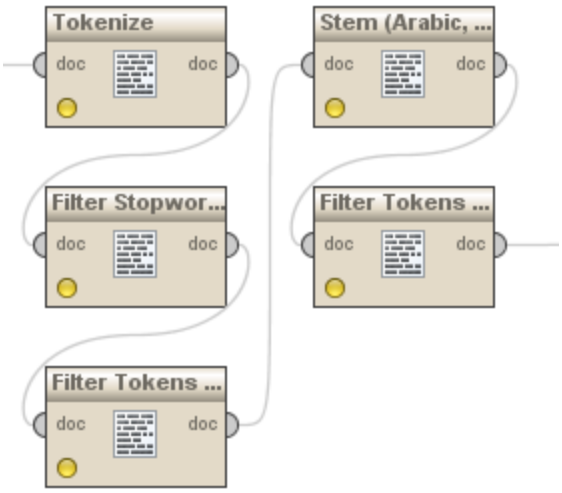

Figure 5.1: Preprocessing steps on Rapidminer

We examined the results of the preprocessing steps on the whole data set (4400 records). The results are illustrated in table 5.1.

Table 5.1: Number of tokens after each step of preprocessing

Process	**Number of resulted tokens after process**	
Tokenization	13,127	
Stop word filtering	12,911	
filtering tokens	12,816	
Processes after using stemming	**Number of tokens after process (using root stemming)**	**Number of tokens after process (using light stemming)**
Stemming	3141	6,495
Filtering tokens	3,078	6,317

We could notice form the results that the processes stop word filtering and filtering tokens, slightly reduced the number of tokens, while stemming reduced the number of tokens to 6495 (around 50% of the tokens) when using light stemming and 3141 (around 25% of the tokens) when using root stemming. This complies with previous experiments on Arabic text [69][70], which showed that stemming greatly reduces vector sizes (features) more than light stemming. Therefore, we decided to use root stemming in the first level, because it would reduce time and effort for manually checking the wordlist.

5.4.2 Feature extraction for generating objective wordlist:

In this step, we extracted a domain relevant wordlist from the corpus; that is frequently used words and phrases (unigram and bigram).

We used generate n-gram process with a parameter of 2 for n. The input of this process is the tokens resulted from preprocessing step in section 5.4.1. And the results are a list of unigram and bigram, where unigrams are represented by tokens consisting of one word, and bigrams are represented by every two contiguous words of the text.

N-grams process as well as, the preprocessing processes represent sub-processes of one larger process that takes the data set and generates word vectors from string attributes,

reviews in our case, this process is called process documents from data, figure 5.2 illustrates the process. In this process, we used term occurrences as a term weighting schema for the represented vectors, because in this level, we need the commonly used words in the domain of teachers' appraising rather than the distinguished words.

For extracting the final wordlist, we removed words with term occurrences less than 20, and then manually we chose the most relevant words.

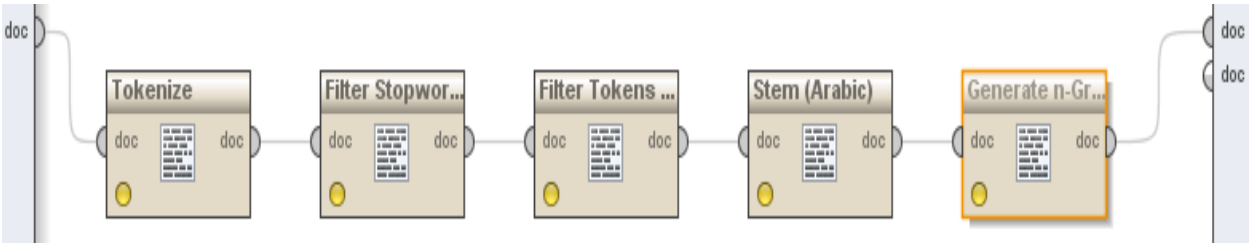

Figure 5.2: Sub processes of Process document from data process

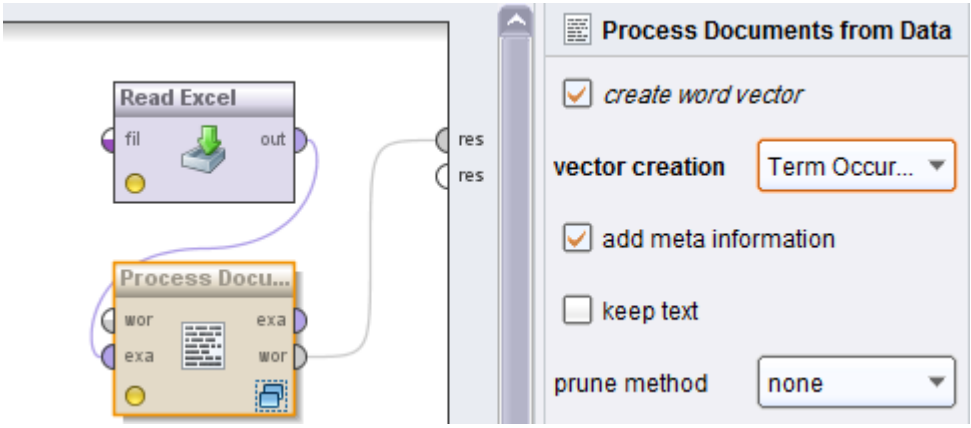

Figure 5.3: Processing Documents Process

In this experiment, we split the dataset into data for generating the wordlist and data for testing. We took into account that, the larger the training data, the better the classifier [71], in our case the larger the data set for generating the wordlist, the better the wordlist, and the larger the test data, the more accurate the error estimate [71]. Therefore, we spilt the dataset three times, as illustrated in table 5.2, in order to decide the best number of records to use for wordlist extraction and for testing.

Table 5.2: Splits of data into input data and testing data

Split Number	Number of records for wordlist extraction	Number of records for testing
1	4400	0
2	4000	400
3	2900	1500

First time, we took all the corpus (4400 records of data). We preformed the preprocessing and feature extraction processes for extracting the objective wordlist, then removed tokens with occurrences less than 20, and finally manually chose the most relevant wordlist. We came up with a list of 206 tokens (the list is illustrated in appendix A).

Second time, we reduced the data for wordlist extraction to 4000 records, so that we could have testing data; that was not used in the wordlist extraction process. As we did in the first time, we preprocessed the dataset, extracted unigrams and bigrams. However, before completing the process of wordlist extraction, which is removing tokens with occurrences less than 20 and manually selecting the relevant tokens, we compared the tokens with the extracted wordlist from the first time; wordlist from all the data i.e. 4400 records. We found that the extracted wordlist is part of the tokens. Therefore, proceeding in the process would lead us to the same wordlist.

Another time, we reduced the size of the data for wordlist extraction to 2900 records, which represents 2/3 of the overall data. We tokenized the data and generate unigrams and bigrams, then checked if the wordlist, extracted at the first time, is part of the tokens. Also at this time, we found that the extracted wordlist is part of the tokens.

The interpretation of these results is, as we mentioned before, that when people comment on different aspects of an entity, the vocabulary that they use usually converges. Moreover, this is well supported because of the large size of data we used.

These results led us to take the choice of using the whole data set as input data, i.e. data for extracting the relevant wordlist, as reducing the input data will always lead to the same wordlist.

5.4.3 Subjectivity Detection

In this step, we checked how much does the answer contains words from the objective wordlist. We calculated the relevance percent, which is the number of words in the review that also exists in the relevance wordlist, divided by the length of the review.

$$\text{Relevance Percent} = \frac{|\{ \text{words from review} \} \cap \{ \text{words from wordlist} \}|}{|\{ \text{Words from review} \}|} \quad \text{.....5.1}$$

Figure 5.4 illustrates examples of real reviews with their relevance percent. We could decide if the answer is relevant or not by checking whether the relevance percent is greater than a threshold. In section 5.4.4, we decide the value of the threshold.

Relevance	**0%**
Review	kl kj ui u u ui i i i bi i iu u i iuh oij oih i kub u hj iu kj kiu ib u iu u i iu i i u uvft tfv po ou uyg
Relevance	**0%**
Review	ينتينبتي بنيت بنيت بني تبنيت بشمكثهب نين يته بثتبنيتب هيت هبت يهبت بهيت هيت
Relevance	**13%**
Review	أذكر الأعمال البارزة للموظف التي أدت إلى تحقيق الأهداف وتجاوز معدلات الأداء. أذكر الأعمال البارزة للموظف التي أدت إلى تحقيق الأهداف وتجاوز معدلات الأداء.
Relevance	**17%**
Review	المساهمة في توزيع البسكويت علي الطالبات
Relevance	**18%**
Review	الاهتمام بالجانب العملي- الالمام والمعرفة بالمادة العلمية- المبادرة- النشاط والالتزامالاهتمام بالجانب العملي- الالمام والمعرفة بالمادة العلمية- المبادرة- النشاط والالتزامالاهتمام بالجانب العملي- الالمام والمعرفة بالمادة العلمية- المبادرة- النشاط والالتزامالاهتمام بالجانب العملي- الالمام والمعرفة بالمادة العلمية- المبادرة- النشاط والالتزامالاهتمام بالجانب العملي- الالمام والمعرفة بالمادة العلمية- المبادرة- النشاط والالتزام-المشاركة في لجنة التخصصات
Relevance	**19%**

Review	بالرغم من كبر السن الا انها تسلم اعمالها قبل الوقت المحدد لها - لم تأخذ اجازات خلال العام الا ما ندر و عند الضرورة القصوى - تشارك في المعرض الفني الختامي باعمال مميزة - تحرص على التعلم و تطوير الذات من خلال المشاركة في دورة بوربوينت على الرغم من بعد المسافة عن مكان الدورة
Relevance	**20%**
Review	تقوم المعلمة بحث الطالبات على معرفة التراث الفلسطيني والمحافظة على صوره في الملبوسات والأثاث من خلال حث الطالبات على المشاركة بعمل المطرزات الجميلة التي تعكس بعض صور التراث والهوية الفلسطينية وتغرس في نفوسهن حب العمل والدقة فيه والحرص على سلامة الأداء

Figure 5.4: examples of real reviews with their relevance percent

5.4.4 Evaluation and results:

We used the whole data set for word list extraction because we came up with the same one after splitting it many times. Also in testing, we decided to start by using the whole data set, and check if the results are satisfying or not. If it were satisfying, we would stop here; else, we would split the data set into input data i.e. data for word list extraction and data for testing, so that we would modify the wordlist and then perform the test on the testing data.

We tried the values of 10, 15, 18, 20, 25 as the threshold of relevance percent, and calculated the measurements of precision, recall, and F-measure (results are illustrated in table 5.3)

As we could see from table 5.3, the F-measure for the subjective reviews, with the threshold of 10 was 0.8, then at the threshold of 15, it increased to 0.82, and it increased to 0.88 when the threshold increased to 18. However, at the threshold of 20, it decreased to 0.83, also at 25 it decreased to 0.39. These results led us to choose the threshold value of 18.

Since the value of F-measure is high, we would not modify the wordlist and retest the data.

Table 5.3: measurements of accuracy with different values for threshold

Threshold	Subjective Class		
	Precision	Recall	F-measure
10	1	0.67	0.80
15	0.9	0.75	0.82
18	0.85	0.92	0.88
20	0.71	1	0.83
25	0.24	1	0.39

5.5 Second level of subjectivity:

This set of experiments aim to detect if the reviewer is duplicating the same answer to different employees. For this purpose, we used the similarity measurement process of text mining, and evaluated the results with the help of domain experts.

5.5.1 Preprocessing data:

In this step, we followed the same steps of text preprocessing that we followed in the first level in section 5.4.1; that is tokenization, removing stop words, filtering tokens less than 3 characters and stemming. For stemming, we compared the two types of stemming; light stemming and root stemming in order to decide which is better to use in this level.

As a data set, we used the answers of four reviewers; each reviewer appraised more than 20 employees (table 5.4 illustrates number of employees for each reviewer)

Table 5.4: number of employees for each reviewer for the second level of subjectivity detection

Reviewer number	Number of records (employees)
1	21
2	41
3	37
4	26

5.5.2 Similarity Measurement:

We used the similarity measurement process of Rapidminer, illustrated in figure 5.5. We chose the cosine similarity which finds the similarity between each two documents (answers) as described in section 4.7.2.

As we mentioned in section 5.4.1, we need to decide which stemming type is better to use. We made experiments using the two types and compared between them. Experiments showed that similarity percent using root stemming in most cases is greater than similarity percent using light stemming. We investigated the results and concluded that light stemming failed to detect similarity between some long words such as:

"اولوياته، اولويات"

"تستثمر، يستثمر"

Therefore, we decided to use root stemming in this level.

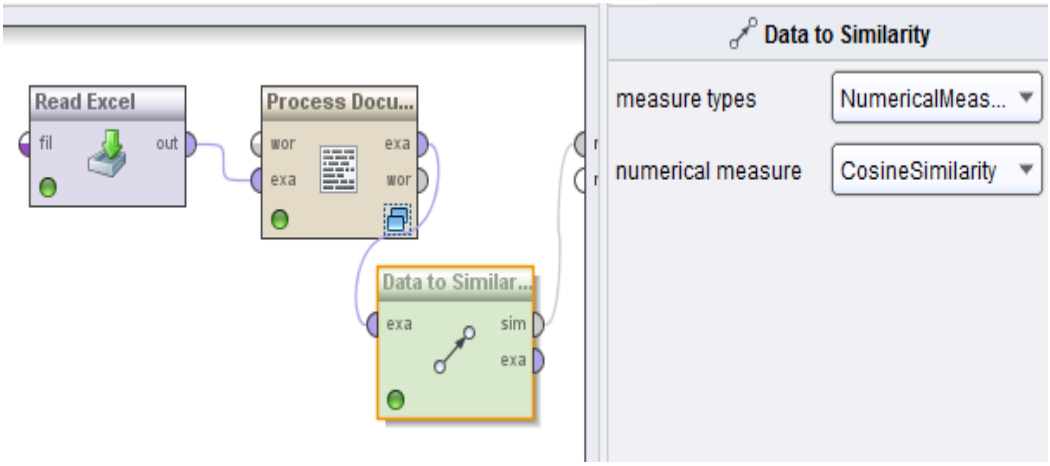

Figure 5.5: Similarity Measurement Process

5.5.3 Subjectivity Detection

Figure 5.6 illustrates examples of real answers with similarities equal to 95%, 90%, 85%, 80%, 75%, 70%, 65%.

Similarity	**95%**
First Answer	متمكنة من المادة العلمية متميزة في الأداء الصفي . تلتزم بمواعيد الحضور والانصراف . تستثمر أوقات الدوام المدرسي في صالح العملية التعليمية . تلتزم بالأنظمة والتعليمات المدرسية.
Second Answer	متمكنة من المادة العلمية تلتزم بمواعيد الحضور والانصراف . تستثمر أوقات الدوام المدرسي في صالح العملية التعليمية . تلتزم بالأنظمة والتعليمات المدرسية.
Similarity	**90%**
First Answer	تستثمر أوقات العمل بطريقة كفؤ - تتعاون بشكل إيجابي مع الطلبة وتكون بمثابة قدوة حسنة لهم . تشارك في الأنشطة المدرسية
Second Answer	يلتزم بالأنظمة والتعليمات – ويستثمر أوقات العمل بطريقة كفؤ - يتعاون بشكل إيجابي مع الطلبة ويكون بمثابة قدوة حسنة لهم .
Similarity	**85%**
First Answer	متمكنة من المادة العلمية متميزة في الأداء الصفي . تلتزم بمواعيد الحضور والانصراف . مخلصة ومتفانية في عملها في تعليم المبحث تستثمر أوقات الدوام المدرسي في صالح العملية التعليمية . تلتزم بالأنظمة والتعليمات المدرسية.
Second Answer	تستثمر أوقات الدوام المدرسي في صالح العملية التعليمية تلتزم بالأنظمة والتعليمات المدرسية تتقبل التوجيهات وتلتزم بها متمكنة من المادة العلمية ومتميزة في الأداء
Similarity	**80%**
First Answer	تلتزم بالأنظمة والتعليمات المدرسية. متميزة في الأداء الصفي . متمكنة من المادة العلمية تستثمر أوقات الدوام المدرسي في صالح العملية التعليمية .

Second Answer	متمكنة من المادة العلمية تلتزم بمواعيد الحضور والانصراف تستثمر أوقات الدوام المدرسي في صالح العملية التعليمية
Similarity	**75%**
First Answer	لديه قدرة متميزة على تقبل التوجيهات - متحمس لتحمل أية مسئولية، يعمل بشكل منظم ويرتب أولوياته بشكل جيد - لديه قدرة عالية للتعامل مع المواقف الطارية
Second Answer	يعمل بشكل منتظم ويرتب أولوياته بشكل جيد – يوزع أولوياته بشكل جيد- - لديه قدرة متميزة على تقبل ا لتوجيهات
Similarity	**70%**
First Answer	متميزة في الأداء الصفي ملتزمة بالانظمة والتعليمات المدرسية تتقبل التوجيهات وتنفذها
Second Answer	متميزة في الأداء الصفي . تلتزم بمواعيد الحضور والانصراف . تستثمر أوقات الدوام المدرسي في صالح العملية التعليمية . تلتزم بالأنظمة والتعليمات المدرسية وتنفذها .
Similarity	**65%**
First Answer	متمكنة من المادة العلمية متميزة في الأداء الصفي .تلتزم بالأنظمة والتعليمات المدرسية. مخلصة ومتفانية في تعليم المبحث
Second Answer	متمكنة من المادة العلمية ومبدعة ومتميزة في الاداء الصفي .تحقق نتائج متميزة في امتحانات الثانوية العامة في مبحث الجغرافيا. تلتزم بألأنشطة والتعليمات المدسية

Figure 5.6: Examples of similar real answers with their similarity percent

5.5.4 Evaluation and results:

The result of this process was the similarity percent for each two answers of the same reviewer. From these results, we came up with around 1500 records representing the

similarity between each two answers for each reviewer. We analyzed these results with the help of Eng. Iyad Abu Safia, in order to find the threshold of similarity that reviews with this similarity or greater, would be considered duplicated. We found that answers with similarities greater than or equal to 85% could be considered as subjective reviews, as these answers contain many duplicates. We noticed that, even if answers with percent less than 85% and greater than 70% contains some duplicates, it contains distinct non duplicated features. This made us consider that at this interval, a reviewer is not duplicating his reviews nor making reviews carelessly. Therefore, we would not consider them as subjective reviews. Eng. Abu Safia's point of view was that in the domain of teachers' appraisals, the opportunity of duplication is high, because all teachers are working in the same domain and the same work. They may share the same behavior and the same accomplishments. Therefore, we chose a high percent for similarity (85%) to indicate subjective reviews in this domain.

5.6 Third level of subjectivity:

This set of experiments aim to detect a higher level of subjectivity by determining whether the textual review is a meaningful answer to the question or not. For example, we took the question:

"**اذكر الاعمال البارزة للموظف التي أدت الى تجاوز معدلات الاداء**"

"What are the key accomplishments of the employee that made him exceed performance rates?". In domain experts' point of view, review is considered to be meaningful if it mentions a clear and concrete accomplishment. For example:

"**- تهتم بهندامها وأناقتها بشكل دائم وتظهر بمظهر أنيق محتشم إسلامي رزين.-تمثل سلوكياتها قدوة صالحة للطالبات فهي خلوقة مؤدبة تتصرف بحكمة واتزان.-تقية تراعي الله في عملها ولا تبالي بمراقبة المسؤلين وانما تعمل بما يرضي الله وما تمليها عليها اخلاقيات المهنة**."

This example is not a meaningful review, because it describes the general behavior of the teacher rather than mentioning a concrete accomplishment.

An example of a meaningful review is:

"**خبرة المعلمه الطويله منحتها عدة ميزات اهمها *عمل خطط علاجيه قابله للتنفيذ والمتابعه *تبنى عدد من الطالبات الضعيفات جدا باللغه العربيه وتعليمهن اساسيات اللغه *دائما تبحث عن تطوير ذاتها *لديها مهارة توزيع المنهاج واستغلال الوقت على مدار السنه الدراسيه**"

5.6.1 Data preprocessing:

In this step, we used the whole data set and we followed the same steps of text preprocessing that we followed in the first and second levels of subjectivity; that is tokenization, removing stop words, filtering tokens less than 3 characters and stemming. We compared also between light stemming and root stemming. In addition, in this step, we labeled the answers to be either meaningful (objective) or meaningless (subjective) based on the instructions of domain experts, we mentioned before, that is; an objective review mentions clear and concrete employee accomplishments.

5.6.2 Machine Learning processes (classification):

We fed the labeled reviews (training data) to machine learning algorithms, so that they could learn how to classify the new coming data. We tried three algorithms for classification: Support Vector Machine (SVM), Naïve Bayes (NB) and K-Nearest Neighbor (KNN)

5.6.2.1 Support Vector Machine (SVM):

We used the SVM algorithm of rapid miner as illustrated in figure 5.7

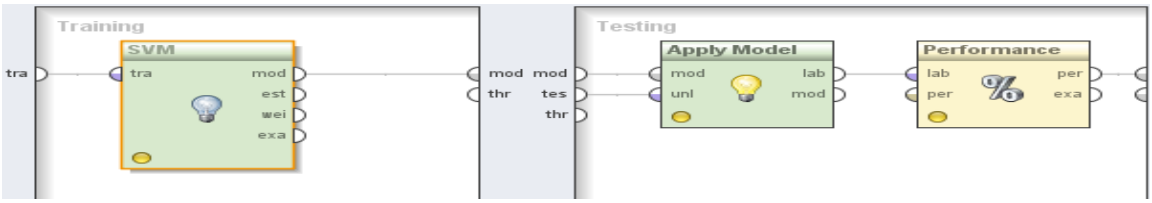

Figure 5.7 Applying SVM for classification

5.6.2.2 Naïve Bayse (NB):

We used the NB algorithm of rapid miner as illustrated in figure 5.8

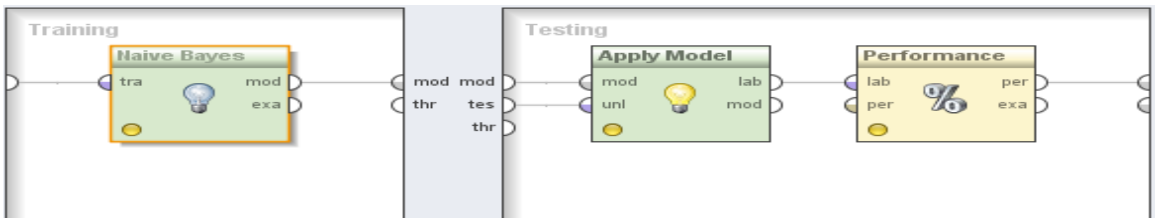

Figure 5.8 Applying NB for classification

5.6.2.3 K-Nearest Neighbor (KNN):

We used the K-NN algorithm of rapid miner as illustrated in figure 5.9

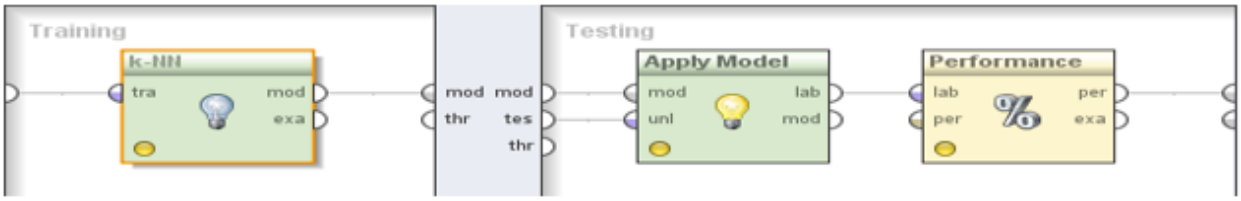

Figure 5.9 Applying K-NN for classification

5.6.3 Subjectivity Detection:

In the previous step, we used the algorithms of classification to build the model for classification. In this step, we used the model to classify the testing data as seen in figure 5.10

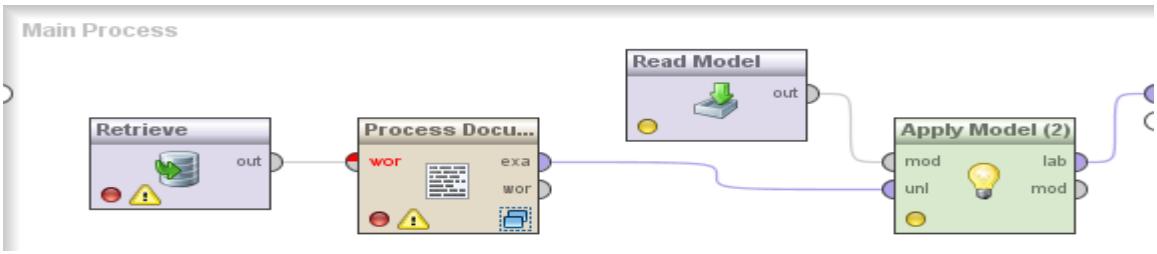

Figure 5.10 applying classification model for data

5.6.4 Evaluation

As the data set is not too large (rule of thumb is 5000), we decided to use the 10 fold cross validation method in splitting the data. So that, we would take the average of the evaluation measurements, and this would be more accurate.

In our comparison between the classifiers, we used the measurements of precision, recall, f-measure and accuracy. In addition, we compared using light stemming versus stemming. Table 5.5 shows the results of the first fold, where we used 4000 records for training and 400 for testing.

Table 5.5: the results of classification for first fold validation

Algorithm	Stemming Type	Subjective reviews measurement			Objective reviews measurement			
		Precision	Recall	f-measure	Precision	Recall	F-measure	Accuracy
SVM	Stemming	0.73	0.94	0.83	0.94	0.74	0.83	0.83
	Light stemming	0.79	0.94	0.86	0.94	0.81	0.87	0.87
KNN	Stemming	0.67	0.85	0.76	0.86	0.71	0.78	0.77
	Light stemming	0.71	0.83	0.77	0.85	0.74	0.80	0.79
NB	Stemming	0.61	0.92	0.73	0.90	0.56	0.69	0.72
	Light stemming	0.72	0.85	0.78	0.88	0.74	0.80	0.80

We noticed from the results, that light stemming is always better than stemming, so we decided to use light stemming in the rest of folds.

After completing the 10 folds, we came up with the average of them, illustrated in table 5.6.

Table 5.6: average of 10 fold classification

Algorithm	Subjective reviews measurement			Objective reviews measurement			
	Precision	Recall	f-measure	Precision	Recall	f-measure	Accuracy
SVM	0.78	0.91	0.84	0.92	0.81	0.86	0.85
KNN	0.75	0.81	0.78	0.84	0.79	0.81	0.80
NB	0.72	0.82	0.76	0.84	0.75	0.72	0.78

As we can see from the table, SVM achieved the highest accuracy (85%) and f-measure (84% for subjective class and 92% for objective class), then KNN with accuracy (80%)

and f-measure (78% for subjective class and 81% for objective class), then the NB with accuracy (78%) and f-measure (76% for subjective class and 72% for objective class).

In addition, we noticed that the precision of the objective class with the SVM classifier is high (92%). This means that the number of false classification of the objective class is small (only 8%).

Therefore, we decided to select SVM algorithm to be used in our domain.

5.7 Summary:

This chapter described experiments results and analysis for detecting the three levels of subjectivity in staff appraisals (irrelevance, duplication and insignificance). We used the text mining methods: feature extraction, similarity measurements and classification. In the first level of subjectivity we used feature extraction to generate objective wordlist, so that we would classify reviews that contains a percent of relevant words less than 18% to be subjective. The f-measure for this level was 88%. In the second level, we used similarity measurements to specify the percent of similarity between each two reviews among reviews of the same reviewer. With the help of domain experts, we decided the percent of similarity that let us consider a review to be subjective, if it contains greater than or equal to this percent. In the third level, we used classification methods to detect reviews with insignificant meaning. We compared three classifiers and found that SVM achieved the highest average accuracy (85%). We found that these methods are efficient in detecting subjectivity in terms of experts' approval for the second level and the high F-measure for the first and third levels (88% and 84% respectively).

Chapter Six

Conclusion and Future Works

6.1 Conclusion:

In this research, we proposed a text mining based approach for detecting subjectivity in staff performance appraisals. We used as a case study to evaluate our approach, teachers' appraisals of the Palestinian government for two years, consisting of 4400 records.

The approach detects subjectivity at three levels; irrelevance to domain, duplicated reviews and insignificant meaning reviews. We used different opinion mining technique for each level. For the first level, we used feature extraction by using unigrams and bigrams in order to generate an objective wordlist. For the second level, we used similarity measurement. And for the third level, we used classification.

According to our experiments, we found that the approach is effective regarding our evaluations; where we used: expert opinion, precision, recall, accuracy and F-measure. In the first level we reached the F-measure of 88%, and in the second level, we used the expert staff opinion, where they decided the percent of duplication to be 85%, and in the third level, we compared between three classifiers (SVM, KNN and NB), our experiments showed that SVM achieved the best average accuracy (85%), and best average F-measure (84%)

6.2 Future Works:

Our work could be developed to detect more clues of subjectivity; for example, by analyzing the reviewers' answers in the textual part of appraisal, understanding what reviewers are talking about, and trying to search for a contradiction with the non-textual part of appraisal. Also, we could work on other domains and a larger data set. We could look for other clues that could help HR in other areas. Also, the work could be developed further to handle other languages.

According to the notes of discussion committee of the thesis, consisting of Dr. Rawia Awadallah and Dr. Ahmad Mahmoud, our methodology could be enhanced as follows:

- The levels of subjectivity could be detected as a pipeline rather than separately; that is only the objective reviews from one level are passed to the next level.

- In order to generalize our work, in the first level of subjectivity detection, the wordlist could be extracted from other general corpuses.
- Splitting the data into training data and testing data with respect to the year of evaluation, would make more sense. In other words, we could use the appraisals of one year as training data and test on the appraisals of the other year.
- For the second level of subjectivity, we could evaluate the percent of relevance that experts staff chose by applying it on other data.

References:

[1] Piazza F., Strohmeier S., "Domain-Driven Data Mining in Human Resource Management: A Review", *in Proceedings of Data Mining Workshops (ICDMW), IEEE 11th Int. Conf.*, 2011.

[2] Suriyakumari V., Kathiravan A. V., "An Ubiquitous Domain Driven Data Mining Approach For Performance Monitoring in Virtual Organizations Using 360 Degree Data Mining & Opinion Mining", *in proceedings of Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 IEEE Int. Conf.,* 2013.

[3] Richards L., "What Are the Problems With Performance Appraisals?", Small Business by Demand Media available: http://smallbusiness.chron.com/problems-performance-appraisals-1913.html [Accessed: 22nd October, 2015].

[4] Banfield A., "Unspeakable Sentences: Narration and Representation in the Language of Fiction", Routledge & Kegan Paul, 1982

[5] Quirk R., et, al., "a Comprehensive Grammer of the English Language", Longman, 1985.

[6] Wilson T., "Fine Grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States", PHD Dissertation, CS Dept., Pittsburgh Univ., 2007.

[7] VeeraKarthik M., Elamparithi M., "Enhance the Text Clustering using an Efficient Concept-Based Mining Model", Advanced research in Computer Science & Technology, vol. 2, no. 3, 2014.

[8] Dell Inc. "Statistics: Methods and Applications". Available: http://documents.software.dell.com/Statistics/Textbook/Text-Mining Last update: 14th May, 2015 [Accessed: 22nd, October 2015].

[9] Gaikwad S., Chaugule A., et. al., “Text Mining Methods and Techniques”, International Journal of Computer Applications, vol.85, no.17, 2014.

[10] Singh V., Dubey S.K.., "Opinion Mining and Analysis: A Literature Review", *in proceedings of Confluence the Next Generation Information Technology Summit (Confluence), 2014 IEEE Int. Conf.,* 2014.

[11] Pang B., Lee L., "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval, vol.2, no.2, 2008.

[12] Veselovská K., “Sentence-Level Polarity Detection in a Computer Corpus”, *in proceedings of WDS'11 of Contributed Papers,* 2011.

[13] Bing L., “Sentiment Analysis and Opinion Mining”, Morgan & Claypool Publishers, 2012.

[14] Das A., Bandyopadhyay S., “Theme Detection an Exploration of Opinion Subjectivity”, *in proceeding of Affective Computing and Intelligent Interaction and Workshops (ACII), IEEE Int. Conf.,* 2009.

[15] Sharma S. K., “Handbook of HRM Practices: Management Policies and Practices“, Global India Publication, 2009.

[16] Spitzer D., “Transforming Performance Measurement: Rethinking the Way we Measure and Drive Organizational Success”, AMACOM, 2007.

[17] Bâra A., Lungu I., “Improving Decision Support Systems with Data Mining Techniques", Advances in Data Mining Knowledge Discovery and Applications, Intech, 2012.

[18] Fayyad U., Piatetsky-Shapiro G., et. al., “The KDD Process for Extracting Useful Knowledge from Volumes of Data”, Communications of the ACM, vol.11, no.39, 1996.

[19] Fayyad U., "Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases", *in Proceedings of Scientific and Statistical Database Management, IEEE Int. Conf.* 1997.

[20] Abu Hammad A., "An Approach for Detecting Spam in Arabic Opinion Reviews", M.S. Thesis, IT Dept., IUG Univ., Gaza, 2013.

[21] Verma V. K., Ranjan M., et. al., "Text Mining and Information Professionals Role, Issues and Challenges*", in Proceedings of Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), IEEE Int. Conf.* 2015.

[22] Nidhi, Vishal G., "Recent Trends in Text Classification Techniques", International Journal of Computer Applications, vol.35 no.6, 2011.

[23] Salloum W, Habash N, "Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation", *in Proceedings of Empirical Methods in Natural Language Processing,* 2011.

[24] Al Najjar M., "Automated Complaint System Using Text Mining Techniques", M.S. Thesis, IT Dept, IUG univ, Gaza, 2013.

[25]] Gupta V., Lehal G., "a Survey of Text Mining Techniques and Applications", Emerging Technologies in Web Intelligence, vol.1, no.1, 2009.

[26] Turney P., Pantel P., "From Frequency to Meaning: Vector Space Models of Semantics", Journal of Artificial Intelligence Research, vol.37, no.1, 2010.

[27] Zhengwei Q., Gurrin C., et. al., "Term weighting approaches for mining significant locations from personal location logs", *Proceedings of Computer and Information Technology, IEEE Int. Conf.*, 2010.

[28] Hotho A., Nürnberger A., et. al., “A brief survey of text mining “, GLDV Journal for Computational Linguistics and Language Technology, vol.20, no.1, 2005.

[29] Polamuri S., “Supervised and Unsupervised Learning”, dataaspirant. Available: http://dataaspirant.com/2014/09/19/supervised-and-unsupervised-learning/ [Accessed: 24th, October, 2015]

[30] Wu X., Kumar V., et. al., “Top 10 algorithms in data mining”, Knowledge Information Systems, vol.14, no.1, 2008.

[31] Yottamine Analytics, "Analytics: The Machine Learning Advantage", 2011 Available: http://yottamine.com/machine-learning-svm. [Accessed 1st September., 2015].

[32] Liu B., ”Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data”, springer 2011.

[33] Saralegi X., San I., et. al., “Cross-Lingual Projections vs. Corpora Extracted Subjectivity Lexicons for Less-Resourced Languages”, *Proceedings of Computational Linguistics and Intelligent Text Processing,* 2013.

[34] Seasr, “Text Mining Overview” Available: http://www.seasr.org/documentation/text-mining-overview/ [Accessed: 24th, October, 2015]

[35] Hou X., Dong Y, et. al., "Application of Fuzzy Data Mining in Staff Performance Assessment*", in Proceedings of Conference Machine Learning and Cybernetics, IEEE Int. Conf.*, 2007.

[36] Youzheng C., Ming G., "Data Mining to Improve Human Resource in Construction Company", in *Proceedings of Business and Information Management,. ISBIM '08. Int. Seminar*, 2008.

[37] Kapoor B., “Business Intelligence and its Use for Human Resource Management”, The Journal of Human Resource and Adult Learning, vol.6, no.2, 2010

[38] Backer S., "Data Mining Moves to Human Resources", Business Week, 11th,March, 2009 available: http://www.businessweek.com/stories/2009-03-11/data-mining-moves-to-human-resources. [Accessed: 1st January, 2015].

[39] Bridgwater A., “IBM: 80 percent of our global data is unstructured (so what do we do?)”, Computer Weekly, 26th,Oct. 2010 available: http://www.computerweekly.com/ blogs/cwdn/2010/10/ibm-80-percent-of-data-is-unstructured-so-what-do-we-do.html. [Accessed: 2nd January, 2015].

[40] Kopackova H., Komarkova J., et. al., “Decision making with textual and spatial information", Wseas Transactions on Information Science & Applications, vol.5, no.3, 2008.

[41] Hoffmann L., “Mine Your Business”, Communications of the ACM, vol.53., no.6, , 2010.

[42] Bolasco S., Canzonetti A., “Understanding Text Mining: a Pragmatic Approach”, *in Proceedings of the NEMIS 2004 Final Conference*, 2004.

[43] Dave K., Lawrence S., et. al., " Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", *in Proceedings of the 12th Int. Conf. on World Wide Web*, 2003.

[44] Zabin J, Jefferies A., "Social media monitoring and analysis: Generating consumer insights from online conversation", Aberdeen Group Benchmark Report, 2008.

[45] Bloom K., Garg N, “Extracting Appraisal Expressions”, *in Proceedings of North American Chapter of the Association for Computational Linguistics*, 2007.

[46] Wang X., Fu G., "Chinese Subjectivity Detection Using A Sentiment Density-Based Naive Bayesian Classifier", *in Proceedings of Machine Learning and Cybernetics (ICMLC), IEEE Int. Conf.*, 2010.

[47] Tang H., Tan S., et. al., "A survey on Sentiment Detection of Reviews", Expert Systems with Applications, vol.36, no.7, 2009.

[48] Lu B., Tsou B. K., "Combining a Large Sentiment Lexicon and Machine Learning For Subjectivity Classification", *in Proceedings of the 9th Int. Conf. on Machine Learning and Cybernetics,* 2010.

[49] Pang B., Lee L., "A Sentiment Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", *in Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics,* 2004.

[50] Riloff, E., Wiebe J., et. al., "Learining extraction patterns for subjective expressions", *in Proceedings of Empirical Methods in Natural Language Processing,* 2003.

[51] Yu H., Hatzivassiloglou V., "Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences", *in Proceedings of Conference on Empirical Methods in Natural Language Processing,* 2003.

[52] Biyani P., Caragea C., et. al., "I want what I need! Analyzing Subjectivity of Online Forum Threads", *in Proceedings of the 21st ACM Int. Conf. on Information and knowledge management,* 2012.

[53] Verma T., Renu, et. al., "Tokenization and Filtering Process in RapidMiner", International Journal of Applied Information Systems (IJAIS), vol.7, no.2, 2014

[54] C. Ramasubramanian, R. Ramya, "Effective Pre-Processing Activityes in Text Mining using Improved Porter's Stemming Algorithms", International Journal of advanced research in Computer And Communication Engineering, vol.2, no.12, 2013.

[55] Saad M., Ashour W., "Arabic Morphological Tools for Text Mining", *in Proceedings of 6th International Conference of Electrical and Computer Systems*, 2010.

[56] Monali P., Sandip K., "A Concise Survey on Text Data Mining", International Journal of Advanced Research in Computer and Communication Engineering, vol.3, no.9, 2014.

[57] Szczepaniak P. S., Segovia J., et. al., "Studies in Fuzziness and Soft Computing: Intelligent Exploration of the Web", springer 2003.

[58] Singhal A., "Modern Information Retrieval: A Brief Overview", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2001.

[59] Aggarwal C., Zahi C., "A Survey of Text Classification Algorithms", Mining Text Data, Springer 2012.

[60] Zhou G., Zhang M., et. al., "Hierarchical Learning Strategy in Relation Extraction Using Support Vector Machines", *in Proceedings of Third Asia Information Retrieval Symposium*, 2006.

[61] Rapidminer Documentation, "Support Vector Machine (LibSVM)" Available: http://docs.rapidminer.com/studio/operators/modeling/classification_and_regression/Svm/support_vector_machine_libsvm.html [Accessed: 29th August, 2015].

[62] Rapidminer Documentation, "Naïve Bayes" Available: http://docs.rapidminer.com/studio/operators/modeling/classification_and_regression/bayesian_modeling/naive_bayes.html [Accessed: 29th August, 2015].

[63] Rapidminer Documentation, "K-NN" Available: http://docs.rapidminer.com/studio/operators/modeling/classification_and_regression/lazy_modeling/k_nn.html [Accessed: 29th, August, 2015].

[64] Han J., Kamber M., "Data Mining Concepts and Techniques", Elsevier Inc., 2012.

[65] Chawla N. V., Hall L. O., et. al., "Wrapper-based Computation and Evaluation of Sampling Methods for Imbalanced Datasets", *in Proceedings of 1st Int. workshop on Utility-based data mining*, 2005.

[66] Hand D., Mannila H., et. al., "Principles of Data Mining", Massachusetts Institute of Technology, 2001.

[67] Maimon O., Rokach L., "Data Mining and Knowledge Discovery Handbook", springer 2010.

[68] Herschel G., Linden A., et. al., "Magic Quadrant for Advanced Analytics Platforms" Available:http://www.gartner.com/technology/reprints.do?id=12AHPOU0&ct=150225&st=sb [Accessed: 24th,October, 2015].

[69] Saad M. K., "The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification", M.S. Thesis, Computer Engineering Dept., IUG Univ., Gaza, 2010.

[70] Duwairi R., Al-Rafai M., "Feature Reduction Techniques for Arabic Text Categorization", American Society for Information Science and Technology, vol.50, no.11, 2009.

[71] Witen I. H., Frank E., et. al., "Data Mining Practical Machine Learning Tools and Techniques", Morgan Kaufmann, 2011.

Appendix A

Objective Wordlist

Table A.1: objective word list for first level of subjectivity.

Stemmed token	Origin	Stemmed token	Origin	Stemmed token	Origin
Lcd	ICDL	مون	امانة	قرر	قرارات
حيا_نسب	احياء مناسبات	جلس	مجالس	حصا	حصص
حلل_حوي	تحليل المحتوى	غور	غيور	حصص_ضيف	حصص اضافية
سبر_ذكي	سبورة ذكية	سور	سير،مسير	طابور	طابور
درس_قوا	دروس تقوية	شرع	مشروع	أثر	أثر
زور_بدل	زيارات تبادلية	طوع	تطوع	حمد	حميدة
نصح	نصح	رجع	مراجعة	متع	يتمتع
وظب	مواظبة	كشف	كشافة	ثقف	تثقيق
سبر	سبورة	لوح	لوحات	فهم	فهم
سلب	اسلوب	ولف	ملف	نظف	نظافة
زون	تزيين	حفل	حفلات	جدي	مستجدات
عبأ	عبئ دراسي	توج	انتاج	هنأ	مهنية
جرب	تجارب	غدر	مغادرة	برنامج	برنامج
جدر	جدارة، جدارية	طرح	طرح	حوب	حب، محبة
عني	معنويا	مهن	مهنية	صحي	صحية
ودي	ودية، دائما	حبب	محبوب	صباح	صباح
دعم	دعم	عقد	عقد	ولي_مور	اولياء الأمور
بسط	تبسيط المادة	شجع	تشجيع	عرض	عرض
نقش	مناقشة	هجأ	منهجية	حوج	احتياجات
لقي	القاء	ألف	ألف	تقن	اتقان
أدائه	ادائه	قنأ	تقنيات	صحح	تصحيح امتحانات
ورش	ورش	حاسوب	حاسوب	هرأ	مهارات
أدب	أداب	بذل	بدل جهد	درج	تدرج، درجة
روح	روح	رشد	ارشاد	نهج	منهج
لعب	لعب	طوب	طيبة	رقأ	ارتقاء، ورقيا
وثق	توثيق	وزر	وزاري	عزز	عزز
خوذ	اتخاذ	دوأ	أدوات	فود	وفود
غرس	غرس	وتستثمر	تستثمر	ذيع_درس	اذاعة مدرسية
جوح	انجاح	مركز	مركز	قدي	قدوة
شرح	شرح	ركز	مركز	رتب	ترتيب
غادر	مغادرة	حوث	يحث	خلص	اخلاص
صلي	مصلى	عبر	تعبير	حول	تحمل
فوز	فوز	ندا	نادي	وقف	مواقف طارئة
طلع	اطلاع	نجز	انجاز	كفؤ	كفؤ
قصر	تقصير	قرح	مقترحات	ورق_عمل	اوراق عمل
مهر	مهارة	ميد	مادة	عمد	اعتماد

Stemmed token	Origin	Stemmed token	Origin	Stemmed token	Origin
وجب	واجبات	فوق	تفوق	خطط	تخطيط
جدد	تجديد	نتج	نتائج	دور	ادارة
أسس	أسس	صفا	صفية، لاصفية	نشط	انشطة
خصص	تخصيص	وسل	وسائل	نظم	انظمة
بدع	ابداع (مهارات ابداعية)	علق	علاقات،علاقة	طلب	طلاب
مول_عوم	مال عام	وصل	تواصل	درس	مدرسة
حدث	حديثة	حصص	حصص	علم	معلم
وكل	موكل، وكيل	درأ	اداري	جمع	مجتمع
قون	قوانين	وضح	توضيحي	شرف	اشراف
عضو	عضو	سول	اسئلة	حمل	تحمل
نوب	مناوبة	ضعف	ضعف	لجن	لجنة
رعي	مراعاة ، رعاية	فصل	فصل	جبي	واجبات
مجل	مجال	ثري	اثراء	بحث	بحث،مبحث
حدد	تحديد	وظف	توظيف	غيب	غياب
ملك	ممتلكات، املاك	عون	تعاون	طور	تطوير
كلف	مكلف	سبق	مسابقات	سعد	المساعدة ، مستعد
درب	تدريبات، تدريب	طرا	طارئة	علج	علاج
سجل	سجل حضور	فرق	فروقات، فريق	صرف	انصراف
لوب	اسلوب	هدف	هدف	ذكي	ذكي
هوم	مهام، مهم	زمل	زملاء	كرم	تكريم
كتب	كتاب	ضبط	ضبط	جدول	جدول
خلق	اخلاق	هتم	اهتمام	نجح	نجاح
ضغط	ضغط	صفف	صف	صول	اتصال
عرف	معرفة	خبر	مختبر	حلل	تحليل
جزأ	اجازات	ربأ	تربية	تستثمر	تستثمر
جهد	جهد	ثمر	استثمار	دقق	دقيقة
بكر	باكرا	بدر	المبادرة	سلك	سلوكيات
كفأ	كفاءة	حسن	حسن، حسنة	سأل	مسئوليات
نمي	تنمية	وعد	مواعيد	دوم	دوام
نسق	تنسيق	حصل	تحصيل	حضر	حضور
سمر	مستمر	قوم	تقويم	سهل	تسهيل
عطي	اعطاء	قدر	قدرة	فكر	تفكير
		صلح	مصلحة	سوا	مستوى