

Sentiment Analysis Algorithm for Arabic Reviews on the Movies Domain

خوارزمية لتحليل المشاعر للتعليقات العربية على نطاق الافلام

Prepared By:

Khalid Waleed Nassar Al-mansoori

Supervisor:

Prof. Riyad Al-Shalabi

This thesis is submitted as the Partial Fulfillment of the Requirement for the
Master Degree in Computer Science

Faculty of Computer Sciences and Informatics

Amman Arab University

June / 2017

Authorization

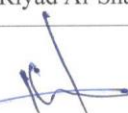
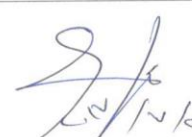


Form (9)

College of Scientific Research and Graduate Studies

Authorization

We, the undersigned, pledge to grant Amman Arab University for discretion in the publication of the academic content of the dissertation, so that the intellectual property rights of Master's thesis be back to the university in accordance with the laws, regulations and instructions relating to intellectual property and patent.

| Advisor Name | Student Name |
|--|---|
| Prof. Riyad Al-Shalabi | Khalid Waleed Nassar Al-mansoori |
| Signature:  Date: July 2, 2017 | Signature:  Date: 2/2/17 |


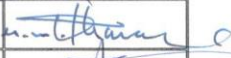

شارع الأردن - موبص - هاتف: 6954 0040 +962 7 - ص.ب. 2234 عمان 11953 - الأردن
 Jordan Street - Mubia - Telephone +962 7 8094 0040 - P.O.Box 2234 Amman 11953 - Jordan
 Email: waaga@bau.edu.jo / Web: www.bau.edu.jo

Committee Members Decision

Committee Members' Decision

The thesis entitled: "Sentiment Analysis Algorithm for Arabic Reviews on the Movies Domain" was submitted by the student, Khalid Waleed Nassar Al-mansoori was examined and approved on 14/6/2017.

Committee Members

| Name | | Signature |
|------------------------|------------------|--|
| Prof. Riyad Al-Shalabi | Chair/Advisor |  |
| Dr. Akram Mashaykhi | Member |  |
| Prof. Asem Al-Shekh | External/ Member |  |

الاهداء

اهدي ثمرة جهدي المتواضع

الى والدي ووالدي حفظهم الله ورعاهم...

الى زوجتي رفيقة دربي التي كانت عوناً وسنداً لي طيلة فترة دراستي...

الى اخواني أسأل الله تعالى ان يوفقهم...

الى اساتذتي الكرام الذين ساهموا في تكويني منذ المرحلة الابتدائية الى يومنا هذا...

ACKNOWLEDGMENT

I want to thank ALLAH for his blessings that help me achieve my dream. I would like to thank my supervisor Prof. Riyad Al-Shalabi who supported and helped me to complete this thesis. I would like to thank him because he was always available when I needed his help. Also, I would like to thank everyone who contributed to the elaboration of this thesis.

Table Of Content

| | |
|---|---------|
| Authourization | II |
| Committee Members Decision | III |
| IV | الإهداء |
| ACKNOWLEDGMENT | V |
| Table Of Content..... | VI |
| Contents | VII |
| LIST OF FIGURES..... | IX |
| List Of Tables..... | XI |
| Abstract..... | XII |
| XIII..... | الملخص |
| Chapter one Introduction | 1 |
| Chapter Two Literature Reviews..... | 5 |
| Chapter 3 Research Methodology..... | 15 |
| Chapter four Experimntal and Result | 33 |
| Chapter Five Conclusions and Recommendations For Future Work..... | 57 |
| References | 59 |

Contents

| |
|--|
| Subject |
| CHAPTER ONE (Introduction) |
| 1.1 Introduction |
| 1.2 Research Problem |
| 1.3 Reserch Questions |
| 1.4 Significance Of The Study |
| 1.5 Research Model |
| 1.6 Limitations and Delimitations of the Study |
| 1.7 Thesis Organization |
| CHAPTER TWO (Literature Survey) |
| 2.1 Sentiment Analysis |
| 2.2 Movies Reviews |
| 2.3 Arabic Language Reviews |
| 2.4 Related Work |
| CHAPTER THREE (Research Methodology) |
| 3.1 Introduction |
| 3.2 Movie Rating for Arabic Reviews System |
| 3.3 Procedures and Methodologies |
| 3.3.1 Build Sentiment Lexicon |
| 3.3.2 Collect Sentences Online |
| 3.3.3 Translate English Reviews to Arabic |
| 3.3.4 Split the Sentence into Words |
| 3.3.5 Matching Process |
| 3.3.6 Assign Label to The Sentences |
| 3.3.7 Check the Accuracy |

| |
|---|
| CHAPTER FOUR |
| The Experimental Works |
| 4.1 Introduction |
| 4.2 Movie Rating For Arabic Reviews System |
| 4.3 Frame Works Movie Rating For Arabic Reviews System |
| 4.3.1 Main Menu Screen |
| 4.3.2 Read File Menu Screen |
| 4.3.3 Add Comment Menu Screen |
| 4.3.4 Add Rule Menu Screen |
| 4.3.5 About Menu Screen |
| 4.3.6 Execute Menu Screen |
| 4.3.7 Message Alarm Screens |
| 4.3.7.1 Read File Alarm Screen |
| 4.3.7.2 Add Comment Alarm Screen |
| 4.3.7.3 Execute Alarm Screen |
| 4.4 Examples for Movie Rating for Arabic Reviews System |
| 4.4.1 Example 1 - Sentiment Analysis for Arabic Reviews of (The Boss Baby 2017) Movie |
| 4.4.2 Example 2 - Sentiment Analysis for Arabic Reviews of (THE CIRCLE 2017) Movie |
| 4.4.3 Example 3 - Sentiment Analysis for Arabic Reviews of Many Movies |
| CHAPTER FIVE |
| Conclusions And Recommendations For Future Work |
| 5.1 Introduction |
| 5.2 Conclusions |
| 5.3 Recommendations for Future Work |
| REFERENCES |
| REFERENCES |

LIST OF FIGURES

| Figure number | Figure name |
|---------------|--|
| 1 | Research Model steps |
| 2 | Flowchart of the research steps |
| 3 | Positive lexicon words example |
| 4 | Negative lexicon words example |
| 5 | Neutral lexicon words example |
| 6 | Sentiment lexicon model flowchart |
| 7 | The programing code for split the sentence into words |
| 8 | The First Stage for macting process |
| 9 | The programing code for the First Stage for macting process |
| 10 | The Second Stage for macting process |
| 11 | The programing code for The Second Stage for macting process |
| 12 | Main menu screen |
| 13 | Read file menu screen |
| 14 | Add comment menu screen |
| 15 | Add rules menu screen |
| 16 | About menu screen |
| 17 | Execute menu screen |
| 18 | Alert message the path cannot be null |
| 19 | Finish loading file |
| 20 | Alert message save change |
| 21 | Finish alert message |
| 22 | The matching process |
| 23 | The matching process in oracle form |
| 24 | Result execute for the boss baby movie, page1 |
| 25 | Result execute for the boss baby movie-page 2 |

| | |
|----|--|
| 26 | Result execute for the boss baby movie-page 3 |
| 27 | Result execute for the boss baby movie-page 4 |
| 28 | Result percentage for the boss baby movie |
| 29 | Result execute for the circle movie-page 1 |
| 30 | Result execute for the circle movie-page 2 |
| 31 | Result execute for the circle movie-page 3 |
| 32 | Result percentage for the circle movie |
| 33 | Example 3 result percentage |
| 34 | The accuracy results |

List Of Tables

| Table number | Table name |
|--------------|-------------------------------------|
| 1 | Number of lexicon words |
| 2 | Examples for count words in lexicon |
| 3 | Examples for many movies |
| 4 | The accuracy results |

Sentiment Analysis Algorithm for Arabic Reviews on the Movies Domain

Prepared by

Khalid Waleed Nassar Al-Mansoori

Supervised by

Prof. Riyad Al-Shalabi

Abstract

Sentiment analysis has become popular in recent years and has been used in many life aspects where the people reviews are very important to know their reaction on any product. Regarding movies, these impressions are important for people to choose the best film to watch.

Movie reviews are a very important tool for measuring the performance of the movie, most of the sentiment analysis research in those reviews are in English language, but our goal is how to analyze the feelings and reviews using Arabic language.

In this thesis, an Arabic lexicon for movies was created and the lexicon is used in applying movies sentiment analysis.

Movie Rating for Arabic Reviews System (MRARS) was built for sentiment analysis for Arabic movies reviews using Oracle Language.

The lexicon contains all the feelings words about movies which are positive, negative and neutral sentiment.

We collected about 1250 words, the number of positive words is 655, the number of negative words is 441 and the number of neutral words is 154.

The proposed system showed excellent performance varied between 94% - 100% and the accuracy can be raised after adding new words to the lexicon.

خوارزمية لتحليل المشاعر للتعليقات العربية على نطاق الافلام

إعداد

خالد وليد نصار المنصوري

اشراف

الأستاذ الدكتور رياض الشلبي

الملخص

أصبح تحليل المشاعر منتشراً في السنوات الأخيرة وتستخدم في كثير من مجالات الحياة حيث تعليقات الناس مهمة جداً لمعرفة مشاعرهم حول أي منتج. وفي ما يتعلق في الافلام فهذه الانطباعات مهمة للناس لاختيار افضل فلم لمشاهدته.

تعليقات الافلام هي أداة مهمه جداً لقياس تأثير الافلام ومعظم بحوث تحليل المشاعر في تلك التعليقات هي في اللغة الانجليزية ولكن هدفنا هو تحليل المشاعر باستخدام اللغة العربية هذه الرسالة تناولت هذا الموضوع و وضعت معجماً عربياً خاصاً للافلام ، الذي سيستخدم لتحليل المشاعر للافلام، وهذا المعجم يحتوي على جميع كلمات المشاعر للافلام والتي هي جيدة، سيئة، ومحايدة .

تم جمع حوالي 1250 كلمة، عدد الكلمات الإيجابية هو 655، عدد الكلمات السلبية هو 441 وعدد الكلمات المحايدة هو 154.

تم بناء نظام تصنيف الافلام للتعليقات العربية (م.ر.ا.ر.س) لتحليل المشاعر لتعليقات الافلام العربية باستخدام لغه اوراكل لكتابة البرنامج. واطهر النظام المقترح أداءً ممتازاً يتراوح بين 94% إلى 100% ويمكن أن تتاثر الدقة بعد إضافة كلمات جديدة إلى المعجم.

Chapter one

Introduction

1.1 Introduction

Sentiment analysis is used in many life aspects and has become popular in recent years. Where the reviews of the people are very important to know their feelings about a particular product. There is too much data on the internet and it is difficult to assess it personally, for example, movie reviews essay or opinion about a particular figure. Sentiment analysis can be handy in the fields of politics, where the politicians care about people opinions regarding their policy. Companies can benefit from this technique to increase their knowledge about people review related to their performance, products and what products will be best sellers and spread in the markets (Vishwanathan, 2014); (Vinodhini, 2012).

In present days, people seek their friend's impression about a certain movie, whether it is worth or not before they decide to spend money watching it in the cinema. These impressions are important for people to choose the best film to watch (Pang and Lee, 2008).

Movie reviews are very important tool for measuring the performance of movies and at the same time famous actors are interested in knowing people's feelings about the movies they made. Also, it assists producers (film production companies) to learn about films that were produced and people's reaction about movies and recognizing possible future improvement they can resolve. Each language has difficult aspects in sentiment analysis that varies from one language to another where each language has its rules and characteristics in analyzing emotions. (Nabil and et al., 2014).

There are many research that has been done in sentiment analysis on the internet sites in general, news articles and social media sites and how to express feelings because the informal language was written by people and message-length containing of microblogging. The sentiment lexicons have proved very useful in sentiment analysis and knowing what each person feels about a particular subject. (Kouloumpis and et al., 2011)

Most of the sentiment analysis research is in English language, but our goal is how to analyze the opinions using Arabic language.

1.2 Research Problem

After extensive research, there isn't any research that has explored sentiment analysis for Arabic reviews on the movie domain. To best of my knowledge, there is no any Arabic sentiment lexicons for the Arabic language on the movie domain. Sentiment analysis is more accurate if applied to a certain domain and for a specific language. In general, there is limited research on sentiment analysis in Arabic.

1.3 Research Questions

This research answers the following questions:

How to build the lexicon for Arabic reviews on the movies domain?

How to increase the lexicon accuracy ?

How can the sentiment analysis be treated in the context of Arabic?

Can this system be effective when handling movie reviews ?

1.4 Significance of the Study

This research is important to check if the movie is highly rated easily. It is also important for movie owners and can predict the revenue of the movie.

This research explored sentiment analysis for Arabic reviews on the movie domain. An Arabic lexicon for movie domain will be created and the lexicon will be used to apply sentiment analysis to movie.

1.5 Research Model

Figure (1) summarizes the research model steps in present work:

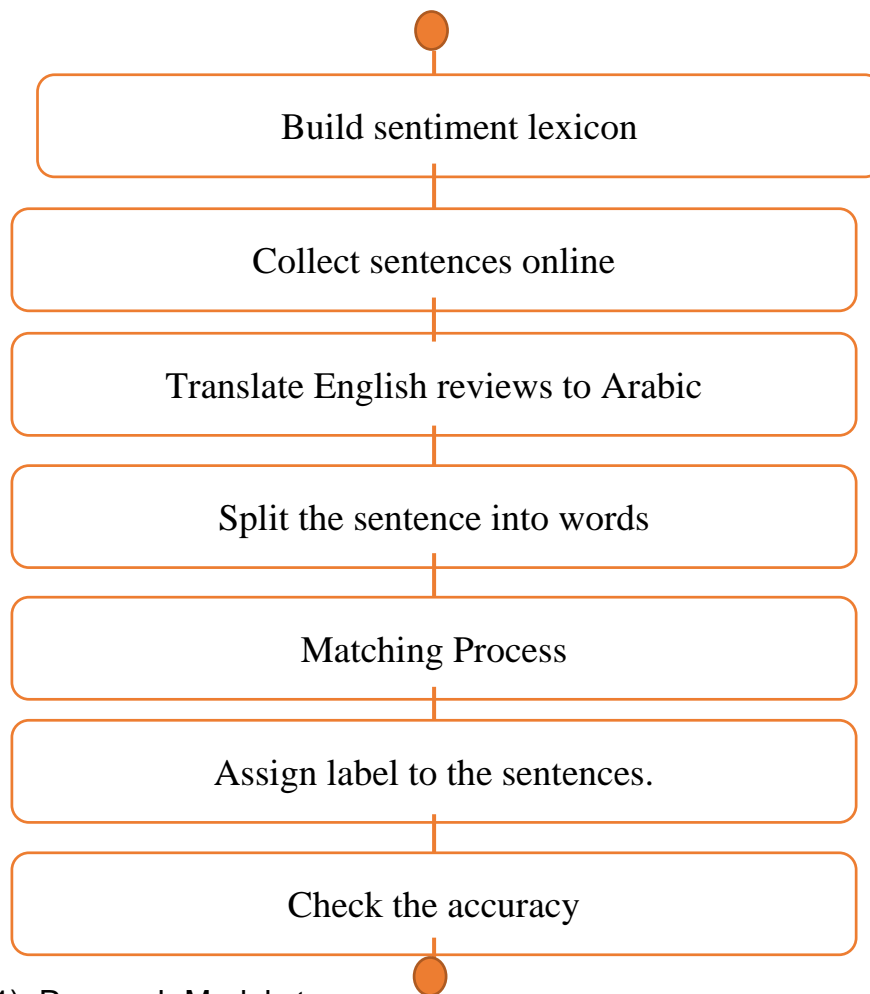


Figure (1): Research Model steps

1.6 Limitations of the Study

This research deals with sentiment analysis for Arabic reviews on the movies domain. The most anticipated restrictions and challenges are:

Choosing the best translation for English reviews, since this translation should be efficient for this study that specialized in sentiment analysis. This translation should express the sentiment accurately and gives the correct meaning after being translated to Arabic language.

Time required to collect movies with rich reviews that contain lots of keywords to provide clear audience impression about the movie.

Finding suitable movies that inspire Arab audience, considering that not all foreign or English movies are within Arab audience concern.

The lack of obvious researcher in the area of the thesis, there are many studies talking about sentiment analysis in general, but absence of studies that are specialized in sentiment analysis regarding the movie domain, specially that this study is the first of its kind to analyze movies Arabic reviews.

Writing the programming code for the system that contains a rich database of keywords and can extract words from the reviews to be compared with the database and then determine the movies quality from audience point of view.

1.7 Thesis Organization

This thesis contains research related to creating a system to analyze Arabic reviews. This research explores sentiment analysis on the movie domain.

This thesis includes five chapters as follows:-

Chapter one presents the research questions, research model, limitations and delimitations of the study and thesis organization.

Chapter Two explains the main concepts of the sentiment analysis presents some of the literature reviews that are related to this thesis.

Chapter Three presents the proposed methodology that has been followed in the practical part of this thesis and represents it through algorithms and flowcharts.

Chapter Four is devoted for the experiments and result

Chapter Five summarizes the main conclusions of this thesis and suggesting recommendations for future.

Chapter Two

Literature Reviews

2.1 Sentiment analysis

Sentiment analysis or Opinion mining is the computational study of people's opinions, evaluation, point of view, and emotions toward entities, issues, facts, topics and their attributes. Sentiment analysis tend to take advantage of building sentences and extract the words from this sentences that we need, where we can extract these words through Text analysis of news, documents from the political, movie reviews, twitter or from the Internet in general, where sentiment analysis can determine the best feelings and emotions to their holders. (Maks and Vossen, 2012).

One approach of sentiment analysis labeled texts and use the machine learning to classify the texts to positive and negative polarity .(Nielsen,2011).

We observed that the automated sentiment analysis system determines positive opinions when the context of the news is good as talking about peace. The negative opinions determine when the context of the news speaks about negative things such as the financial crisis and the wars were beginning. (Balahur and et al., 2013)

Sentiment analysis is an application of natural language processing and analysis texts from public data. Although the sentiment analysis is more effective at specific domains because each domain has its own words meanings. Some words contain positive or negative feelings such as the words 'good' or 'bad', but some words depend on the specific domain. An example for this is the word 'early' which may reflect negative subjectivity in movies as in the instance "The movie was displayed too early!", then again, when describing a parcel service such as "The parcel arrived early", this is most likely a positive sentiment. (Altrabsheh, 2016)

Sentiment analysis is used to determine a particular text i.e determining whether the expressed opinion is positive, negative or neutral. However, it is used to detect emotion, which plays an important role in the analysis of movies, for example, the positive movie reviews help filmmakers to increase production like this movie which makes people happy, more efficiency and increase their awareness, where emotions have the ability to energize people about good thinking. For example, feelings of sadness and anger have a negative impact on society.

(Altrabsheh, 2016)

Sentiment analysis research hve grown considerably in the last decade, mainly due to the availability of rich text resources such as social networking sites, blogs and micro-blogs, movie reviews and product reviews.

The websites contain many documents, search of information is very important for users and the most important information found in the texts spread on the sites, and to know if this information is useful or not to the user. The reviews of people have become to decide that. Movie review is also the most important information for users who go to the movie.

(Tsutsumi and et al., 2007)

Opinion can be expressed in different forms like the Amazon site for reviewing products, or movie review sites such as RottenTomatoes. Usually, the reviews on these sites are long, consisting of several paragraphs, and there are other types of sites that contain short messages such as social networking sites Twitter and Facebook. (Yessenov and Misailovic, 2009)

2.2 Movies Reviews

The movie review is very important for revenues that come from them. Knowing the advantages and disadvantages factors for any movie is significant to avoid the weakness that movie includes and help produce future movies that satisfy the audience and can increase the profit. (Joshi and et al., 2010)

The created reviews over the web is considered as a main resource about most products including the movies. Most people pay attention to what is written. They care about the products to take an idea about them for making a decision depending on online reviews and many studies proved that online reviews are most reliable and authentic for the user. (Koh and et al., 2010)

Movie reviews have the unique characteristics when a user writes a comment about the film, the user may not comment on the elements of any film directly (such as the scenario, the music and the effects within the film), but also comment on the producers and actors and what the user sees about performance of each actor. The advantages of movie reviews are richer than in other products reviews, Therefore, the movie reviews are more difficult than products reviews. (Zhuang, and et al., 2006)

2.3 Arabic Language Reviews

Arabic is a Semitic language which is very rich of information and has many meanings and characteristics different from other languages cts such as derivation, inflection, and agglutination.

Arabic is a morphologically rich language, the Arabic language inherently has a high number of variable word forms leading to data sparsity.

The Arabic language is both challenging and interesting because of its history, the strategic importance of its people, the region they occupy, and its cultural and literary heritage.

As the official language of 22 countries, Arabic is spoken by more than 300 million people, and is the fastest-growing language on the web and with the great increase of the Middle East community for using the websites and writing comments on many topics, especially in social media sites such as politics, news, show different types of movies and sports, the Arab sentiment analysis has become very important about knowing what the Arab community feeling about these topics. (Korayem, and et al., 2012)

However, progress in the analysis of Arab sentiment are slow compared to English and other languages. (Abdul-Mageed and et al., 2014) (El-Beltagy and Ali, 2013) (Refaee and Rieser , 2014)

2.4 Related work

A study by (Vishwanathan, 2014) with title " Sentiment Analysis of Movie Reviews" shows sentiment analysis has become one of the most important feelings research areas in the search for prediction and classification especially in research about sentiment analysis for movie review .

The study was interested in the analysis of texts and understanding the meanings of sentences using appropriate tokenization and thereafter classifying. A set of data has been built from the website 'Rotten Tomatoes' which provides movie reviews. They analyzed the reviews into positive or negative using word stem tokenization.

A study by (Nabil and et al. ,2014) with title " LABR: A Large Scale Arabic Sentiment Analysis Benchmark " research paper discusses the LABR, the largest sentiment analysis dataset to-date for the Arabic language. It consisted of over 63,000 book reviews. The users rated it on a scale from 1 to 5 stars. Properties of the dataset were investigated, and statistics were presented. A dataset was used for two tasks: sentiment polarity classification; and rating classification. A sentiment lexicon was created that contains many sentiment words indicating feelings and explore its effectiveness which is not compatible to movie domain.

A study by (Alotaibi, 2015) with title " Sentiment Analysis in the Arabic Language Using Machine Learning " The aim of their study was to analyze the sentiment in Arabic. They use nonlinear machine learning classifiers for their study. In another aspect, their study discusses the effect of negation in Arabic character and polarity classification. Two methods were suggested that prove negation in the Arabic language. The first method is that any sentence containing negation words is considered a negated sentence. The second method depends on a dynamic method. A number of comment negation dataset is required to create a model to establish the effect of the negation on the sentence.

The research claims that sentiment analysis has become in recent years an important topic, especially in research related to natural language processing and machine learning where they utilized a wide range of companies and politicians and users of social networking sites to find out people's feelings about a particular topic.

Research in Arabic on the topic of sentiment analysis remains in its infancy, therefore, there must be cooperation between research in this topic.

A study by (Zhang, 2014) with title " Text Mining for Sentiment Analysis " study shows that a few years ago with the development of internet services and the emergence of social media, people used social media most of their time and wrote their feelings about any topics. Their article identified four types of challenges to determine feelings basic sentiment expressing unit, a paucity of labeled data, domain dependence, and author modeling and they used two approaches, ReNew and Arch, to address these challenges.

ReNew, is a framework used in sentiment analysis; It can collect a large number of data to automatically generate a domain-specific sentiment lexicon and make an exclusive sentiment. Arch is a probabilistic model in sentiments analysis. It summarizes the sentiment analysis of the authors and measures the similarities between the sentiments.

A study by (Mejova, 2012) entitled " Sentiment analysis within and across social media streams " creates a comparison literature review on sentiment analysis. Her thesis began by introducing a rich multidimensional model based on affect control theory, researched in Messages and numerous texts on Twitter and blogs. She suggested that the better way to represent data is to build data-driven sentiment classifier.

A study by (Na and et al., 2010) entitled " Comparing sentiment expression in movie reviews from four online genres " research focuses on the characteristics and differences between the expression sentiment expression in movie review documents from four online opinion genres that are blog postings, discussion board threads, user reviews, and critic reviews.

In their publication, they researched many movie reviews from different sources of Internet networks. They analyzed 520 movie reviews which compares four types of normalizations. The analysis focused on document and sentence length, part-of-speech distribution, vocabulary, aspects, and analyzing selected Movie reviews to Positive or Negative . Their research main focus was about analyzing movie review documents.

A study by (Yang and et al., 2015) entitled " Sentiment analysis for Chinese reviews of movies in multi-genre based on morpheme-based features and collocations" researched sentiment analysis in the Chinese language which is more difficult than other languages such as English and European languages because Chinese language contains a lot of difficult words and phrases. Characters in the Chinese language are mono section where the words and phrases might be on a single form or grouped to form new words and expressions; their study classified movie reviews found on Yahoo Taiwan and utilized Pointwise Mutual Information (PMI) collocations.

They proposed a method for generating morphological rely on Chinese domain that is directly taken from the set of data without any previously identified sentimental resources. They built morpheme-based classifiers applicable in a different type of movies which has proven to be better than other classifiers in performance. It is based on keywords and also identifies compounds that have several semantic polarities depending on contexts.

If the approach was adopted in English language, the sentiment analysis in the Chinese language will become mainly biased in many forms.

Where the use of texts based on indivisible morpheme, the results appeared that the method can achieve high precision, especially across different types of films .

A study by (Kechaou and et al., 2013) entitled " A novel system for video news' sentiment analysis " the study aims to improves access to the news and people's opinions on a particular movie and compiling their views for the classification of films according to people feelings that were good or bad using Hidden Markov Model (HMM) and Support Vector Machine (SVM) hybrid learning method. A model that was conducted on the databases of people's opinions with different aspects films that include Information Gain (IG), Mutual Information (MI) and CHI-statistic (CHI).

A study by (Hasan, 2011) entitled " Proximity-based sentiment analysis " After research about sentences was written on social media sites, movie sites, and book sites they analyze these sentences into good or bad.

There are different levels: word, sentence, and document level especially based on common machine learning techniques such as Support Vector Machine (SVM), Naive Bayes (NB), and Maximum Entropy (ME).

In their paper, they used a new way of sentiment analysis called proximity-based. Their study focused on three different word proximity-based features, namely, proximity distribution, mutual information between proximity types and proximity patterns. Their approach applied to the analysis of movie reviews domain.

A study by (Al-Ayyoub and et al., 2016) entitled " Hierarchical Classifiers for Multi-Way Sentiment Analysis of Arabic Reviews " Most of the research focused on sentiment analysis by extracting words from sentences and analysis that positive or negative, Multi-Way Sentiment Analysis focuses on sentiments conveyed through a rating system (e.g., a 5-star rating system). A hierarchical classification structure used with this approach where each node performs a different classification sub-problem and the decision from it may lead to the invocation of another classifier.

In their paper, they used divide-and-conquer hierarchical structure of classifiers achieves better results than the use of existing flat classifiers for the Multi-Way Sentiment Analysis problem. Also, they focused on the Arabic Language, where there is little research in Multi-Way Sentiment Analysis of Arabic reviews, and collected a large scale Arabic Book Reviews (LABR) dataset.

unfortunately, the baseline experiments on this dataset had very low accuracy.

In their paper they compared two different hierarchical structures with the flat structure using different core classifiers based on standard accuracy measures to using the mean squared error (MSE).

The results show that, in general, hierarchical classifiers give significant improvements (of more than 50% in certain cases) over flat classifiers.

A study by (Alsmearat and et al., 2015) entitled " Emotion Analysis of Arabic Articles and Its Impact on Identifying the Author's gender "the study focuses on knowing the gender of the author by analyzing the text based on text content, there are many researchers on this topic in several languages but did little on the same topic in the Arabic language In their paper, they compared two approaches, the first approach is the Bag-Of-Words (BOW) and the second approach is based on computing features related to sentiments and emotions.

The goal is to confirm the validity of the common stereotype that female authors tend to write in a more emotional way than male authors. The results show there are no conclusive evidence that true for their dataset.

A study by (Basari and et al., 2013) entitled "Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization "At present days, social media and websites are the online discourse, where people contribute to writing, publishing, and blogging. Among the most widely used social media site is Twitter.

Twitter command includes reviews about movies, producers, and politics

Opinion mining refers to the texts mining to determine or classify whether the movie is good or bad based on the content message, where their research consists use the binary classification, which is classified into two classes.

Those classes are positive and negative. The positive class shows good message opinion; otherwise, the negative class shows the bad message opinion of certain movies.

This argument is based on the accuracy level of SVM with the validation process uses 10-Fold cross-validation and confusion matrix. The hybrid Particle Swarm Optimization (PSO) is used to improve the selection of the best parameter in order to solve the dual optimization problem.

A study by (Yessenov & Misailovic, 2009) entitled "Sentiment analysis of movie review comments "the researcher presents an empirical study of efficacy of machine learning techniques in classifying text messages by semantic meaning by (Yessenov & Misailovic) search about movie review comments from social network Digg (Digg formerly had been a popular social news website, with support for sharing content to other social platforms such as Twitter and Facebook) as our data set and classify text by subjectivity/objectivity and negative/positive, The process of extracting the text features through using large movie reviews corpus, threshold and using WordNet synonyms knowledge, (Yessenov & Misailovic) evaluate their effect on accuracy of four machine learning methods - Naive Bayes, Decision Trees, Maximum-Entropy, and K-Means clustering.

A study by (Kechaou, and et al.,2013) entitled " A novel system for video news' sentiment analysis "The spread of different types of multimedia information on the websites has been greatly but research is still in this topic at the beginning, therefore the purpose of their paper is to find a new way for sorting out and classifying various types of news videos and media texts based on sentiment analysis. Where vedio news are defined and categorized into good and bad form through suggested Hidden Markov Model (HMM) and Support Vector Machine (SVM) hybrid learning method, This proven that the feature-selection-combining method, encompassing the Information Gain (IG), Mutual Information (MI) and CHI-statistic (CHI), performs the best classification .

Chapter 3

Research Methodology

Introduction

This chapter presents a detailed description of the research methodology , description of the system built representing it through algorithms and flowcharts for this thesis, the sentiment includes emotions, opinions, and speculations, among others.

Sentiment have unique characteristics that and distuinctive them from other attributes, for example sentiment classification usually deals with two classes good or bad sentiment (positive vs. negative), a range of polarity (e.g. star ratings for movies), or even a range in strength of opinion.

Sentiment Analysis has many names, it's often referred to subjectivity analysis, opinion mining, and appraisal extraction, with some connections to affective computing (computer recognition and expression of emotion) since sentiment and opinion often refer to the same idea. (Mejova, 2009)

Sentiment analysis is an important topic that helps companies, filmmakers, and others to know what people feel about their product, by analyzing their comments to know if the product is good or bad (positive or negative)? Each topic differs from others in expressed way, for example, who speaks and expresses about politics and wars different when speaks and expresses about Sports and Education. It is important to create a special lexicon for each topic.

The use of social media sites, websites in the Arab world and the follow comments for movies become widely spread, it became important to sentiment analysis about Arabic reviews for movies.

This helps filmmakers to know what people feel about their movies and also benefit people to know if the movie is good or bad before watching them so as to not waste time, effort and money and also benefit everyone who is interested in movies.

In this thesis, we built a system to analyse the Arabic reviews on the movies domain for positive, negative or neutral sentiment , and build a lexicon for Arabic reviews on the movie domain .

Movie Rating for Arabic Reviews System (MRARS)

The researcher have built a system of sentiment analysis for Arabic reviews on movie domain (MRARS) Movie Rating for Arabic Reviews where the researcher have built using program code in Oracle.

Oracle in 1979 was the first company to commercialize a relational database, and the relational software, now called Oracle Database.

Oracle database is a collection of data treated as a unit. The purpose of a database is to store and retrieve related information. A database is the key to solve the problems of information management., and it is the most flexible and cost effective way to manage information and applications.

(Michele , 2005).

Oracle is a database that responds very well with excellent performance in demanding environments. Oracle is a major database which along with its added features passes the ACID test, which is important in insuring the integrity of data. This is very important because data is the heart of any system in organization.

(<http://www.learn.geekinterview.com/database/oracle/advantages-of-using-oracle.html>)

We built a lexicon, where it contains all the feelings words about movies divided into three main sections as follows:

The first section containing the positive words like (جيد ، سعيد).

The second section containing the negative words like (سيء ، مزعج).

The third section containing the neutral words like (مقبول ، مناسب).

The system has two methods for reading sentences, the first method is to write the sentence through the interface and the second-way is to read the sentences stored in a file.

When writing or inserting a sentence into the system, it split the sentence into words and matching the word with the words in the lexicon.

The matching process consists of two main stages:

The first stage is based on the full words of the sentence. If all the words of sentence are completely matching the words in the lexicon, then consider them as full positive, full negative and full neutral.

The second stage is to be based on the some words of the sentence. If some the words of sentence are matching the words in the lexicon, then consider them as positive, negative and neutral sentiment.

If all or some words of sentence do not match the words in the lexicon, we consider them not available.

After completing the matching process of all the sentences, the system calculates the number, average, and percentage of all reviews for each movie (full positive , positive) , (full negative, negative) and (full neutral, neutral) as shown in equation 1,2,3,4,5,6 .

$$(average\ positive\ rating) = \frac{num\ of\ full\ positive\ sentences + Positive\ sentences}{num\ of\ all\ sentences} \dots\dots\dots eq(1)$$

$$(Percentage\ positive\ rating) = average\ positive\ rating * 100 \dots\dots\dots eq(2)$$

$$(average\ negative\ rating) = \frac{num\ of\ full\ negative\ sentences + negative\ sentences}{num\ of\ all\ sentences} \dots\dots\dots eq(3)$$

$$(Percentage\ negative\ rating) = average\ negative\ rating * 100 \dots\dots\dots eq(4)$$

$$(average\ neutral\ rating) = \frac{num\ of\ full\ neutral\ sentences + neutral\ sentences}{num\ of\ all\ sentences} \dots\dots\dots eq(5)$$

$$(Percentage\ neutral\ rating) = average\ neutral\ rating * 100 \dots\dots\dots eq(6)$$

then the largest Percentage of these will be the final results of the movie reviews rating.

Finally, the system shows the result of all sentences (reviews) of each movie into positive or negative or neutral sentiment and calculates the number, average, and percentage of all reviews for each movie.

Procedures and Methodologies

The main idea of this thesis is to construct an algorithm for Sentiment Analysis for Arabic Reviews on the Movies Domain, the following steps explain how our system works:

- 1-Building sentiment lexicon based on movie reviews.
- 2- Collecting sentences online about movie reviews.
- 3-Translating English reviews to Arabic Language.
- 4- Creating a code to split the sentence into words.
- 5- Matching Process.
- 6- Assigning label to the sentences.
- 7- Checking the accuracy.

Figure (2), shows the flowchart of the research steps of our work

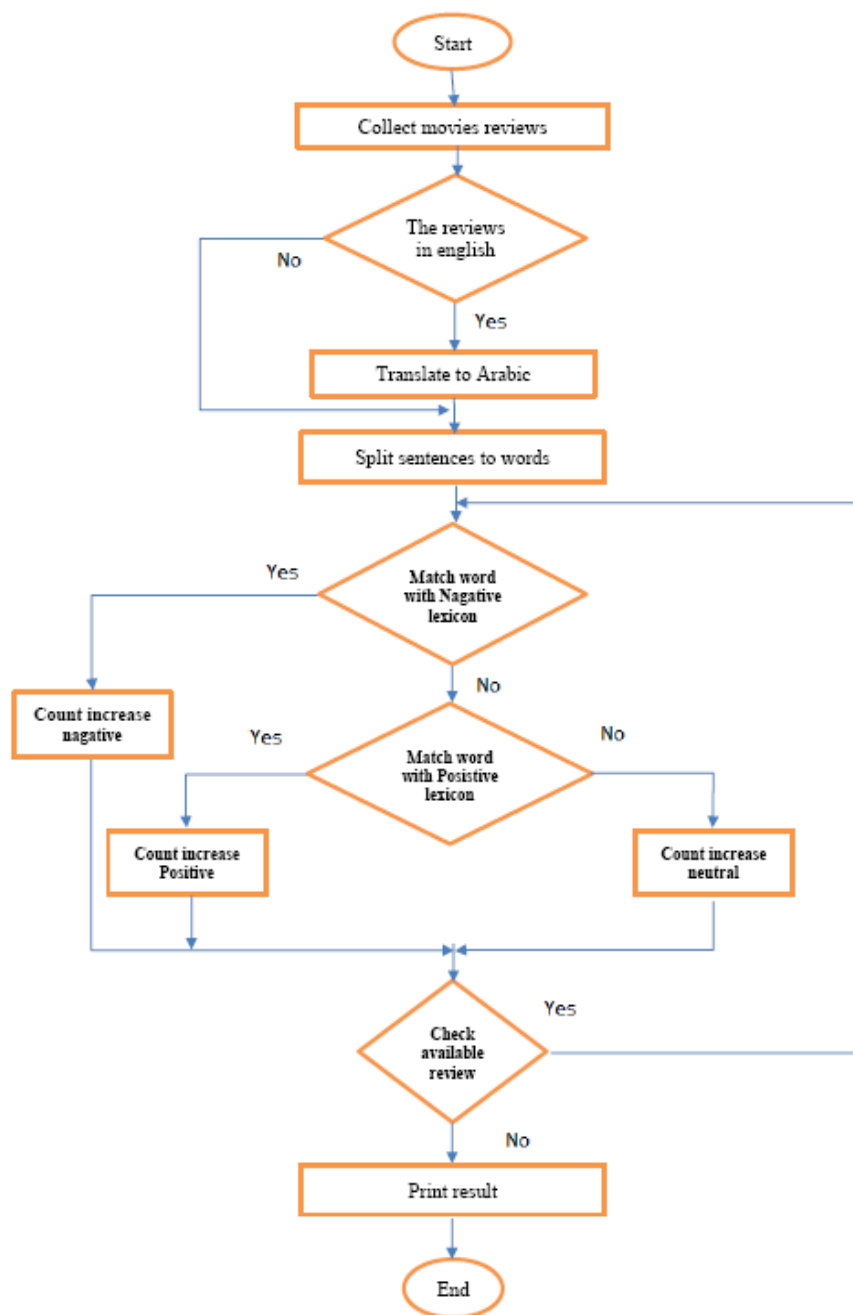


Figure (2): Flowchart of the research steps

Building Sentiment Lexicon

In the beginning, to build a lexicon for Arabic reviews on movie domain we should collect many Arabic reviews as possible about movies domain, we used Data Miner Software to collect the reviews from websites, we collected more than 1551 reviews (Arabic reviews and English reviews which is translated into to the Arabic language) about movies domain, we collected all reviews from many Arabic, English, and Hindi famous movie websites, we asked many people about what famous movie websites and most commonly used.

Then the reacher split the sentence into words and focus on the words that contain feelings to facilitate the construction of the lexicon and lable it to positive, negative and natura using the star system.

The most Web site the reasearcher extracted reviews have a star system, the star system Consists of five stars the reviews that contain five or four stars are positive reviews, if the reviews contain three stars, the reviews are neutral and if the reviews contain one or two stars, the reviews are considered negative.

The lexicon contains words with their polarity label or a number reflecting how much the word expresses each polarity/emotion. The polarity/emotion label can be assigned to the sentence according to the majority score. For example, if the majority of the words were positive or the positive total was higher than the negative total, the sentence will be labeled as positive.

(Heerschop and et al., 2011) (Altrabsheh, 2016)

Figure (3),(4),(5) show the lexicon in the database for our work.

| S. | DESCS | FLAG |
|-----|----------|------|
| 611 | عاطفي | |
| 612 | ممتع | |
| 613 | جميل | |
| 614 | روعه | |
| 615 | جيد | |
| 616 | احب | |
| 617 | اعشق | |
| 618 | قوي | |
| 619 | رائع | |
| 620 | يضحك | |
| 621 | نهج جديد | |
| 622 | مفرح | |
| 623 | ممتاز | |
| 624 | مهم جدا | |

Figure (3): Positive lexicon words example

| SEQ | DESCS | FLAG |
|-----|-------------|------|
| 399 | لايعجيني | |
| 400 | مخيب للامال | |
| 401 | فاشل | |
| 402 | احيطني | |
| 403 | باتس | |
| 404 | هامل | |
| 405 | مضيعه للوقت | |
| 406 | سيئ | |
| 407 | مرفوض | |
| 408 | مكروه | |
| 409 | ممل | |
| 410 | مروعا | |

Figure (4): Negative lexicon words example

| SEQ | DESCS | FLAG |
|-----|----------------|------|
| 134 | مناسب | |
| 135 | مقبول | |
| 136 | متردد لمشاهدته | |
| 137 | ودي | |
| 138 | ممکن إدراکه | |
| 139 | مناسب | |
| 140 | أطمح للاكثر | |
| 141 | متوازن | |
| 142 | مرغوب | |
| 143 | محير | |
| 144 | سطحي | |

Figure (5): Neutral lexicon words example

We collected about 1250 (positive , negative , neutral) words as shown in Table (1) :

Table (1): Number of Lexicon Words

| Data source | Number of words |
|----------------|-----------------|
| Positive words | 655 |
| Negative words | 441 |
| Neutral words | 154 |

The number of words mentioned in the table that represents the three sentiment cases were calculated from the collective reviews and these words are summary without redundancy, also the number of reviews were determined after ignoring the repetition.

Figure (6), shows sentiment lexicon model flowchart and the summary of the steps to build sentiment lexicon, which we explained already in detail.

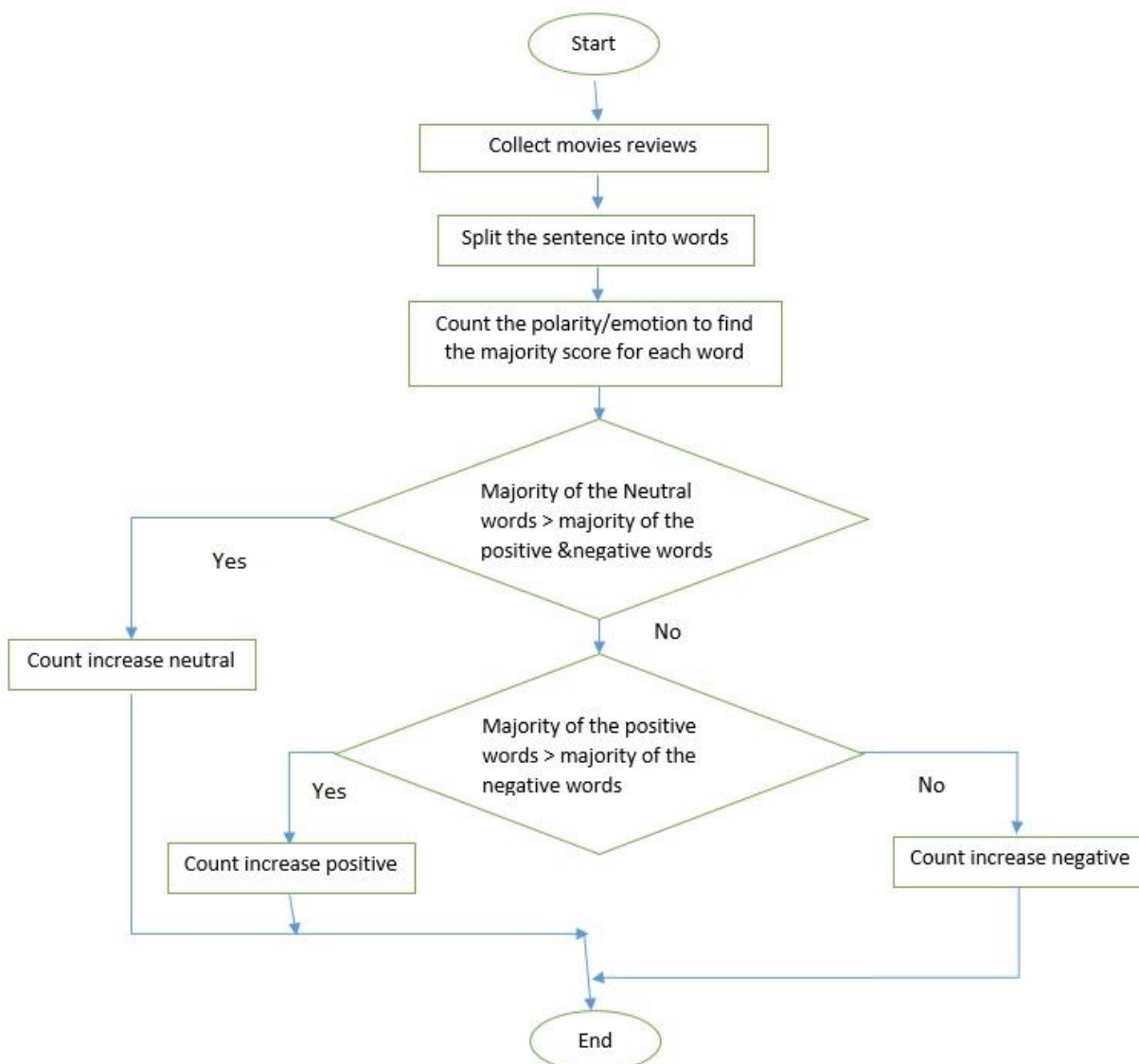


Figure (6): Sentiment lexicon model flowchart

Collecting Sentences Online

Different websites are used for data collection. The researcher has chosen the most popular movie sites. In the beginning, we have collected reviews from popular local sites in Jordan such as (Grand Cinemas Jordan : facebook) , (Cinemas in Jordan: facebook), (Cinemas in Jordan: twitter) and collect reviews from the most famous sites in the Arab world such as (www.elcinema.com) , (www.dardarkom.com) (موفيز لاند : Facebook) , ([Iraqi Movies channel in YouTube](https://www.youtube.com/channel/UC...)) .

Also the researcher has collected reviews from popular sites in the world such as (www.imdb.com) , (www.rottentomatoes.com) , (www.fandango.com), and from the time of India site (timesofindia.indiatimes.com).

Translating English Reviews into Arabic Language

During collection of reviews specialty from the world famous movie sites most reviews were in the English language, We translated these reviews into Arabic language.

the researcher has some difficulty in translating the reviews into Arabic because the reviews related to the person's feelings about the movie. Some reviews were translated by a specialized translator person and some others by the ACE Translate program.

ACE Translate program is a special translation software for languages, it has the ability to translate 59 languages.it has the capability to copy and paste sentences from any source with any language. The software also claims to work with Microsoft Outlook and email systems. (<http://www.toptenreviews.com/business/software/best-translation-software/ace-translator-review/>)

Splitting the Sentence into Words

the researcher created a code in oracle PL/SQL in order to split the sentence into words, to match these words with the words in the lexicon as shown in Figure (7) .

```

declare
    vlength    number;
    sub_space  varchar2(1);
    vfname     varchar2(10000) := '';
    x          varchar2(8000);
    v_word     number(10) := 0;
    y          number(20);
    counter    number(3);
    fname      varchar2(10000);

    w_ser      number := 0;
begin
    for rec in (select seq,full_aname
                from wfname
                order by seq)
loop
    counter := 1;
    begin
        select nvl(length(full_aname) + 1 ,0)
        into vlength
        from wfname
        where seq = rec.seq;
    exception when no_data_found then
        vlength := 0;
    end;

```

```

if vlength = 0 then
  null;
else
  x := rec.full_aname ;
  v_word := 0;
  y := 0;

  for i in 1..vlength
  loop
    y:= y+1;

    select substr(x,y,1)
    into sub_space
    from wfname
    where seq = rec.seq;

    if sub_space <> ' ' then
      vfname := vfname||sub_space;
    else

      if counter = 1 then
        fname := vfname;
        counter := counter + 1;

        goto inst;
      end if;

      v_word := v_word + 1;
      x:= substr(rec.full_aname || ' ',instr(rec.full_aname || ' ', ' ',1,v_word ) + 1);
      vfname := '';
      y :=0;
    end if;
  end loop;
end loop;

```

Figure (7): The programing code for split the sentence into words

Matching Process

In our system, we put the lexicon that we built in the database and classified the word with the lexicon into positive, negative and neutral.

The matching process consists of two main stages:

The first stage is:

If all the words of sentence match the words in the lexicon and refer to negative, we consider them to be full negative.

If all the words of sentence match the words in the lexicon and refer to positive, we consider them to be full positive.

If all the words of sentence match the words in the lexicon and refer to neutral, we consider them to be full neutral.

This stage makes us sure to know what kind of sentiment about this sentence (positive or negative or neutral).

The second stage is:

If some the words of sentence match the words in the lexicon and refer to negative, we consider them to be negative.

If some the words of sentence match the words in the lexicon and refer to positive, we consider them to be positive.

If some the words of sentence match the words in the lexicon and refer to neutral, we consider them to be full neutral.

If the word of sentence does not match the words in the lexicon then the result is not available.

This result helps us to increase the size of lexicon in the future and know what new words we should to enter into the lexicon.

Figure (8) and (9) show the first stage and the second stage of the matching process.

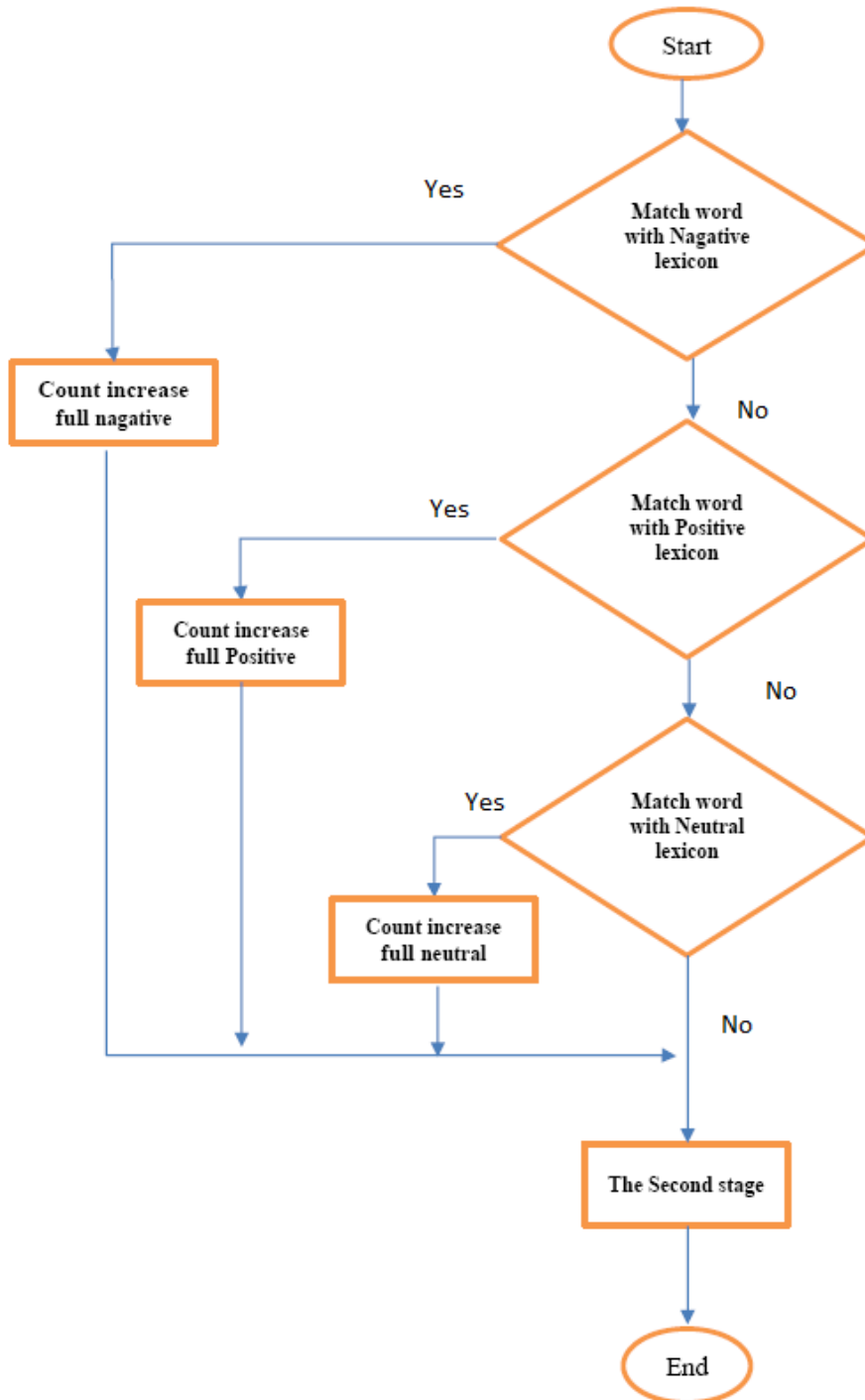


Figure (8): The First Stage for matching process

The flowchart in figure (8) was implemented through the programming code shown in figure (9)

```

PROCEDURE a IS
  x  varchar2(10);
begin
  update wfname
  set flag = null;

  forms_ddl('commit');

  for rec1 in (select * from wfname order by seq)
  loop

    for rec2 in (select * from neg order by seq)
    loop
      if rec1.full_aname = rec2.descs then
        update wfname
        set flag = 'Tneg'
        where seq = rec1.seq;
        forms_ddl('commit');
      end if;
    end loop;

    for rec3 in (select * from pos order by seq)
    loop

      if rec1.full_aname = rec3.descs then
        update wfname
        set flag = 'Tpos'
        where seq = rec1.seq
        and flag is null;
        forms_ddl('commit');
      end if;
    end loop;
  end loop;
end;

```

Figure (9): The programming code for the First Stage for matching process

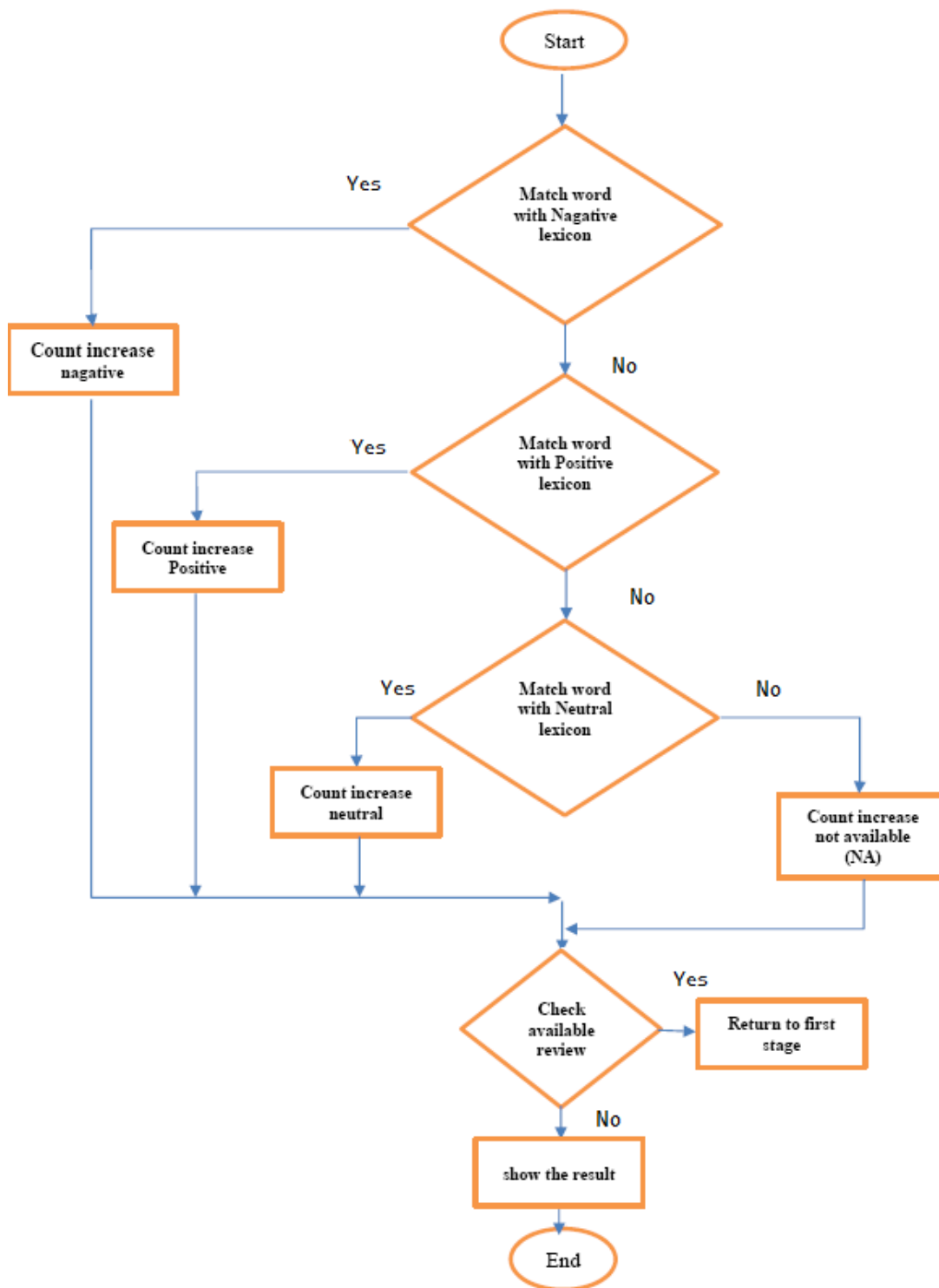


Figure (10): The Second Stage for matching process

The flowchart in figure (10) was implemented through the programming code shown in figure (11)

```

PROCEDURE b IS
  x  number;
begin

for rec1 in (select * from wfname order by seq)
loop

  if rec1.flag is null then
for rec2 in (select * from neg order by seq)
loop
  if rec1.full_aname like '%'||replace(rec2.descs, ' ', '%') then
    update wfname
    set flag = 'neg'
    where seq = rec1.seq
    and flag is null;
    forms_ddl('commit');
  end if;
end loop;

for rec3 in (select * from pos order by seq)
loop
  if rec1.full_aname like '%'||replace(rec3.descs, ' ', '%')
  update wfname
  set flag = 'pos'
  where seq = rec1.seq
  and flag is null;
  forms_ddl('commit');
  end if;
end loop;

  for rec3 in (select * from neut order by seq)
  loop
  if rec1.full_aname like '%'||replace(rec3.descs, ' ', '%') then
    update wfname
    set flag = 'neut'
    where seq = rec1.seq
    and flag is null;
    forms_ddl('commit');
  end if;
  end loop;

  update wfname
  set flag = 'NA'
  where seq = rec1.seq
  and flag is null;
  forms_ddl('commit');

  else
    null;
  end if;

  end loop;
  x := show_alert('msg');
end;

```

Figure (11): The programming code for The Second Stage for matching process

Assign Label to the Sentences

After we had matched all the words of reviews with the lexicon that was previously built to know the type of sentiment if they are positive or negative or neutral.

If the number of words used in the sentence indicates a positive, we considered the review is positive but If the number of words used is negative, the review is negative otherwise the review will be neutral.

After selecting the sentiment of most severe words which specified the sentiment of review we give the appropriate label for reviews.

After we had made the label of all reviews we calculate the number of reviews containing the same label and calculate the average and percentage of them to know the mostly label type of reviews.

Check the Accuracy

The system was checked manually by selecting all reviews within the movie, for example, the total of 100 reviews per movie, we checked them manually .

Chapter four

Experimntal and Result

Introduction

In order to simplify the method and the system (Movie rating for Arabic reviews), in this chapter, a detail description of the whole system with all screens will be discussed.

Examples of the execution of the frame work are represented and the results of the system are shown.

Movie Rating for Arabic Reviews System

A framework is designed using Oracle Forms for the sentiment analysis for Arabic reviews on the movie domain and this framework is implemented using PL/SQL.

The lexicon was built using the standard database “SQL” implemented by TOAD software.

TOAD (tool for Oracle Application Developers) is a software application from Quest Software used for development, database administrators and data analysts use to manage both relational and non-relational databases using SQL. (Jim McDaniel, (2002))

Arabic reviews for movies were obtained from many famous websites.

Frame Works Movie Rating for Arabic Reviews System

All the screens display and the characteristics of each screen with message alarm of our system are explained below.

Main Menu Screen

When running the system, a main menu screen shown in figure (12) will be displayed. The main menu screen contains main features for the system (Execute, Read file, Add comment, Add rule, About)

Read file: This feature adds and reads files (text file) containing the reviews messages of movies.

Add comment: In this feature, we can write reviews of movies.

Add rules: In this feature, we can add words to the lexicon.

About: A brief description of the system.

Execute: This feature analyzes the reviews of movies and show the results.

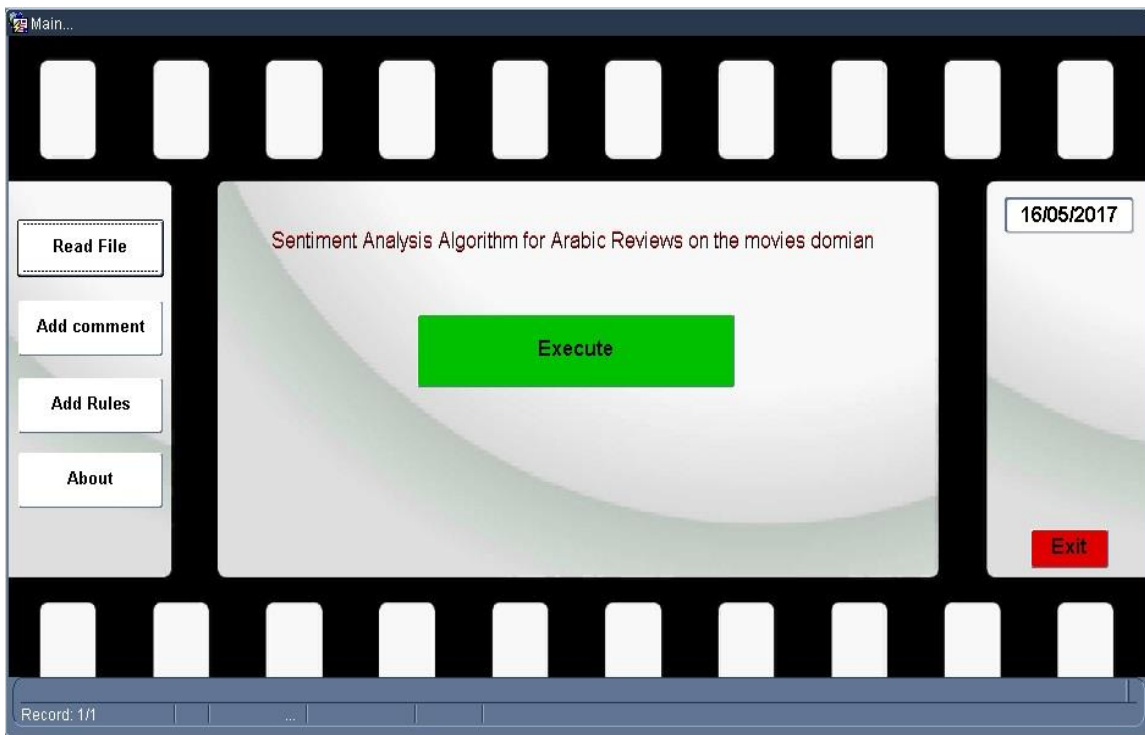


Figure (12): Main menu screen

Read File Menu Screen

When clicking on the button (read file) in the main screen a read file menu screen shown in figure (13) will be displayed, which allow us to add any file containing movie reviews. we click on the browse button lable 1 to choose the file from PC the path for file will show on lable 2 or we can choose the file by write the path on lable 2 and we click on the add file button lable 3 to add the file to the database system and then click on the back button lable 4 to return to the main menu screen to do the execute.

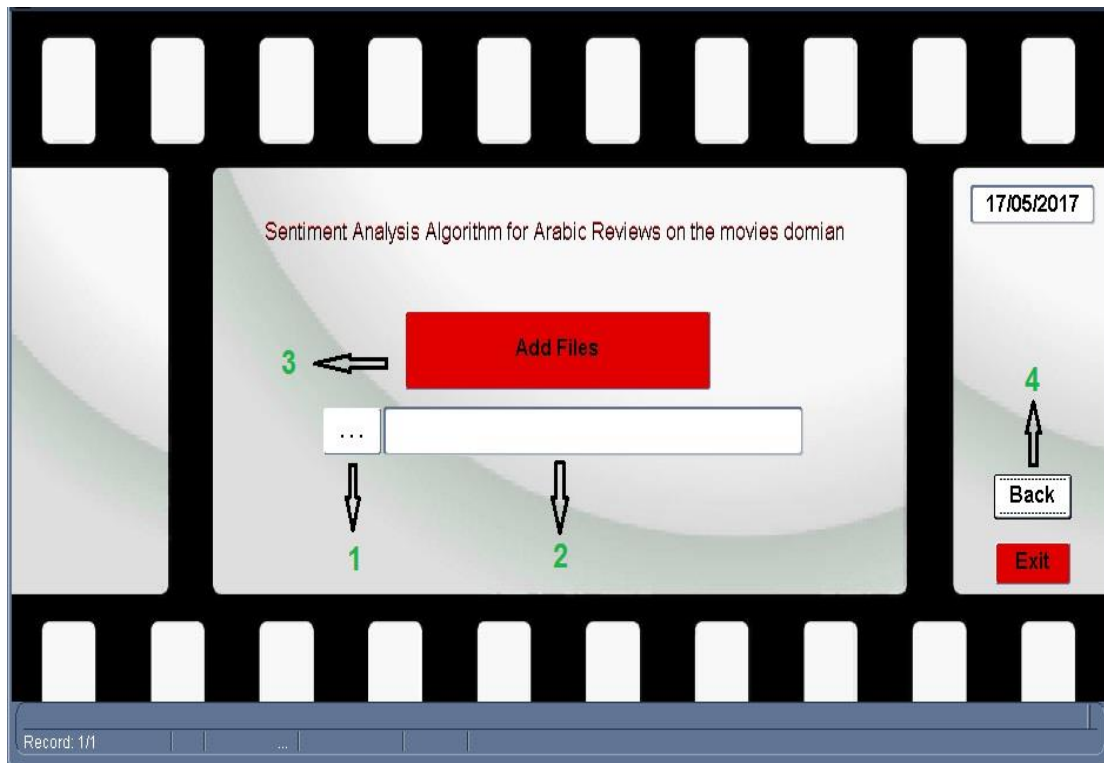


Figure (13): Read file menu screen

Add Comment Menu Screen

When clicking on the button (Add comment) in the main screen add comment menu screen shown in figure (14) will be displayed; We write any comment (review) in label 1 and then click on the add comment button label 2 the result will show directly for analysis review (full positive, positive, full negative, negative, full neutral, neutral) in label flag.

We can save the comment (review) in the system's database by just clicking on the save button label 3 (the sequence number will be immediately corrupted for the comment on the database) and then clicking on the back button label 4 to execute it with the other comments in the database.

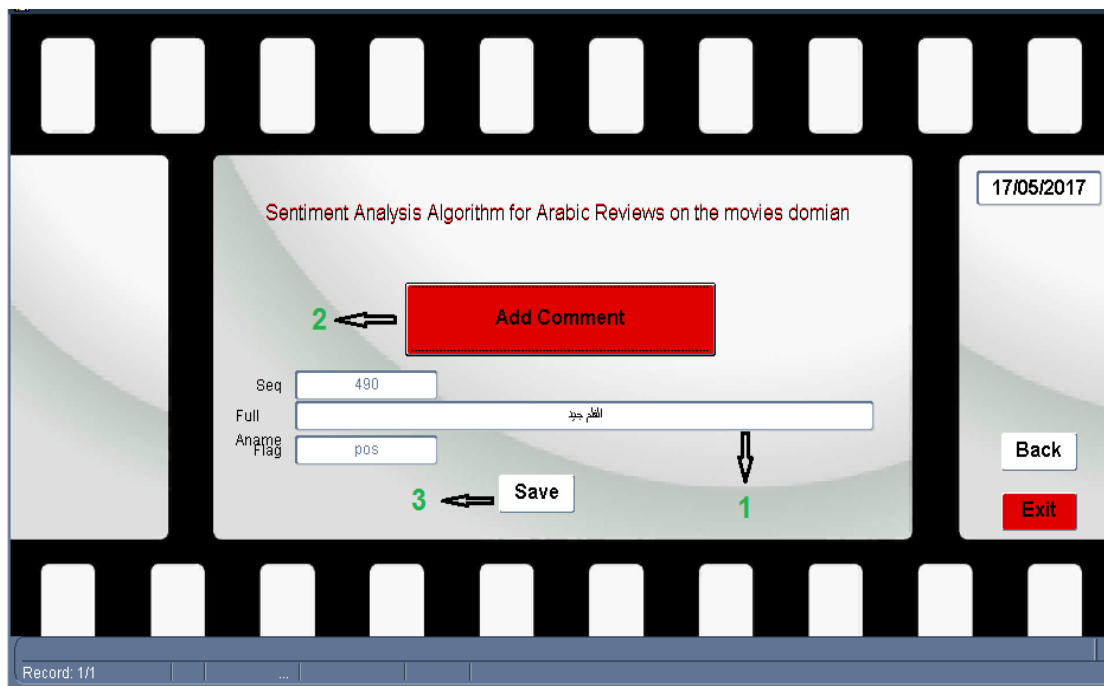


Figure (14): Add comment menu screen

Add Rule Menu Screen

When clicking on the button (Add Rules) in the main screen a add rule menu screen shown in figure (15) will be displayed, we can add words to the lexicon, First, we choose positive or negative or neutral words.

We can also display all the words in the lexicon by clicking on the show button.

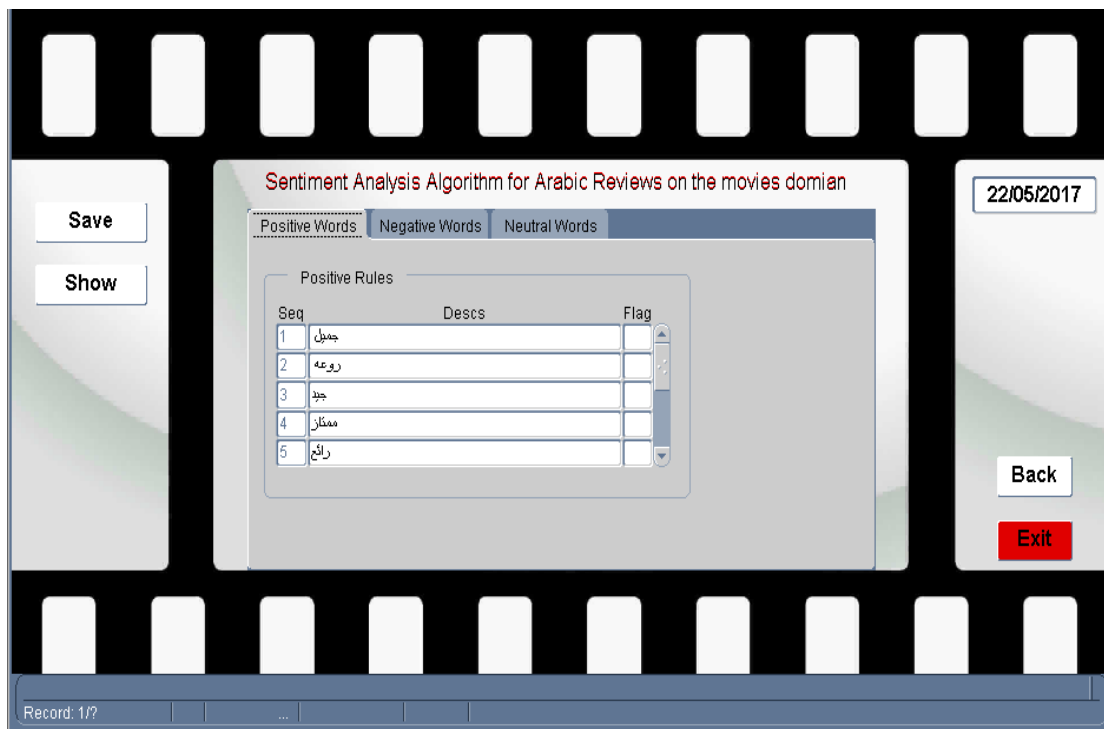


Figure (15): Add rules menu screen

About Menu Screen

When clicking on the (About) button in the main screen about menu screen shown in figure (16) will be displayed, in this screen a brief description of the system will be displayed.

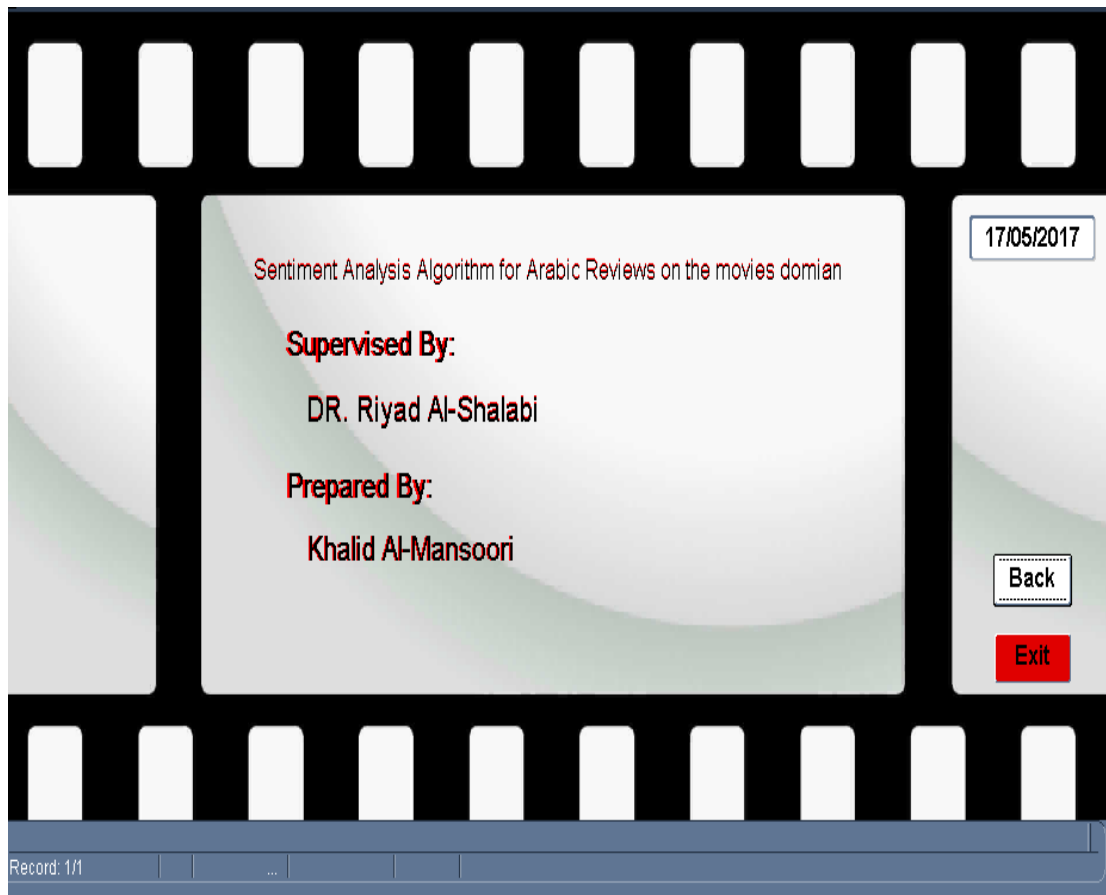


Figure (16): About menu screen

Execute Menu Screen

When clicking on the button (execute) in the main screen execute menu screen shown in figure (17) will be displayed; When we complete adding the file that contains movie reviews or writing a set of comments and saving them in the database, in this screen we click on the execute button table 1, the results will be displayed in a report.

The report covers all the comments, the result for each comment, average, percentage and the final rate for a movie (positive, negative, neutral).



Figure (17): Execute menu screen

4.3.7. Message Alarm Screens

4.3.7.1 Read File Alarm Screen

When we don't choose any file or don't insert the path for file and click on the add file button table 3 the alert message "the path cannot be null" shown in figure (18) will be displayed.

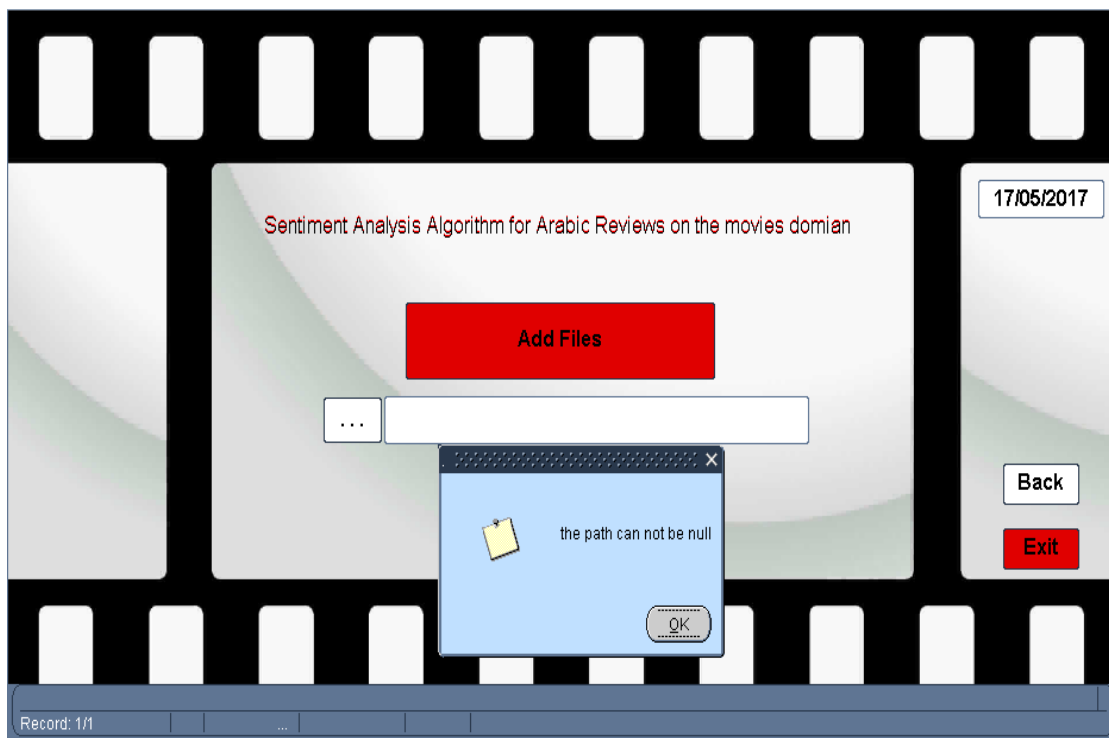


Figure (18): Alert message the path cannot be null

This message alert (finish loading file) will be shown when we insert the file or input the true location path and save into database, screen shown in figure (19) will be displayed.

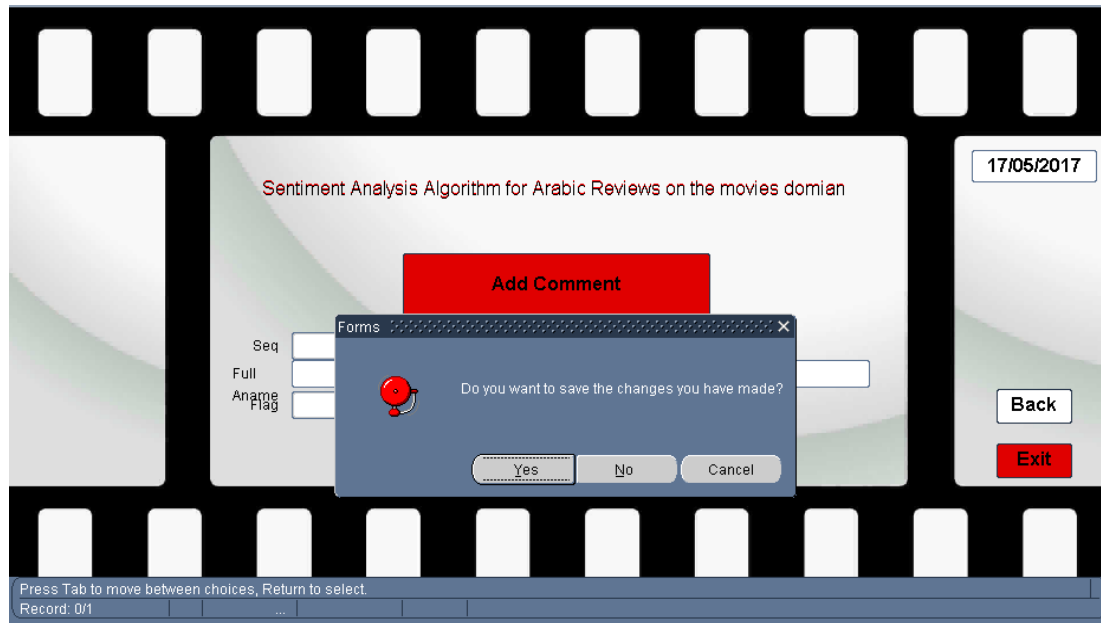


Figure (19): Finish loading file menu screen

4.3.7.2 Add comment alarm screen

This message alert (do you want to save the changes you have made) will show when we don't save comment and click on the back button this alert message in Figure (20) will be displayed.

Figure (20): Alert message save change



4.3.7.3 Execute alarm screen

When this alert appears (finish), it detects that the executing process is completed and there is no any problem in the executing process; The alert message in Figure (21) will be displayed..



Figure (21): Finish alert message

Examples for movie rating for Arabic reviews system

We created three examples of the most famous movies and enter the reviews for the movies in the system and display the results and ratios (positive, negative, neutral) and differences rating between each movie.

4.4.1 Example 1 - sentiment analysis for Arabic reviews of the boss baby 2017 movie

The “boss baby 2017” movie being the most famous movie and most watched of the year 2017.

The researcher has collected many reviews about (the boss baby movie) and we have chosen 102 reviews that contain different sentiment (positive, negative, neutral), where each review shows the result next to it.

Positive = pos

Full positive = Tpos

Negative = neg

Full negative = Tneg

Neutral = neut

Full neutral = Tneut

shows the Procedures steps for sentiment analysis of the boss baby 2017 movie :

Step (1) Collect sentences online about movie reviews:

We chose two sentences :

فلم مرح

هذا الفلم كان رائعا

Step (2) split the sentence into words:

| | | |
|---|-----|-----|
| 1 | مرح | فلم |
|---|-----|-----|

| | | | | |
|---|-------|-----|-------|-----|
| 2 | رائعا | كان | الفلم | هذا |
|---|-------|-----|-------|-----|

Step (3) Matching Process:

Matching these words with the lexicon into positive, negative and neutral, the Figure (22) (23),, shows the matching process

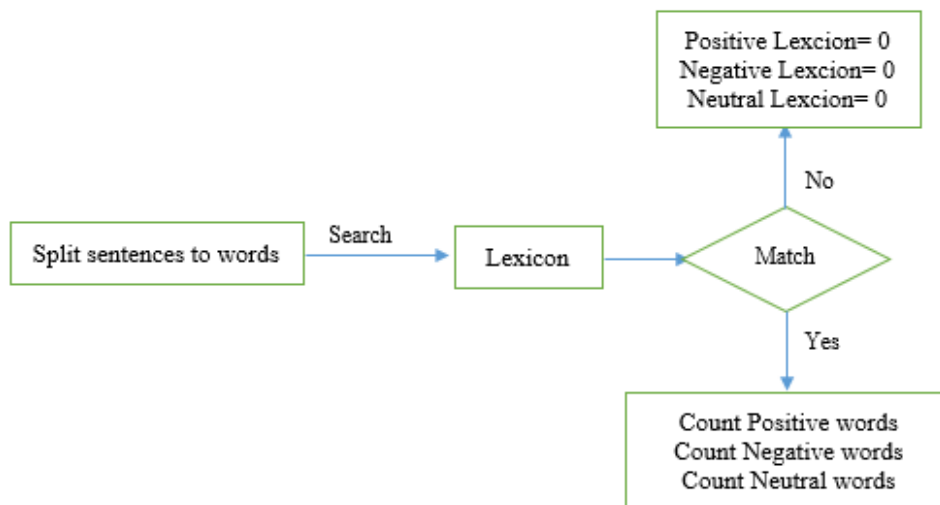


Figure (22): the matching process

Table (2) Examples for count words in lexicon

| Word | Positive Lexicon | Negative Lexicon | Neutral Lexicon |
|-------|---------------------|---------------------|--------------------|
| مرح | 1 | 0 | 0 |
| رائعا | 1 | 0 | 0 |

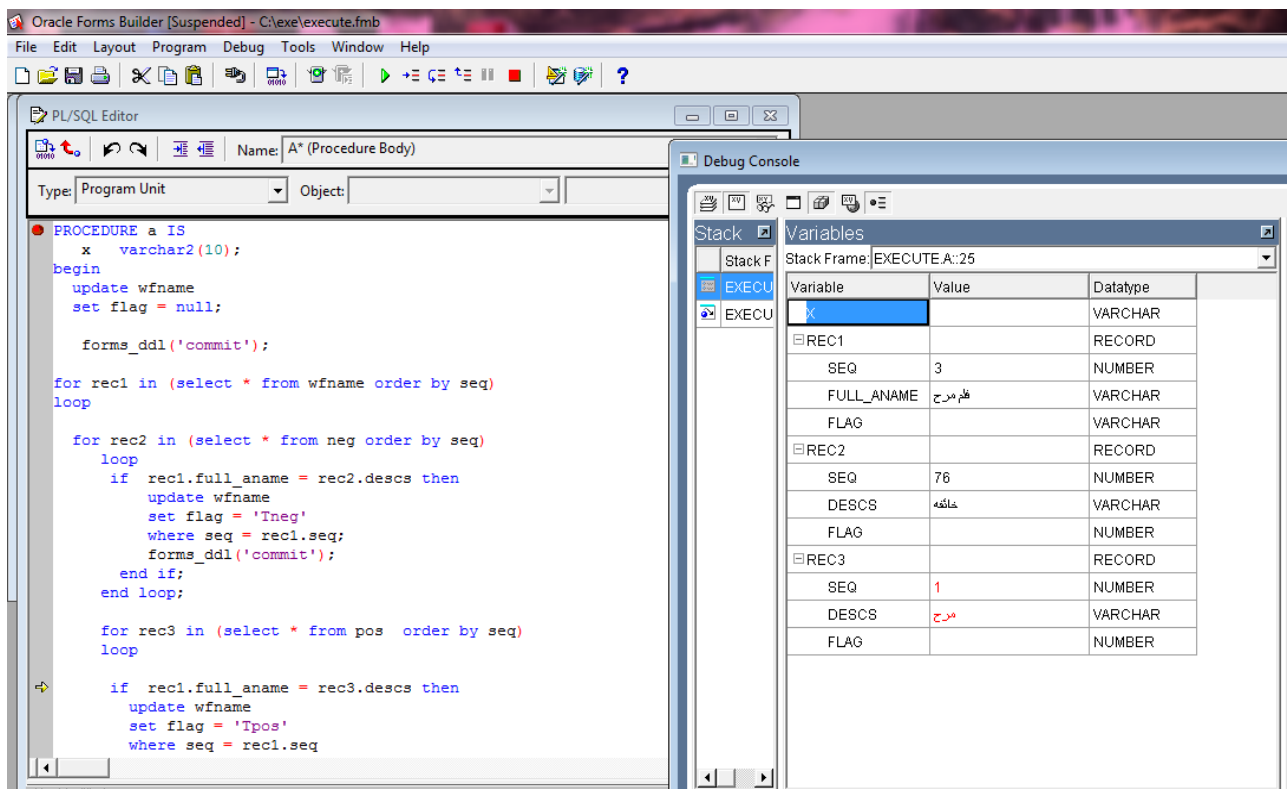


Figure (23): The matching process in oracle form

Figure (24) shows the result execute for the boss baby movie - page1, figure (25) shows result execute for the boss baby movie - page2 and figure (26) shows result execute for the boss baby movie - page3, As shown below:

| Seq | Full Aname | Flag |
|-----|--|------|
| 3 | فلم مرح | pos |
| 4 | قصه القلم سيئة | neg |
| 5 | فلم مضحك | pos |
| 6 | احب هذا القلم | pos |
| 7 | فلم عظيم للعائلة بأكملها | pos |
| 8 | القلم حقا جيد | pos |
| 9 | فيلم عاطفي | pos |
| 10 | فلم ممل جدا | neg |
| 11 | كان مروعا | neg |
| 12 | جدا مضحك | pos |
| 13 | اتصح جميع العائلة بان يشاهدوه | pos |
| 14 | ممتع جدا | Tpos |
| 15 | لا استطيع التوقف عن الضحك | pos |
| 16 | فلم مزعج | neg |
| 17 | هذا القلم اكثر مما توقعت | pos |
| 18 | فلم ممتع | pos |
| 19 | هذا القلم من أطرف الأفلام التي رأيتها على الإطلاق | pos |
| 20 | قضيئا وقت ممتع | Tpos |
| 21 | أنا متأكد من أن جميع البالغين في السينما كان لديهم وقت أفضل من الأطفال | pos |
| 22 | فلم ممتاز | pos |
| 23 | القلم فيه كثير من الترفيه | pos |
| 24 | فلم عظيم | pos |
| 25 | أفضل بكثير من الاستعراضات | pos |
| 26 | حقا استمتحت بمشاهدة القلم | pos |
| 27 | اكره هكذا أفلام | neg |
| 28 | فلم لطيف | pos |
| 29 | رائع جدا | Tpos |
| 30 | مذهل | Tpos |
| 31 | هذا الفيلم كان رائع تماما | pos |
| 32 | اشاهده وانا بسعادة شامة | pos |
| 33 | فلم يتفوق على غيره من الافلام | pos |
| 34 | فلم لا يقاوم | pos |
| 35 | كل يوم اريد مشاهدته | pos |
| 36 | اريد استرجاع النقود التي دفعنها | neg |
| 37 | متحمس جدا لمشاهدته | pos |
| 38 | اعطيه جائزة نوبل | Tpos |
| 39 | فلم مسلي | pos |
| 40 | هذه القلم مذهل بشكل لا يصدق | pos |
| 41 | فلم مشوق | pos |
| 42 | يمتلك ميزة جميلة | pos |
| 43 | فيه تجاوزات كثيرة | neg |
| 44 | فلم واسع الخيال | pos |
| 45 | لقد ظهر محبوب الجماهير | pos |
| 46 | مع جزيل الشكر لمن انتج القلم | pos |
| 47 | عقري من ينتج هكذا قصص | pos |

Figure (24): Result execute for the boss baby movie - page1

| Seq | Full Aname | Flag |
|-----|--|------|
| 48 | فلم فوضوي | neg |
| 49 | فلمى المفضل | pos |
| 50 | يحمل طرائف كثيرة | pos |
| 51 | فلم خلاق | pos |
| 52 | ماهر جدا من صنع فكرة الفلم | pos |
| 53 | فلم متنوع بافكاره | pos |
| 54 | جدير بالثناء | Tpos |
| 55 | جدا قبيحه قصة الفلم | neg |
| 56 | فلم ساحر | pos |
| 57 | فلم مثير | pos |
| 58 | فلم محبوب | pos |
| 59 | استجعت مع الفلم حين مشاهدته | pos |
| 60 | يحتوي على مغامرات كثيرة | pos |
| 61 | فلم مقبول نوعا ما | neut |
| 62 | الفلم له أفضلية عن باقي الافلام | pos |
| 63 | الفلم يحتوي على افكار دقيقة | pos |
| 64 | اريد ان اشاهده مرة اخرى | pos |
| 65 | احببني الفلم كثير | neg |
| 66 | فلم راقى | pos |
| 67 | لقد استمتعت بكل لحظة وانا اشاهده | pos |
| 68 | اعجبني الفلم | pos |
| 69 | الموسيقى فيه جدا جميلة | pos |
| 70 | يستحق المشاهدة | pos |
| 71 | اطفالي استمتعوا في مشاهدته | pos |
| 72 | فلم ليس له مثيل | pos |
| 73 | قصه جدا جميلة عن باقي قصص الاطفال | pos |
| 74 | فلم يحتوي على كوميديا جميلة جدا | pos |
| 75 | شاهدته وانا لا استطيع التوقف عن الضحك | pos |
| 76 | انصح اي شخص يريد ان يضحك فليشاهد هذا الفلم | pos |
| 1 | فيلمى المفضل لهذا العام | pos |
| 2 | احب هذا الفلم من كل قلبي | pos |
| 77 | فلم سخيف | neg |
| 78 | فلم محير | neut |
| 79 | اوجعني راسي هذا الفلم | neg |
| 80 | فيه عنوائيه شديده | neg |
| 81 | بلا هدف | Tneg |
| 82 | فلم غير ملائم | neg |
| 83 | خدعني اعلان الفلم | neg |
| 84 | فلم محبوب جدا | pos |
| 85 | فلم يجذب الفرحه للاطفال | pos |
| 86 | مرفوض | Tneg |
| 87 | فلم غير منطقي | neg |
| 88 | متردد لمشاهدته | neut |
| 89 | خالقه على اطفالى لا اريد ان يشاهده | neg |
| 90 | الفلم يحتوي على مفاهيم خاطئه | neg |
| 91 | اصابني الاحباط | neg |
| 92 | الفلم فيه خطورة على اطفالى | neg |

Figure (25): Result execute for the boss baby movie-page 2

Figure (26): Result execute for the boss baby movie - page 3

| Seq | Full Aname | Flag |
|-----|--|------|
| 96 | مناسب | neut |
| 97 | حقاً رائع اريد مشاهدته مرة اخرى | pos |
| 98 | مذهل مدهل | pos |
| 99 | اقوى افلام الرسوم المتحركة | pos |
| 100 | التمنى لو انققت المال على شئى اخر | neg |
| 101 | اخذت ابني لمشاهدت الفلم و استمتع كثيرا | pos |
| 102 | فلم اوصي الكثير لمشاهدته | pos |
| 1 | فيلمى المفضل لهذا العام | pos |
| 2 | احب هذا الفلم من كل قبلي | pos |

Figure (27): Result execute for the boss baby movie-page 4

Sentiment Analysis Algorithm for Arabic Reviews on the movies domian

Mavie rating for arabic reviews system

| | |
|-------------------------------------|-----|
| Total Number reviews = | 102 |
| Total Number positive reviews = | 72 |
| Total Number negative reviews = | 25 |
| Total Number neutral reviews = | 5 |
| Total Number not avaiable reviews = | 0 |

| | |
|-------------------------------|--------|
| percentage positive reviews = | 70.588 |
| percentage negative reviews = | 24.51 |
| percentage neutral reviews = | 4.902 |
| percentage NA reviews = | 0 |

This movie is positive

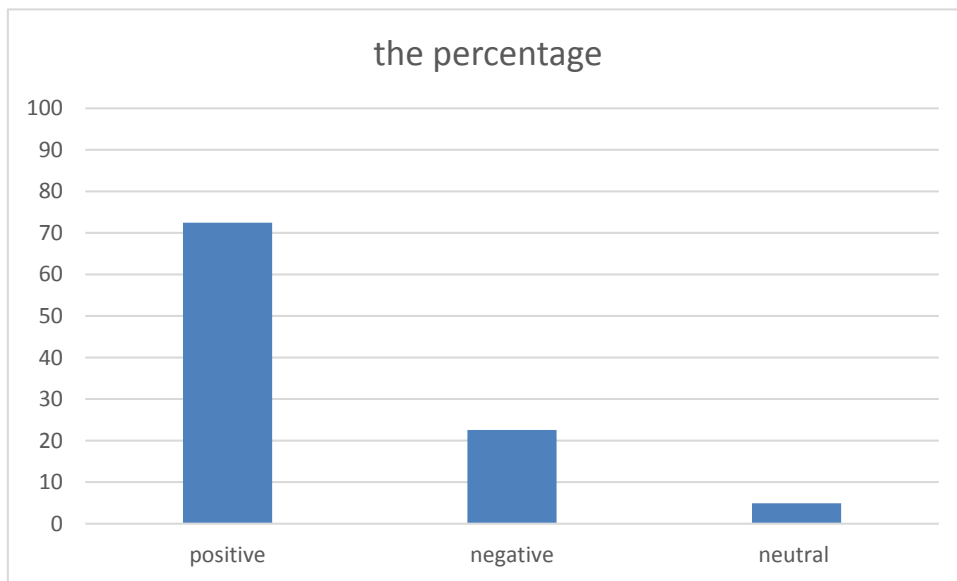


Figure (28): Result percentage for the boss baby movie

Figure (27) result execute for the boss baby movie-page 4 shows the result of sentiments analysis for 102 reviews was mixed with positive, negative and neutral, where a total number of positive reviews is 72, the total number of negative reviews is 25, the total number of neutral reviews is 5 and there is not found any reviews that have no value.

Figure (28) result percentage for the boss baby movie shows the percentage for positive, negative and neutral; we see that the percentage ratio of positive reviews more than the percentage ratio of negative and neutral reviews, finally we consider the overall rating of the movie is positive.

This result helps people to know what the general ratio about this move (positive, negative, neutral) before they watch it, know everyone's feelings about this movie and what advice they provide also this result increases the number of viewers about the movie and make it more famous.

4.4.2 Example 2 - sentiment analysis for Arabic reviews of (THE CIRCLE 2017) movie

The “circle 2017” movie being the most famous and most watched of the year 2017. We have collected many reviews about (the circle) movie and we have chosen 76 reviews that contain different sentiment (positive, negative, neutral), where each review shows the result next to it.

Positive = pos

Full positive = Tpos

Negative = neg

Full negative = Tneg

Neutral = neut

Full neutral = Tneut

Figure (29) result execute for the circle movie - page1, figure (30) and result execute for the circle movie - page2, As shown below:

| Seq | Full Aname | Flag |
|-----|---|------|
| 3 | لاتعجبني هكذا نوع من الافلام | neg |
| 4 | لا احببه على الافلام | neg |
| 5 | فلم جيد | pos |
| 6 | فلم ذو قيمة علميه | pos |
| 7 | فلم سيئ | Tneg |
| 8 | الفلم فاشل | neg |
| 9 | لم افهم قصه الفلم | neg |
| 10 | فلم يائس | neg |
| 11 | نوعا ما جيد | pos |
| 12 | مقبول | neut |
| 13 | فلم هامل | neg |
| 14 | مضيقه للوقت | Tneg |
| 1 | نهايه الفلم لم تعجبني | neg |
| 2 | لايستحق المشاهده | neg |
| 15 | لايوجد فيه اي شعور | neg |
| 16 | نهايه سيئه | neg |
| 17 | فلم لايليق بالتمثيل توم هانكس | neg |
| 18 | فلم رائع | pos |
| 19 | اصبتي الفلم | pos |
| 20 | لا اعلم لماذا هذه التعليقات حول الفلم انه جيد | pos |
| 21 | معروف من الممثل توم هانكس ولكن الفلم سيئ | neg |
| 22 | فلم قوي | pos |
| 23 | لم يعجبني | Tneg |
| 24 | لا اطيق هكذا افلام | neg |
| 25 | اتمنى لو لم اشاهده | neg |
| 26 | جميل بمعنى الكلمه | pos |
| 27 | فلم مفزع | neg |
| 28 | فلم ضعيف | neg |
| 29 | اشق الافلام حول التكنولوجيا | pos |
| 30 | لايطابق الاعلان | neg |
| 31 | مخيب لاملال | Tneg |
| 32 | اسوا فلم شاهدته | neg |
| 33 | فلم يحتوي على نهج جديد | pos |
| 34 | لماذا نلتضرر لهكذا افلام | neg |
| 35 | فلم ضعيف | neg |
| 36 | هذا الفلم يتكلم عن مستقبلنا | pos |
| 37 | كنت اتوقع الافضل | neg |
| 38 | لاتوجد نهايه | Tneg |
| 39 | ليس جيد | Tneg |
| 40 | هل تسخرون منا | neg |
| 41 | يجعلك الفلم تحب التكنولوجيا | pos |
| 42 | فلم ممل | neg |
| 43 | مهم جدا | Tpos |
| 44 | مع الاسف ضيقت وقتي | neg |
| 45 | لا يوجد اي تشويق | neg |
| 46 | احببت الفلم من اول لحضه | pos |
| 47 | تجنيو هذه الافلام | neg |

Figure (29): Result execute for the circle movie-page 1

| Seq | Full Aname | Flag |
|-----|--|------|
| 48 | لا يعجبني الفلم | neg |
| 49 | اصابني النعاس وانا اشاهده | neg |
| 50 | لا اكرت لمثل هذه الافلام | neg |
| 51 | اعتشق هذا الممثل | pos |
| 52 | احب افلام التكنولوجيا | pos |
| 53 | الفلم سيئ الجودة | neg |
| 54 | الفلم جيد | pos |
| 55 | فلم مذهل | pos |
| 56 | الفلم لاصحاب العقول | pos |
| 57 | اتمنى لو حضر ابي معي | pos |
| 58 | ياالله اعتشق افلام التكنولوجيا | pos |
| 59 | القصة جميلة بالفعل | pos |
| 60 | ساضعه في قائمه المفضلات | pos |
| 61 | الفلم محقد | neg |
| 62 | مصطنع جدا | neg |
| 63 | غير ناجح | Tneg |
| 64 | دون المستوى المطلوب وخاصة مع هكذا ممثلين | neg |
| 65 | خدعني اعلان الفلم | neg |
| 66 | مستاء جدا | neg |
| 67 | لا نهاية له | neg |
| 68 | لن اشاهده حتى بدون مقابل | neg |
| 69 | مستقبل رائع لهذا الفلم | pos |
| 70 | عشقت قصة الفلم | pos |
| 71 | الفلم منسوخ من قصة اخرى | neg |
| 72 | فلم يخزي | neg |
| 73 | اكره هذه الافلام | neg |
| 74 | مناسب | neut |
| 75 | الفلم ممكن إدراكه | neut |
| 76 | أطمح للتفوي | pos |

Figure (30): Result execute for the circle movie-page 2

| | |
|--------------------------------------|----|
| Total Number reviews = | 76 |
| Total Number positive reviews = | 26 |
| Total Number negative reviews = | 47 |
| Total Number neutral reviews = | 3 |
| Total Number not available reviews = | 0 |

| | |
|-------------------------------|--------|
| percentage positive reviews = | 34.211 |
| percentage negative reviews = | 61.842 |
| percentage neutral reviews = | 3.947 |
| percentage NA reviews = | 0 |

This movie is negative

Figure (31): Result execute for the circle movie-page 3

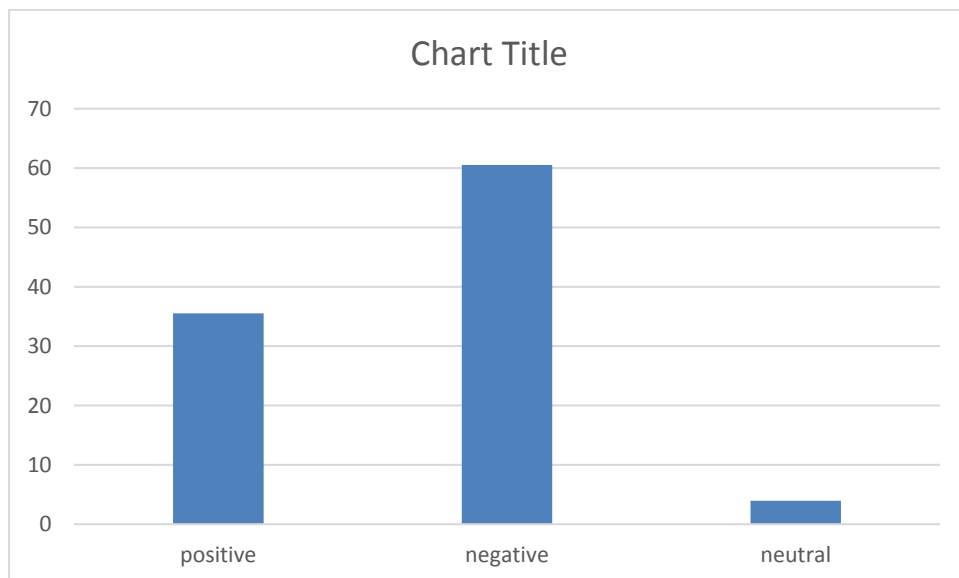


Figure (32): Result percentage for the circle movie

Figure (31) shows result execute for the circle movie shows the result of sentiments analysis for 76 reviews was mixed with positive, negative and neutral, where a total number of positive reviews is 26, the total number of negative reviews is 47, the total number of neutral reviews is 3 and there is not found any reviews that have no value.

Figure (32) shows result percentage for the circle movie shows the percentage for positive, negative and neutral, we see that the percentage ratio of negative reviews more than the percentage ratio of positive and neutral reviews, finally we consider the overall rating of the movie is negative.

This result helps people to know what the general ratio about this move before they watch it, know everyone's feelings about this movie and what advice they provide also this result help the filmmakers to know what people feel about their movie and give him some advice that helps them to make new movies.

4.4.3. Example 3 - sentiment analysis for Arabic reviews of many movies

We have chosen ("MOANA", "الجزيره 2", "The Wall", "عمر وسلمى 3", and "Men Women Children ") movies the most famous and most watched, we have collected many reviews about these movie and calculate the number , percentage of all reviews for each movie and check the accuracy, As shown in table (3) :

Table (3) Examples for many movies

| No | Movie Name | Num of reviews | Number | | | Percentage | | | Final Sentiment result |
|----|--------------------|----------------|--------|-----|------|------------|-------|-------|------------------------|
| | | | Pos | Neg | Neut | Pos | Neg | Neut | |
| 1 | MOANA | 200 | 147 | 35 | 18 | 73.5 | 17.5 | 9.0 | Positive |
| 2 | الجزيره ٢ | 110 | 75 | 26 | 9 | 68.18 | 23.63 | 8.18 | Positive |
| 3 | The Wall | 60 | 12 | 42 | 6 | 20.0 | 70.0 | 10.0 | Negative |
| 4 | عمر وسلمى ٣ | 140 | 114 | 26 | 0 | 81.42 | 18.57 | 0.0 | Positive |
| 5 | Men Women Children | 110 | 21 | 63 | 26 | 19.09 | 57.27 | 23.63 | Negative |

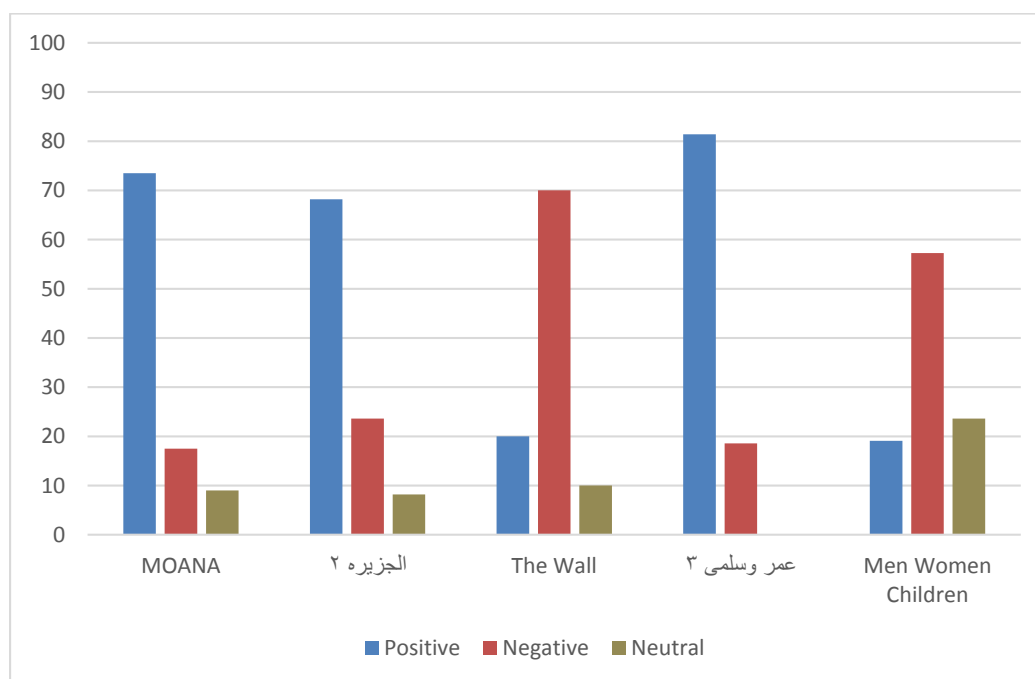


Figure (33): Example 3 result percentage

Table (3) shows the results of all reviews for each movie and the number and percentage of positive, negative and neutral sentiment for each movie.

MOANA movie is an animation movie; the story about a girl in Ancient Polynesia. The results showed that the people reviews were good and they were happy to watch the movie and some of the reviews were neutral, although there were some bad reviews about the movie but the overall sentiment was good.

Al-Jazera 2 (الجزيره 2) Movie is an action movie. The story about Mansur el-Hafni (Ahmed El Sakka Actor) who escape from prison, meets his brother and son, and returns to the island; The reviews display that people were happy to watch the movie and the Actor Ahmed El Sakka was so wonderful , although there are bad reviews, most of them show they did not like the movie because it overstepped the government but the overall felling was good.

The Wall Movie is an action movie the story about two American Soldiers are trapped by a fatal sniper, most people were unhappy to watch the movie and very few liked it, the overall Opinion of the movie was bad.

Omar & Salma3 (عمر وسلمى 3), is a movie tells the story of Omar life. the was lost and reckless young man who has no clear goal in life, and then his life changed and becomes meaningful when seeing Salma, The movie was a great success as a lot of reviews were very good and few people did not like the movie.the general character was good.

Men, Women, Children Movie the story about the filmmakers are trying to prove the negative impact of Internet and social media sites and their destruction of humanity, ignoring all the positives that the Internet has made since their emergence, the reviews showed that people's impression was very bad, although there are neutral and positive reviews but the general impact was bad.

After collecting reviews about movies, as shown table (3) and sentiment analysis of these reviews we checked the accuracy of these reviews, the accuracy process was checked manually, the results of the accuracy verification were uneven for each movie reviews as shown in Table (4) :

Table (4) The accuracy results

| No | Movie Name | Num of reviews | Accuracy |
|----|--------------------|----------------|----------|
| 1 | MOANA | 200 | 94% |
| 2 | الجزيره ٢ | 110 | 96% |
| 3 | The Wall | 60 | 100% |
| 4 | عمر وسلمى ٣ | 140 | 96% |
| 5 | Men Women Children | 110 | 98% |

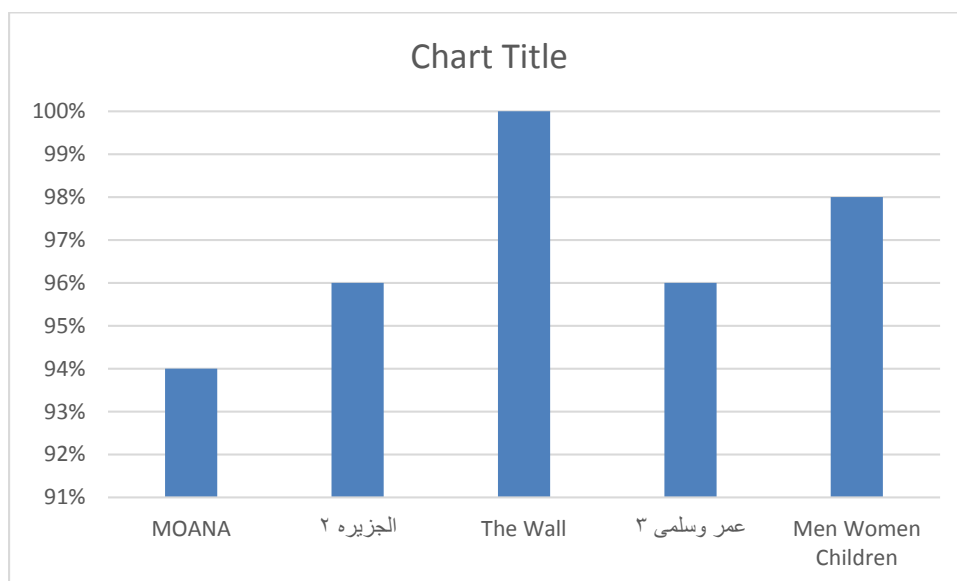


Figure (34): The accuracy results

The accuracy process showing when the fewer reviews (the number of words) are, the more accurate the result are. on other, large number of reviews results in accuracy decrease.

In the future, we can increase the accuracy by increasing the number of words in the lexicon.

Chapter Five

Conclusions and Recommendations For Future Work

Introduction

The database systems are designed and used in all aspects of movie rating in sentiment analysis for Arabic reviews on the movies domain.

This chapter includes, in addition to the conclusions of the present work, recommendations for future work and for further studies on sentiment analysis algorithm for Arabic reviews on the movies domain.

Conclusions

The present research work formulates a technique for sentiment analysis for Arabic reviews on the movies domain. A lexicon was built for Arabic reviews on movie domain. We collected about 1250 (positive , negative , neutral) words. The positive contain 655 words , the negative contain 441 words and the neutral contain 154 words.

Sentiment analysis system was created for Arabic reviews on movie domain (Movie rating for arabic reviews system), with constructing a program code in Oracle language using the TOAD program for SQL to bulid the databaes for the movie rating for arabic reviews system.

Through selecting many Arabic and English movies from many different movies sites and collected many reviews about these movies to test the system. The system matches the reviews whether they are entered into the system within the lexicon that was built and analysis of these reviews to positive, negative and neutral sentiment and show the number, average, and percentage of all each movie reviews then compare the percentages and show the final result that represent the highest one.

The system also check if there are not available words contained in the system database over lexicon and show the number, average, and percentage of these not available words in all reviews for each movie.

The proposed system showed good performance in analyzing the reviews after manually testing and comparing since there are no similar automated system.

Recommendations for future work

This section offers some suggestions and recommendations for future work on the development of the sentiment analysis for Arabic reviews on the Policy and Government domain to make it easier for people to know the general impression about a specific political article and also to help politicians know the impression of people around them.

References

- Abdul-Mageed, M., Diab, M., & Kübler, S. (2014). SAMAR: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language*, 28(1), 20-37.
- Al-Ayyoub, M., Nuseir, A., Kanaan, G., & Al-Shalabi, R. (2016). Hierarchical Classifiers for Multi-Way Sentiment Analysis of Arabic Reviews. *International Journal of Advanced Computer Science and Applications*, 7(2), 531-539.
- Alotaibi, S. (2015). Sentiment Analysis in the Arabic Language Using Machine Learning (Doctoral dissertation, Colorado State University. Libraries).
- Alsmearat K., Shehab M., Al-Ayyoub M., Al-Shalabi R., & Kanaan G. (2015). Emotion Analysis of Arabic Articles and Its Impact on Identifying the Author's gender. The 12th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2015). Marrakech-Morocco.
- Altrabsheh, N. (2016). Sentiment analysis on students' real-time feedback (Doctoral dissertation, University of Portsmouth).
- Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., & Belyaeva, J. (2013). Sentiment analysis in the news. arXiv preprint arXiv:1309.6202.
- Basari, A., Hussin, B., Ananta, I., & Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering*, 53, 453-462.
- El-Beltagy, S., & Ali, A. (2013). Open issues in the sentiment analysis of Arabic social media: A case study. In *Innovations in information technology (iit)*, 2013 9th international conference on (pp. 215-220). IEEE.
- Goyal, A., & Parulekar, A. (2015) Sentiment Analysis for Movie Reviews CSE255/Fall2015/Assignment-2/MovieSentimentAnalysis.
- Hasan, S., & Adjeroh, D. (2011). Proximity-based sentiment analysis. In *Applications of Digital Information and Web Technologies (ICADIWT)*, Fourth International Conference on the (pp. 106-111). IEEE.

Heerschop, B., Hogenboom, A., & Frasinca, F. (2011). Sentiment lexicon creation from lexical resources. In International Conference on Business Information Systems (pp. 185-196). Springer Berlin Heidelberg.

<http://www.erpgreat.com/oracle-database/advantage-of-oracle-database.htm>. Access date at 17/5/2017.

[https://en.wikipedia.org/wiki/Toad_\(software\)](https://en.wikipedia.org/wiki/Toad_(software)). review at 17/5/2017.

<http://www.learn.geekinterview.com/database/oracle/advantages-of-using-oracle.html>. Access date at 17/5/2017.

<http://www.toptenreviews.com/business/software/best-translation-software/ace-translator-review/>. Access date at 17/5/2017.

Jim McDaniel, (2002) "Various (depending on the database used)": Toad Pocket Reference for Oracle plsql 1st Edition.

Joshi, M., Das, D., Gimpel, K., & Smith, N. (2010). Movie reviews and revenues: An experiment in text regression. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 293-296). Association for Computational Linguistics.

Kechaou, Z., Wali, A., Ben Ammar, M., Karray, H., & Alimi, A. (2013). A novel system for video news' sentiment analysis. Journal of Systems and Information Technology, 15(1), 24-44.

Koh, N., Hu, N., & Clemons, E. (2010). Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. Electronic Commerce Research and Applications, 9(5), 374-385.

Korayem, M., Crandall, D., & Abdul-Mageed, M. (2012, December). Subjectivity and sentiment analysis of arabic: A survey. In International Conference on Advanced Machine Learning Technologies and Applications (pp. 128-139). Springer Berlin Heidelberg.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg!. Icwsm, 11(538-541), 164.

- Maks, I., & Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, 53(4), 680-688.
- Mejova, Y. (2012). "Sentiment analysis within and across social media streams." PhD (Doctor of Philosophy) thesis.
- Mejova, Y. (2009). Sentiment analysis: An overview. Comprehensive exam paper, available on <http://www.cs.uiowa.edu/~ymejova/publications/CompsYelenaMejova.pdf> [2010-02-03].
- Michele, (2005). "Oracle Database Concepts, 10g Release 2 (10.2)", Oracle , p5.
- Na, J., Thura Thet, T., & Khoo, C. (2010). Comparing sentiment expression in movie reviews from four online genres. *Online Information Review*, 34(2), 317-338.
- Nabil, M., Aly, M., & Atiya, A. (2014). LABR: A Large Scale Arabic Sentiment Analysis Benchmark. ArXiv preprint arXiv: 1411.6718.
- Nielsen, F. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2), 1-135.
- Refaee, E., & Rieser, V. (2014). An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. In LREC (pp. 2268-2273).
- Tsutsumi, K., Shimada, K., & Endo, T. (2007). Movie Review Classification Based on a Multiple Classifier. In PACLIC (pp. 481-488).
- Vinodhini, G., & Chandrasekaran, R. (2012). Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6).
- Vishwanathan, S. (2014). Sentiment Analysis of Movie Reviews. In Proceedings of 3rd IRF International Conference, 10th May-, Goa, India.
- Yang, H., & Chao, A. F. (2015). Sentiment analysis for Chinese reviews of movies in multi-genre based on morpheme-based features and collocations. *Information Systems Frontiers*, 17(6), 1335-1352.
- Yessenov, K., & Misailovic, S. (2009). Sentiment analysis of movie review comments. 6.863 Spring 2009 final project, CSAIL.
- Yessenov, K., & Misailovic, S. (2009). Sentiment analysis of movie review comments. *Methodology*, 1-17.

Zhang, Z. (2014). Text Mining for Sentiment Analysis

<http://www.lib.ncsu.edu/resolver/1840.16/9940>.

Zhuang, L., Jing, F., & Zhu, X. Y. (2006, November). Movie review mining and summarization. In Proceedings of the 15th ACM international conference on Information and knowledge management (pp. 43-50). ACM.