COMPARISON OF MACHINE LEARNING WITH LEXICON-BASED APPROACHES FOR SENTIMENT OF ARABIC STUDENT FEEDBACK

OMAR AHMED NASSIF AL JASSEM

OCTOBER 2017

COMPARISON OF MACHINE LEARNING WITH LEXICON-BASED APPROACHES FOR SENTIMENT OF ARABIC STUDENT FEEDBACK

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES OF
ÇANKAYA UNIVERSITY

BY

OMAR AHMED NASSIF AL JASSEM

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF COMPUTER ENGINEERING

INFORMATION TECHNOLOGY PROGRAM

OCTOBER 2017

Title of Thesis: **Comparison of Machine Learning with Lexicon-based Approaches for Sentiment of Arabic Student Feedback**

Submitted by: **Omar Ahmed Nassif AL JASSEM**

Approval of the Graduate School of Natural and Applied Sciences, Çankaya University.

Prof. Dr. Can ÇOĞUN
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Prof. Dr. Erdoğan DOĞDU
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. James LITTLE
Co-Supervisor

Asst. Prof. Dr. Tolga PUSATLI
Supervisor

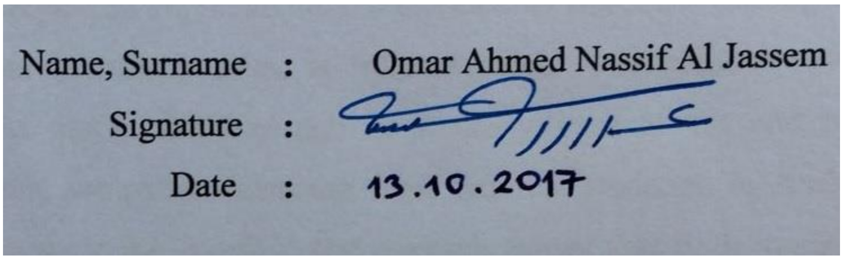
Examination Date: 13/10/2017

Examining Committee Members

Assoc.Prof.Dr. Reza HASSANPOUR	(Çankaya Univ.)	
Asst.Prof.Dr. Tolga PUSATLI	(Çankaya Univ.)	
Asst.Prof.Dr. Burkay GENÇ	(Hacettepe Univ.)	

STATEMENT OF NON-PLAGIARISM

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.


Name, Surname : Omar Ahmed Nassif Al Jassem
Signature :
Date : 13.10.2017

ABSTRACT

COMPARISON OF MACHINE LEARNING WITH LEXICON-BASED APPROACHES FOR SENTIMENT OF ARABIC STUDENT FEEDBACK

AHMED, Omar

M.S., Information Technology Department

Supervisor: Asst. Prof. Dr. Özgür Tolga PUSATLI

October 2017, 29 pages

Sentiment analysis has become an important area of research for machine learning and lexicon-based approaches. Currently there is little research to determine which approach is better across several factors, such as accuracy, ease of use and pre-requisites. In this research, we collect feedback from different students, in Arabic, expressing their feelings about the courses. The research shows that both machine learning and lexicon-based are good ways of solving the problem. Perhaps a dedicated algorithm is best, but requires initial work in setting up. Machine learning can be good but the user needs to find the right algorithm and training set. Overall, the commercial offering has high accuracy and ease of use. We use methods in the lexicon approach that give contrast values of accuracy. One of these tools is based on the ArSenL (Arabic sentiment dictionary), which produces high accuracy. The second tool of Lexicon is Lexalytics that displays not so high accuracy because its dictionary of Arabic language does not cover all Arabic words. Machine learning gives less sentiment accuracy depending on the training data and it needs time and effort to build a model.

Keywords: Sentiment analysis, Lexicon-based, Machine learning, Arabic lexicon, Arabic language.

ÖZ

ARAP ÖĞRENCİLERİN GERİ BİLDİRİMLERİNDEKİ DUYGULAR İÇİN SÖZLÜK TABANLI MAKİNE ÖĞRENME YAKLAŞIMLARININ KARŞILAŞTIRILMASI

AHMED, Omar

Yüksek Lisans, Bilgi Teknolojileri Anabilim Bölümü Dalı

Tez Yöneticisi: Yrd. Doç. Dr. Özgür Tolga PUSATLI

Ekim 2017, 29 sayfa

Duygu analizi, makine öğrenme ve sözlük tabanlı yaklaşımlar için önemli bir araştırma alanı olmuştur. Günümüzde, doğruluk, kolay kullanım ve ön koşul gibi etkenler arasında hangi yaklaşımın seçilmesi üzerine az araştırma vardır. Bu çalışmada, farklı öğrencilerden, dersler hakkında ne hissettiklerini belirten Arapça geri bildirimler topladık. Araştırma, makine öğrenme ve sözlük tabanlı yaklaşımların her ikisinin de problem çözmede iyi şeçenekler olduklarını gösteriyor. Makine öğrenme iyi olabilir ancak, kullanıcı doğru algoritmayı ve eğitim setini bulmalıdır. Genel olarak, ticari sunumun yüksek doğruluğu ve kolay kullanımı vardır. Sözlük yaklaşımında doğrulukta zıt değerler veren yöntemler kullandık. Bu araçlardan biri, Arapça duygusal sözlük olan, yüksek doğruluk veren, ArSenL üzerine kuruludur. İkinci araç olan Lexican, o kadar yüksek doğruluk göstermeyen Lexalytics'dir; yüksek doğruluk gösterememe nedeni, Arapça sözlüğünün Arapça'daki tüm sözcükleri içermemesidir. Makine öğrenme, eğitim setine bağlı olarak daha az doğruluk göstermekte ve bir model kurabilmek için daha çok zaman ve gayret gerektirmektedir.

Anahtar sözcükler: Duygu analizi, Sözlük tabanlı, Makine öğrenme, Arapça sözlük, Arapça dili

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation and gratitude to Dr. James Little and Dr. Tolga Pusatli for support and encouragement during my work for this thesis. They directed me to a broad range of resources on the web and in the library. They answered all my questions and helped me to narrow my search. I thank them for their constructive and positive criticism and for patiently reviewing my work. I would also like to thank Ali Khudhair for his unlimited support during the beginning of this work.

I owe special gratitude to my family. Words cannot express how grateful I am to my father, my mother, brothers, and sisters in my home country. Special thanks to my family, my wife, sons for their continuous support. Lastly, for all my friends who support me at Çankaya University, I give them my thanks.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES

LIST OF TABLES

TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ArSenL	Arabic Sentiment Lexicon
DT	Decision Tree
ML	Machine Learning
NB	Naive Bayes
NLP	Natural Language Processing
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

1.1 Background

Student Feedback

Student feedback covers many different situations that students may encounter with the lectures. Typically, feedback is gathered and analyzed at the end of the lecture, which has benefits for both lecturers and students, such as improving the lecturers' teaching ways and understanding students’ learning behavior. Students’ feedback enhances communication between them and their lecturer by making the lecturer fully understand the students’ opinion [7].

The common way to obtain student feedback is to ask them questions about the expressed lecture, for example, if they grasped a specific section of the lecture. Nevertheless, this method is not suitable for all students, such as low self-confidence and shy ones [8]. The other way to accumulate feedback from students is through social media. Statistics show that 95.4% of 19 years old and 93.5% of 18 years old in the United States used online social media regularly [9]. This survey also showed that in 2010, 52.1% of academics used Twitter, as well as over 470 universities using Twitter and Facebook to contact their students.

Sentiment Analysis

Some words have an obvious positive or negative sentiment, "good" or "bad" for example. Sentiment analysis is a way to process vast amounts of online feelings in social media. It is mainly used to specify the polarity of a given text by identifying if the expressed opinion is positive or negative. Polarity is considered as the positivity degree of each word in the sentence, whereas the sentiment is the summation score of these polarities in the sentence [18]. Some words may take a sentiment that depends on the context, such as the word ‘early’ which produces negative sentiment in education field in terms of time as in the example “The lecture is too early!”. Yet, when explaining a parcel service as in “The parcel arrived early”, it has a positive sentiment.

We can express sentiment in many ways:

- Sentiment classification of a sentence: this level classifies the sentence into negative, positive or neutral. This level is related to our work since we collected feedback from students in form of a couple of sentences.
- Sentiment classification of a document: it tries to classify the entire document as being positive, negative or neutral. This classification may be used when we need to predict opinions of the document as one unit, for example if we want to be informed about comments on news review a document level analysis can tell us if they are positive, negative or neutral.
- Summary of opinion: here, the concern is to identify all opinions in various documents. Then it tries to give a brief overview about the positive or negative feedback. For instance, it tries to conclude and extract information within rates: "70% of people liked i-phone" [2].

Twitter is considered one of the top social networking websites because it is a great source for capturing people's opinions about different topics. There are sentiment researches which take their data from Twitter (such as Neethu M S et al. [1], M. Al-Ayyoub et al [3], M. Althobaiti [5], Altrabsheh et al. [9], J Jmal et al. [10] and more) and those texts are very similar to student feedback we work on.

In the last decade, sentiment analysis research has grown due to the existence of rich text resources represented by social network sites, product reviews and blogs. In the educational domain, sentiment analysis focuses more on e-learning than classroom feedback [7, 8].

There are two main approaches, which can detect sentiment automatically: machine learning and lexicon-based. The machine learning approach to sentiment analysis is one of classifying text. It can be processed by training the classifier on a collection of sentiment-labeled text. Thereafter the model can classify new text. The lexical approach determines sentiment based on analysis of individual words and/or phrases based on a dictionary. Sentiment dictionaries are used to match words or phrases from the text. Then their sentiment weights are counted and aggregated to give an overall rating.

Until now, it was not clear which approach is better with respect to performance because each one depends on a number of variables such as the size of the data, general or specific domain of data and the effort and time for building. M Abdul-Mageed et al [4] showed that in most of the cases the machine learning approaches outperformed the lexicon-based approaches for particular features for example identifying whether the text is objective or subjective before identifying its polarity. Ahmed Abbasi et al [24] stated that their proposed features and techniques gave accuracy over 95% using SVM on two languages English and Arabic. A. Khan et al [27] claimed that their proposed method classifies English sentences from reviews and blog comments using lexicon-based approach. The score of each sentence is obtained from SentiWordNet [10] (a large-scale sentiment lexicon for English language) to calculate its sentiment as positive, negative or neutral. This achieved an accuracy of 86.6% at the sentence level. We cannot say that this is bad accuracy because we obtained 83% from Lexalytics and we still consider this as good accuracy. They applied their approach on one domain just like how we applied Lexalytics on only student feedback.

We will review many more researches in chapter 2 in more details.

1.2 Research Question and Contribution

The aim of this thesis is to present a comparison between the two approaches of sentiment analysis for us to evaluate their performances on Arabic student feedback. Our aim is to answer the following research questions:

- Which is better, machine learning or lexicon?
- What is the ease of use of each method?

- How do commercial sentiment analysis tools (such as Lexalytics) work on Arabic short text?
- What are the best machine learning algorithms and training sets from different domains?

The contribution of our research is to make the first comparison between three main sentiment approaches (Machine learning, Lexicon-based and Lexalytics tool) on Arabic student feedback, since little research has been done in this area.

CHAPTER 2

STATE OF THE ART

2.1 Sentiment Analysis Approaches

There are two main approaches to making a sentiment analysis, machine learning and lexicon-based. The machine learning approach utilizes algorithms to categorize the data depending on linguistic features such as the words and the grammatical structure. The lexicon-based approach classifies the dataset directly using a dictionary of words and phrases. In the following sections, we will review other works on both approaches and their algorithms and lexicon dictionaries in more detail related to our research in terms of Arabic student feedback and tweets sentiment analysis. These are the important aspects of our research.

2.2 The Lexicon-Based Approach

The lexicon-based approach relies on sentiment lexicons such as SentiWordNet, which are needed to classify a given English text by the words and phrases it contains. One such lexicon is ArSenL (Arabic Lexicon Sentiment) [19]. It is a dictionary of Arabic words built based on Arabic Word-Net (AWN), English WordNet (EWN), SAMA and SentiWordNet (ESWN), because there is no direct Arabic lexicon for sentiment. These are the lexicons which are ArSenL-based approach was built based on, because there is no direct Arabic lexicon as we mentioned. Arabic WordNet (AWN) is a collection of Arabic words that is also built on English WordNet. A lexicon consists of many words, each word in it has been assigned a number representing its polarity (negative, positive or neutral) as well as strength measurement [11].

Lexicon-Based Approach for Student Feedback

Nasim, Z. et al [25] is perhaps the closest to ours. They tried to classify student feedback, but in English using a hybrid approach. A hybrid approach combines the use of sentiment dictionary and machine learning methods in one system. The sentiment lexicon was the modified version of MPQA (Multi Perspective Question Answering) lexicon. In addition, the machine learning algorithms they used were decision tree-based and SVM. Their results showed that the accuracy obtained from their proposed hybrid approach is 93%, which is the same value of SVM on our student feedback.

There have been a few other research studies on English student feedback using a lexicon approach. Some, such as MacKim et al. [12], looked much deeper than what we are doing by trying to identify emotions. Their research categorized five emotions: surprise and joy had positive polarities, while sadness, fear and anger belong to negative polarity. The highest accuracy was 84.7% for the joy emotion from Fairy Tales dataset (The data set is composed of three stories for children including Grimms', Hans Christian Andersen's, and Beatrix Potter's stories) using WordNet dictionary.

Lexicon-Based Approach for Twitter Analysis

M. Al-Ayyoub et al. [3] built a very large sentiment lexicon and a lexicon-based sentiment analysis tool. They claimed that the results show that the proposed tool performs very well by reaching an accuracy nearly 86.89% on Arabic general tweets just like one of our datasets.

M Fernández-Gavilanes et al [26] proposed an approach to predict sentiment in online text messages such as tweets and reviews in English. Their system is not directly comparable to ours due to the difference in datasets and their lexicon. They applied their system on The Obama-McCain Debate tweets, while our dataset was tweet like student feedback. We modified ArSenL-based approach and they created a new sentiment lexicon PolarityRank 40 (PR40) using a total of 40 positive and negative seeds of words. They reached an accuracy of 74.80%.

Zhou et al. [22] proposed an expanded lexicon with domain-dependent opinion words, abbreviations and informal opinion expressions for analyzing Twitter texts. They developed SentiStrength (SS) (a lexicon-based classifier that incorporates a booster word list, an emoticon list, an idiom list, a negation word list, a question word list, a slang list and a general opinion word list) to reach an accuracy of 88.29%. We did not add in any special student sentiment extensions in our research, as none were apparent in how students expressed themselves.

2.3 The Machine Learning Approach

Machine learning is a technique, which tries to predict the sentiment of a piece of text based on historical examples of text with the sentiment known.

In Machine learning, two types of datasets are required: a training dataset and a test dataset. A classifier (such as Naive Bayes and Support Vector Machines) learns from the training set and the accuracy in classification can be evaluated using the test set.

A classifier is an algorithm that separates words or sentences according to their polarity to negative or positive. There are many classifiers in machine learning approach such as Naive Bayes, Maximum Entropy, Support Vector Machines, Decision Tree, K Nearest Neighbors and Sequential Minimal Optimization. The main function of these classifiers is to predict sentiment that may or may not be accurate, and then the precise prediction of the classifier compared to the ground truth (human classification).

There are many machine learning classifiers, but the well-known ones are:

- Support Vector Machine: is commonly used in text classification by examining the data. It estimates the text by linearly splitting the data in a feature space, such as given training data, annotated with the class of the data. The algorithm then constructs a model that symbolizes the data as points in space. A vector hyperplane represents these points. The points are mapped to make the gap among the various classes as big as possible. New data to be tested is mapped to the exact same space and then classified by detecting which part of the hyperplane they fall on.

- Naïve Bayes Classifier (NB): It has been used because of its simplicity elegance, speed and robustness in both training and classifying stage. It learns the probability of each attribute, generates a probabilistic model of the features and gives the class label. The probabilistic model is utilized to anticipate the class of new instances utilizing the highest posterior probability.
- Decision tree classifier (DT): As the name implies, DT breaks down a dataset into smaller subsets while at the same time an associated DT is incrementally developed. The final result is a tree with decision nodes and leaf nodes. It develops the tree using a certain measurement of information that requires the maximum information when the number of the two main leaves is equal.

Machine Learning for Student Feedback

Dhanalakshmi V. et al. [8] also used RapidMiner. They focused on using Opinion Mining technique for classifying the student feedback (on six courses) obtained during a survey with respect to various features of teaching and learning. Their results showed that the highest accuracy average went to NB with 99.11%, the other classifiers obtained average of accuracy as follows: SVM 91.41%, K-NN 98.13% and lastly NN 98.52%. Altrabsheh et al. [9] also worked on mining feedbacks of students collected from social media like Twitter using NB and SVM. They found out that Naive Bayes and SVM techniques were superior for education data. They introduced their system architecture termed Systems Analysis for Education (SA-E) as well.

Machine Learning for Twitter Analysis

Neethu M S et al. [1] tried to analyze English Twitter posts about electronic products such as mobiles, laptops etc. by classifying the tweets as positive and negative. They used three machine learning algorithms SVM, Nave Bayes, Maximum Entropy with accuracies of 90%, 89.5% and 90% respectively.

B. Li et al. [13] used general English tweets as dataset for training Naive Bayes and SVM classifiers to extract the sentiment. The accuracy was 80% for both algorithms. Zhang et al. [15] proposed a new sentiment analysis method for Twitter. Firstly, a lexicon-based approach (called Holistic lexicon-based approach built based on WordNet) is adopted to automatically classify the tweets. Afterwards, a sentiment classier (SVM) is trained to assign sentiment polarities for the newly identified tweets. The training data for the classier is the result from the lexicon-based method. The accuracy the obtained was 85.4%.

2.4 Comparison of Sentiment Analysis Techniques

The machine learning approaches require a big labeled training data set (text with sentiment) whereas lexicon uses unlabeled dataset (text with no prior sentiment). Here, we will describe more studies about the main techniques employed by researchers along with the achieved accuracy. In addition, the data given by the authors is presented too.

There are many similarities between our work and Abdulla, N. et al [14]. They applied their approaches on Arabic tweets on various topics such as politics and arts, whereas we analyzed general and politics tweets datasets. An important similarity is that the size of tweets is like our student feedback because it is focused set of few words. They also used a lexicon approach by building their own dictionary and employing RapidMiner to run the machine learning algorithms. Their results show that the accuracy acquired from the lexicon-based is 59.6%, while machine learning algorithms obtained an accuracy of 87.2% SVM, 81.3% NB, 51.45% KNN and 50% D-tree respectively. O. Appela et al. [23] presented a hybrid approach and a comparison between their approach and machine learning two algorithms (Naive Bayes and Maximum Entropy) on two different datasets Twitter and movie review. They enhanced SentiWordNet sentiment lexicon dictionary to estimate the polarity and strength of sentences as well as natural language processing (NLP) and fuzzy sets to build the system. Naive Bayes classifier acquired accuracy of 67.8% on Twitter and 67.1% on movie dataset whereas Maximum Entropy acquired 67.5% and 67.5%. the proposed approach obtained accuracy of 88% on Twitter and 86.5% on movie review. Geetika Gautam et al. [11] also used Twitter as a data source to classify customers' reviews by employing the three algorithms. as well as a lexicon approach using WordNet. Accuracies obtained were Naive Bayes 88.2%, Maximum Entropy 83.8%, SVM 85.5% and WordNet 89.9%. Asmaa Mountassir et al. [28] carried out a study to address the problem of unlabeled data sets in machine learning sentiment classification in an Arabic context by labeling the data manually. They collected their data set from online forums of Aljazeera’s web site regarding politics domain. Their results showed that the accuracy they acquired was 91.2% using NB classifier representing the first research and 65.7% using NB and 61.4% using SVM representing the second research respectively.

Basically, the performance of lexicon approaches depends on contents of lexicon dictionary. If there is a small number of words in the dictionary, then it leads to noticeable decrease in the performance of the lexicon model. On the other hand, the performance of machine learning algorithms depends on how well the classifier is trained.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In order to determine the performance characteristics of the various sentiment analysis approaches (Lexicon and Machine learning), we compare them on a student feedback dataset. This chapter presents a detailed description of the research methodology for this comparison. Figure 1 first shows the flowchart of the methodology. In the next sections, we will describe each stage of the methodology.

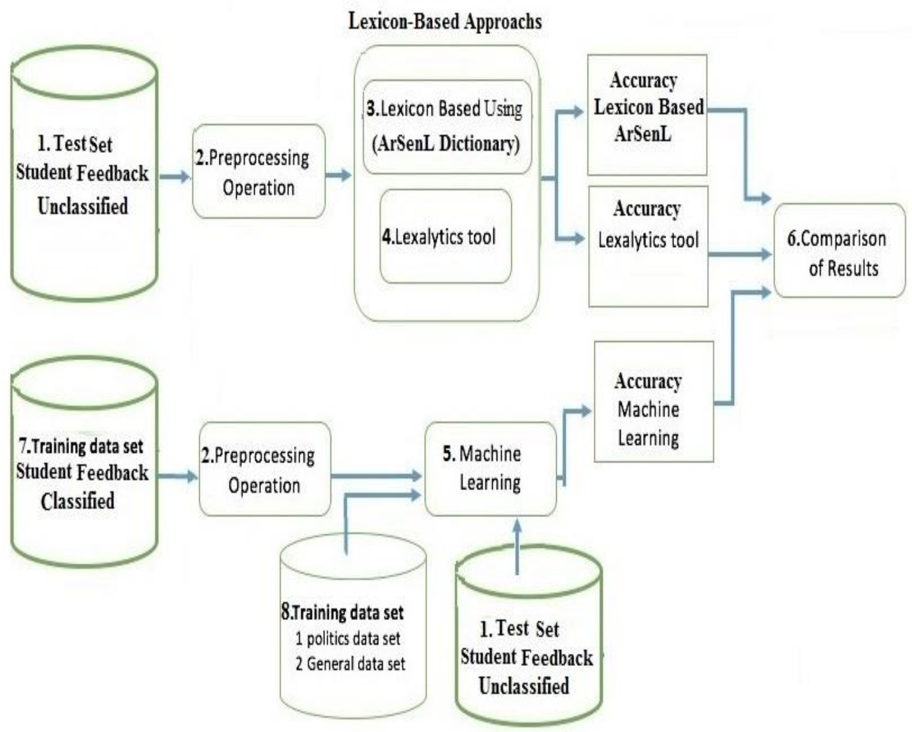


Figure 1 Flowchart of the Methodology

3.2 The Datasets (boxes 1, 7 & 8)

The creation of the datasets is the first step in our methodology. Datasets comprise typically of pieces of text with possibly a sentiment value (positive, negative, neutral) associated with them.

Lexicon-Based Datasets

The main testing dataset (box 1) is collected from student feedback from different courses. It does not initially have a sentiment value, but a ground truth is established that is later used for accuracy.

Machine Learning Based Datasets

The training sets (box7) for machine learning are firstly student feedback from different courses plus ground truth. In addition, (box 8), we collect other data sources as general and political tweets, which includes their sentiment values. A test set (box 1) is also required for external accuracy test data (student feedback in our case study), which was not used in building the model.

3.3 Lexicon-Based Approach (i) (Using ArSenL-Based Approach) (boxes 2 and 3)

After collecting the data from the students, we clean and prepare the text for classification by first extracting a word list. The lexicon-based approach takes each word and tries to make a match into the ArSenL database, which contains a very large number of Arabic words. These are sentiment bearing Arabic words, which also carry a sentiment value. However, many words carry no sentiment and so are not present and hence ignored. Therefore, we can add together the individual sentiment values to calculate an overall sentiment score for the text. This approach is based on the thesis of A. Khudhair [16] and we use that software. In the process of our research, we have significantly increased his sample size from 55 to 122 pieces of student feedback.

3.4 Lexicon-Based Approach (ii) (Lexalytics Tool) (box 4)

Lexalytics [6] is a piece of commercial software, which provides an interface for measuring sentiment from unstructured text. Their method is called phrase-based sentiment and uses a similar dictionary with a sentiment score for every word and phrase. Lexalytics supports the Arabic language and is an important reason behind our choice, as it is one of the few which provides this as well as it is easy to deal with short sentences just like our student feedback. Whereas other tools also support Arabic yet with less efficiency. One other possibility was to use Weka. However, it does not recognize Arabic characters in the ASCII format, thus it needs Buckwalter transliteration method to encode Arabic characters. Another key factor Weka is poorer than RapidMiner is that it does not categorize the text directly, it uses an intermediary language. After the student feedback has been uploaded, the tool rates the feedback as positive, negative, or neutral and assigns a score to show how strong that sentiment is.

3.5 Machine Learning (box 5)

We choose RapidMiner [21] to build our model. It is an open source platform and provides many algorithms that allow the user to build and evaluate classification models. RapidMiner also provides a graphical way to design and execute analytic workflows. The main reason for choosing RapidMiner over other similar tools such as Weka [17] is that RapidMiner supports the Arabic language and deals with the Arabic very well and performs quickly. For example, RapidMiner accepts excel sheet file of data directly without any conversion to Arabic and it provides Graphical User Interface. On the other hand, Lexalytics supports the least features for Arabic language, for example it supports only sentiment and entities (people, places, companies, brands, @mentions, #hashtags, and job titles) while it supports Summarization, POS Tagging, Sentiment, Boolean Queries, Categorization, Intentions, Entities and Themes for English language. We chose Lexalytics as a representative software which has a good representation. We were not doing a review of all commercial lexicon software.

The basic premise of machine learning is to use algorithms that can take data and analyze looking for patterns. This will tell us something about the problem, which was not evident.

3.5.1 Machine Learning (i) Process Flow

The first step in machine learning is building the model, then using labeled student feedback data training to train our model using the collected student feedback. Figure 2 below shows the stages of a machine learning approach using RapidMiner. We will explain each stage.

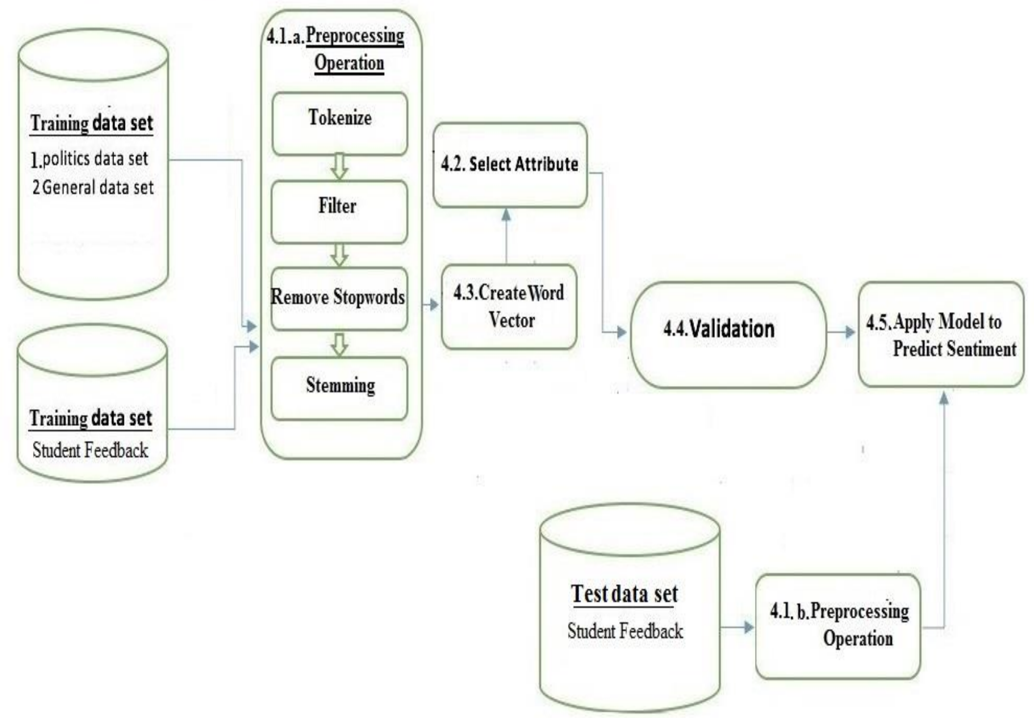


Figure 2 Flowchart of Machine Learning

3.5.2 Preprocessing Operations (box 4.1a & b)

Preprocessing consists of four sequential steps: tokenize, filter, removal of stopwords and stemming.

Tokenize means splitting the text of the feedback into words or tokens. Tokens are separated from each other by spaces or other special characters.

Filter This operator filters tokens based on their length in characters. Minimum and maximum number of characters that a token can contain, are parameters of this operator.

Removal of stopwords is a process of deleting common words from the text that generally do not contribute to the meaning of a sentence (hence its sentiment). This filter also removes single letter words.

Stemming is a process of removing the suffix of the word and returning its root. The sentiment classifier should not treat the words "يحب", "تحب" and "يحبون"as different words, especially when they have the similar sentiment and meaning. The stemmer will return the root “حب (which means "admire")” and then the frequency of this word will be 3 instead of three words with frequency 1 because it will eventually delete all the suffixes and prefixes for Arabic words.

3.5.3 Create Word Vector (box4.2)

This process takes each token (word) in the text and generates a vector that represents each piece of text numerically. The feedback vector represents each possible word and a value whether the word is in that piece of feedback or not.

3.5.4 Select Attribute (box 4.3)

The Select Attribute analysis tries to identify which attributes affect the resulting accuracy most and which do not. From this, it is possible to create a smaller problem, which may give improved accuracy. We include this step as one of the common filtering techniques within machine learning, yet we only explore this option in a limited way, to determine whether it improves accuracy or not, using a default setting. This operator selects attributes to keep and to eliminate. Different filter types may be selected in the parameter attribute filter type and only attributes satisfying this condition type are kept while remaining are discarted. Its methodology works by evaluating group of attributes and recording their accuracy then recommending the best group.

3.5.5 Validation (box 4.4)

The Validation process has two parts: a training part and a testing part. The training phase is used to build a model by selecting an algorithm and providing example data (word vectors plus sentiment). Then the same model may be applied to another dataset (unknown sentiment) for prediction. The performance of the model is measured during the testing phase with respect to the accuracy. We consider four built-in algorithms/classifiers for evaluating the accuracy of sentiment in this domain. The classifiers are Decision Tree classifier (DT), Naïve Bayes (NB), Support Vector Machines (SVM) and Sequential Minimal Optimization (SMO).

3.5.6 Apply Model to Predict Sentiment (box 4.5)

A model based on one of the machine learning algorithms is applied to a new test data set of unclassified data. It will predict the sentiment value (positive, negative or neutral). We then compare this with the ground truth to assess the accuracy.

3.5 Machine Learning Approach (ii)

This part of the process is used to measure the performance based on general and political data training. Machine learning (i) is used on student feedback. We use the general and political tweet data to train the model to predict the sentiment. We see how the performance of the model, when using training data from different sources, has an impact on the accuracy.

3.6 Comparison of Results (box 6)

We compare the accuracy (the final score of the sentence in comparison to the ground truth) which we obtain from three different approaches, lexicon-based (ArSenL), Lexalytics sentiment tool and machine learning (three training sets).

CHAPTER 4

RESULTS

This chapter presents the results of the proposed four models built using Machine Learning approach with four algorithms (SVM, NB, DT and SMO) and Lexicon approach (with ArSenL-based approach and Lexalytics). It shows results to compare the accuracy.

4.1 Lexicon-Based Approach (Using ArSenL-Based Approach)

We used the ArSenL-based approach on student feedback to obtain a sentiment. As we can see from Table 1, we obtained a 100% accuracy when comparing each lexicon result with the ground truth. This means that all sentiment predictions are exactly the same as the human opinion.

Table 1 Results from ArSenL-based approach

No	Arabic Feedback	English Translated	Pos_Score	Neg_score	Sentiment	Ground Truth
1	الاستاذ جيد جدا ومتفاعل مع الطلاب والمواضيع العلمية جيدة جدا وستستفيد منها في اعداد الاطروحة.	Professor is very good and interactive with the students very good scientific subjects and will benefit them in the preparation of thesis.	1.59	0.39	Positive	Positive
2	المحاضرة ممتعة مع معلومات كبيرة عن تعريف المنهج.	Enjoyable lecture with a lot of information about the definition of the curriculum.	0.74	0.12	Positive	Positive
3	هناك تشويش ادى الى عدم فهمنا للمحاضر.	There is confusion in the ideas about subject and I am cannot understand.	0.11	0.55	Negative	Negative
4	تضمنت المحاضرة الكثير من المعلومات عن المواضيع وقد تعلمت اشياء كثيرة ومفيدة عن هذه المواضيع والاستاذ مرح في المحاضرة وقد شعرت بالراحة.	The lecture included a lot of information from the subjects, and I learned many useful things about this subjects and professor frolic in the lecture I felt comfortable.	0.75	0.12	Positive	Positive
5	الاستاذ لديه معلومات كثيرة عن المادة وسيجعلنا نفهمها جيدا.	A professor has a lot of information about the lecture and will make us understand it well.	0.80	0.05	Positive	Positive
6	شرح المحاضرة جيد جدا والاستاذ يحاول على ايصال الفكرة الى طالب.	Explain very well to lecture and professor tries to deliver the idea to the student.	0.94	0.20	Positive	Positive
7	المواضيع العلمية جيدة والاستاذ جيد.	Good scientific subjects a good professor.	1.09	0.05	Positive	Positive
8	لاننا نحتاج الى امثلة من الواقع العملي في الحياة لذلك المواضيع صعبة.	Subjects need to summarize and we need a lot of time so I did not understand anything.	0.11	0.87	Negative	Negative
9	زخم المعلومات ادى الى عدم فهمنا للمحاضرة.	Information momentum led to did not understanding the lecture.	0.21	0.47	Negative	Negative
10	لاننا نحتاج الى امثلة من الواقع العملي في الحياة ادى الى عدم فهمنا . للمحاضرة	We did not understand because we need examples of practical reality in life.	0.16	0.33	Negative	Negative
11	المادة صعبه والاستاذ لايستطيع توضيح المواضيع الى الطلاب.	The Subjects difficult and professor cannot explain the topics to students.	0.20	0.88	Negative	Negative
12	المحاضرة تحتوي على مواضيع صعبة و كثيرة واخشى الحصول على نتائج سيئة في الامتحان	The lecture contains the difficult subjects and many, fear getting bad results in the exam.	0.14	0.80	Negative	Negative
13	المحاضرة تحتوي على مواضيع كثيرة واخشى ان تكون صعبة جدا.	The lecture contains many subjects and I am afraid not understand.	0.29	0.88	Negative	Negative
14	مواضيع كثيرة جدا و صعبه ولا استطيع فهم جميع الامور بصورة كاملة ونرغب بالتركيز على المواضيع الصعبة و المهمة	Too many topics and I cannot understand all the things fully and we want to focus on important topics.	1.07	1.85	Negative	Negative
15	المواضيع صعبة وبحاجة الى المزيد من الامثلة.	The Subjects difficult to understand and we need more examples.	0.04	0.81	Negative	Negative

4.2 Lexicon-Based Approach (Using Lexalytics Software)

Table 2 shows the results that we obtained from the Lexalytics tool. We evaluate our accuracy again by using the same ground truth of the testing data. An 83% of accuracy is good, yet poorer than the hand-built ArSenL-based approach. These are short sentences and so any single error in misclassifying a word has a big effect, compared to a paragraph of text. This seems to be what happened here. The difference between ArSenL and Lexalytics is the type of the dictionary itself, ArSenL-based approach uses a dictionary that contains a large number of Arabic words. The system checks words through the ArSenL where they are matched against the list of words. If we have a matching, then the system returns positive/negative value and stores. While the Lexalytics tool works by calculating the sentiment of each phrase or word in a single text, depending on the weight of each phrase/word in its dictionary. It also looks for specific emotive phrases that carry sentiment. Lexalytics also detects indicators of negative sentiment. In comparison with the lexicon approach in 4.1, Lexalytics does a much deeper analysis of the text; it goes beyond simple word sentiments, as it examines phrases and context to the words. Lexalytics claims good performance at all sentence, paragraph and document level. Lexalytics works well with the big files of words while the ArSenL-based approach has been only designed for short sentences such as tweets and student feedback.

Table 2 Results from Lexalytics Model

No	Arabic Feedback	English Translated	Sentiment Score	Sentiment	Ground Truth
1	الاستاذ جيد جدا ومتفاعل مع الطلاب والمواضيع العلمية جيدة جدا وستستفيد منها في اعداد الاطروحة.	Professor is very good and interactive with the students very good scientific subjects and will benefit them in the preparation of thesis.	0.12	Positive	Positive
2	المحاضرة ممتعة مع معلومات كبيرة عن تعريف المنهج.	Enjoyable lecture with a lot of information about the definition of the curriculum	0.45	Positive	Positive
3	هناك تشويش ادى الى عدم فهمنا للمحاضر.	There is confusion in the ideas about art and I am trying to understand with examples, but I cannot.	-0.94	positive	Negative
4	تضمنت المحاضرة الكثير من المعلومات عن المواضيع وقد تعلمت اشياء كثيرة ومفيدة هذه المواضيع والاستاذ مرح في المحاضرة وقد شعرت بالراحة.عن	The lecture included a lot of information from the subjects, and I learned many useful things about this subjects and professor frolic in the lecture I felt comfortable.	0.51	Negative	positive
5	الاستاذ لديه معلومات كثيرة عن المادة وسيجعلنا نفهمها جيدا.	A professor has a lot of information about the lecture and will make us understand it well	0.93	Positive	Positive
6	لا يتم تحديد ما مطلوب والصعوبة في كيفية التعامل مع النتائج في البرنامج.	It did not specify what is required and how to deal with the results of the program.	0.94	Positive	negative
7	جيدة المحاضرة وتم الاستفادة من جميع المعلومات	The lecture was good to take advantage of the information in the lecture.	-0.42	Negative	positive
8	المحاضرة جيدة من حيث الاعداد والتقديم والاستاذ متمكن من المواضيع	The lecture is good preparation and presentation and Professor versed subjects.	-0.33	Negative	positive
9	المحاضرة جيدة.	The lecture is good.	-0.94	Negative	positive
10	المادة صعبه والاستاذ لايستطيع توضيح المواضيع الى الطلاب.	The Subjects difficult and professor cannot explain the topics to students.	-0.94	Negative	Negative
11	هناك غموض و صعبه في المادة.	There is ambiguity in the subjects.	-0.91	Negative	Negative
12	المحاضرة تحتوي على مواضيع صعبة و كثيرة واخشى الحصول على نتائج سيئة في الامتحان.	The lecture contains the difficult subjects and many, fear getting bad results in the exam.	-0.92	Negative	Negative
13	المحاضرة تحتوي على مواضيع كثيرة واخشى ان تكون صعبة جدا.	The lecture contains many subjects and I am afraid not understand.	-0.94	Negative	Negative
14	مواضيع كثيرة جدا و صعبه ولا استطيع فهم جميع الامور بصورة كاملة وترغب على المواضيع الصعبة و المهمة.بالتركيز	Too many topics and I cannot understand all the things fully and we want to focus on important topics.	-0.56	Negative	Negative
15	المواضيع صعبة وبحاجة الى المزيد من الامثلة.	The Subjects difficult to understand and we need more examples.	-0.42	Negative	Negative

We see that Lexalytics wrongly classified what obviously appears positive or negative statements. Therefore, Lexalytics was unable to achieve 100%. On further examination, the word "good" (in Arabic) appears in many of the statements incorrectly classified. We tried the same sentence in the English version of Lexalytics and found that they were correctly classified. Therefore, it is the Arabic version of "good" that is causing an issue.

In Arabic, the word "good" has several versions and extensions, some are positive and others are negative. This word comes in one written form but different spellings and meanings, and this we believe, presents the difficulty [20] thus it suggests that Lexalytics was fooled by the variants of word "good". We tried using the different variants of “good” and determined what the sentiment of the corresponding sentence was, as shown in table 3 below:

Table 3 Arabic Variation of "good"

Arabic "variant" to good	English Meaning	The lecturer is “variant”	Sentiment
جَيّد	good	المحاضر جيد	Positive
جِيدَ	thirst		Negative
جِيدُ	girl's neck		Neutral

So Lexalytics may have taken the second variant, but ArSenL-based approach on the other hand may assume the word means good. In Table 2, sentences no 7,8 and 9 have the word "جيد" that means "good", however the sentiment for these three sentences is negative using Lexalytics, and we do not see such thing in ArSenL-based approach. ArSenL removes all the diacritics therefore, it considers "جَيّد", "جِيدَ" and "جِيدُ" as "good". We put alternative words into the sentence and showed that Lexalytics can get it right (table 4). There are synonyms for the word "good" in Arabic [20].

Table 4 Arabic Alternatives of "good"

Arabic "alternative" to good	English Meaning	The lecturer is “alternative”	Sentiment
ممتاز	excellent	المحاضر ممتاز	Positive

عظيم	great	المحاضر عظيم	Positive
حسن	well	المحاضر حسن	Positive

4.3 Machine Learning Approach

We build four different models by employing four different algorithms. We also used training sets from different domains such as general, political and student feedback.

4.3.1 Machine Learning with Student Feedback Training Set

We used the same student feedback for both training and testing the model. Note that the testing is done on a new unseen test set. We obtained the following results in table 5 as follows:

Table 5 Accuracy of Using Student Feedback Data Set

Data training	SVM	NB	DT	SMO
Student feedback	93 %	89 %	76 %	61 %

We observe that the accuracies of Support Vector Machine (SVM) and Naive Bayes (NB) are quite higher than the other two algorithms and have reasonable accuracy. The success of these two algorithms on sentiment classification has also been seen in the research of Neethu M S et al. [1], R. Duwairi et al. [2], N. Altrabsheh et al. [7] and more.

4.3.2 Machine Learning Using General and Political Data Training

Table 6 shows the accuracy of training on general and political tweets and applying the resulting model to a feedback test set.

Table 6 Accuracy of Using General and Political Tweets as Training Set

Data training set	SVM	NB	DT	SMO
Political	78 %	72 %	70 %	60 %
General	58 %	57 %	56 %	54 %

We see that again SVM is the best algorithm for both general and political tweets with NB similarly coming second. Generally, accuracies are higher using a political training set than a general one because concepts in politics are perhaps closer to student feedback than general, which has so many different topics. However, both training sets are inferior to the student feedback training set in terms of accuracy. This shows that it is important to train and test a model on the same domain. Still the results were reasonably good for the political set.

In our research, we used different algorithms, but now we consider only NB and SVM because they consistently produced higher accuracies. Table 7 below shows the summary results of three main approaches for sentiment analysis in classifying student feedback.

Table 7 Summary of Accuracy Results

	Data Training	Accuracy
Lexicon-based approach		
ArSenL	-	100 %
Lexalytics	-	83 %
Machine learning approach		
SVM	Political	78 %
Naive Bayes	Political	72 %
SVM	General	58 %
Naive Bayes	General	57 %
SVM	Student feedback	93 %
Naive Bayes	Student feedback	89 %

As we can see, the lexicon-based approach using the ArSenL-based approach is the best for this test set. The different domains (political and general) had the least accuracy because the training data differs from the testing data. The accuracy of Lexalytics was the least in terms of student feedback with accuracy nearly 83%. Machine learning algorithms on the other hand SVM and NB came in between the accuracies of ArSenL-based approach and Lexalytics.

We did not obtain any neutral sentence because all feedback we acquired was either positive or negative. One of the reasons for A. Khudhair's [2016] approach doing well, is that the students have strong views and so it is easy to detect a sentiment. Moreover, in a few words, there is no room for ambiguity or conflicting opinions.

CHAPTER 5

CONCLUSIONS

The main goal of this research is to make a comparative investigation between machine learning and lexicon approaches to classifying the sentiment of student feedback. A hand-built sentiment analyzer based on finding the sentiment of each Arabic word (using ArSenL-based approach) is the best performing with accuracy of 100% in comparison to the ground truth. This approach works well on short Tweet-like sentiments. Lexalytics was the easiest to use as it does not need any preparation, we just put the text we want to find the sentiment and the result comes out quickly. It gave reasonable results if it were not for stumbling over the complexities of some words - a single word in Arabic has several versions and extensions some are positive and others are negative. ML, on the other hand, is also good, but may need the right training set and model when entering a new domain. We trained our models on four different algorithms and on three different domains since the choice of domain is important. Depending on the circumstance of the researcher wanting a quick development time, accurate results and having a budget, there seems to be a tradeoff. Between ease of use and short development time of Lexalytics with least accuracy and working on tailoring an algorithm to student sentiment analysis, both being lexicon-based ML. If the researcher has already skills in ML, then good results can still be obtained using existing datasets and the right selection of prediction algorithm.

In order to allow other researchers to continue on work on sentiment analysis we provide access to the 122-student sentiment feedback texts at:
http://academic.cankaya.edu.tr/~james/SentimentData.xlsx

BIBLIOGRAPHY

Haddi, E. (2015). *"Sentiment analysis: Text, Pre-Processing, Reader Views and Cross Domains"* (Doctoral dissertation, Brunel University London).

Lucas Carstens, (2011), *"Sentiment Analysis - A multimodal approach"*, Master Thesis, Imperial College London

Alhumoud, S. O., Altuwaijri, M. I., Albuhairi, T. M., & Alohaideb, W. M. (2015), *"Survey on Arabic Sentiment Analysis in Twitter"*. International Science Index, 9(1), 364-368.

Khasawneh, R. T., Wahsheh, H. A., Al-Kabi, M. N., & Alsmadi, I. M. (2013). *"Sentiment analysis of arabic social media content: a comparative study"*. In Internet Technology and Secured Transactions (ICITST), 2013 8th International Conference (pp. 101-106). IEEE.

Rajput, Q., Haider, S., & Ghani, S. (2016). *"Lexicon-Based Sentiment Analysis of Teachers' Evaluation"*. Applied Computational Intelligence and Soft Computing, 2016, 1.

Retelny, D., Birnholtz, J., & Hancock, J. (2012). *"Tweeting for class: using social media to enable student co-construction of lectures"*. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion (pp. 203-206). ACM.

Dhir, A., Buragga, K., & Boreqqah, A. A. (2013). *"Tweeters on campus: Twitter a learning tool in classroom?"*. J. UCS, 19(5), 672-691.

REFERENCES

1. **Neethu, M. S., & Rajasree, R. (2013).** *"Sentiment analysis in twitter using machine learning techniques"*. In Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on (pp. 1-5). IEEE.

2. **Duwairi, R., & El-Orfali, M. (2014).** "*A study of the effects of preprocessing strategies on sentiment analysis for Arabic text*". Journal of Information Science, 40(4), 501-513.

3. **Al-Ayyoub, M., Essa, S. B., & Alsmadi, I. (2015).** *"Lexicon-based sentiment analysis of Arabic tweets"*. International Journal of Social Network Mining, 2(2), 101-114.

4. **Abdul-Mageed, M., Diab, M. T., & Korayem, M. (2011).** *"Subjectivity and sentiment analysis of modern standard Arabic"*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers. Association for Computational Linguistics.

5. **Althobaiti, M. (2016).** *"Minimally-supervised Methods for Arabic Named Entity Recognition"*, (Doctoral dissertation, University of Essex).

6. https://www.lexalytics.com/ accessed on August 15, 2017.

7. **Altrabsheh, N., Cocea, M., & Fallahkhair, S. (2014).** *"Sentiment analysis: towards a tool for analysing real-time students feedback.* In Tools with Artificial Intelligence (ICTAI)", 2014 IEEE 26th International Conference on (pp. 419-423). IEEE.

8. **Dhanalakshmi, V., Bino, D., & Saravanan, A. M. (2016).** *"Opinion mining from student feedback data using supervised learning algorithms"*. In Big Data and Smart City (ICBDSC)", 2016 3rd MEC International Conference on (pp. 1-5). IEEE.

9. **Altrabsheh, N., Gaber, M., & Cocea, M. (2013).** *"SA-E: sentiment analysis for education"*. In International Conference on Intelligent Decision Technologies (Vol. 255, pp. 353-362).

10. **Jmal, J., & Faiz, R. (2013).** *"Customer review summarization approach using Twitter and SentiWordNet"*. In Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics (p. 33). ACM.

11. **Gautam, G., & Yadav, D. (2014).** *"Sentiment analysis of twitter data using machine learning approaches and semantic analysis"*. In Contemporary computing (IC3), 2014 seventh international conference on (pp. 437-442). IEEE.

12. **Calvo, R. A., & Mac Kim, S. (2013).** *"Emotions in text: dimensional and categorical models"*. Computational Intelligence, 29(3), 527-543.

13. **Le, B., & Nguyen, H. (2015).** *"Twitter sentiment analysis using machine learning techniques"*. In Advanced Computational Methods for Knowledge Engineering (pp. 279-289). Springer, Cham.

14. **Abdulla, N. A., Ahmed, N. A., Shehab, M. A., & Al-Ayyoub, M. (2013).** *"Arabic sentiment analysis: Lexicon-based and corpus-based"*. In Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on (pp. 1-6). IEEE.

15. **L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B.Liu, 2011.** *"Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis"*, Technical report, HP Laboratories.

16. **Khudhair A. H. (2016),** *"Sentiment Analysis of Student Arabic Feedback throughout the Semester"*, Master Thesis, Çankaya University.

17. http://www.cs.waikato.ac.nz/ml/weka/ accessed on August 8, 2017.

18. **Aldayel, H. K., & Azmi, A. M. (2016).** *"Arabic tweets sentiment analysis–a hybrid scheme"*. Journal of Information Science, 42(6), 782-797.

19. **Muhammad, A. M., Kübler, S., & Diab, M. (2012).** *"Samar: a system for subjectivity and sentiment analysis of Arabic social media"*. In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis (pp. 19-28).

20. https://www.almaany.com/ar/thes/ar-ar/جيد/ accessed on August 30, 2017.

21. https://rapidminer.com/ accessed on September 21, 2017.

22. **Zhou, Z., Zhang, X., & Sanderson, M. (2014).** *"Sentiment analysis on twitter through topic-based lexicon expansion"*. In Australasian Database Conference (pp. 98-109). Springer, Cham.

23. **Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016).** *"A hybrid approach to the sentiment analysis problem at the sentence level"*. Knowledge-Based Systems, 108, 110-124.

24. http://mpqa.cs.pitt.edu/ accessed on 29 September 2017.

25. **Nasim, Z., Rajput, Q., & Haider, S. (2017).** *"Sentiment analysis of student feedback using machine learning and lexicon-based approaches"*. In Research and Innovation in Information Systems (ICRIIS), 2017 International Conference on (pp. 1-6). IEEE.

26. **Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., & González-Castaño, F. J. (2016).** *"Unsupervised method for sentiment analysis in online texts"*. Expert Systems with Applications, 58, 57-75. ISO 690.

27. **Khan, A., Baharudin, B., & Khan, K. 2011.** *"Sentiment classification from online customer reviews using lexical contextual sentence structure"*. In International Conference on Software Engineering and Computer Systems (pp. 317-331). Springer, Berlin, Heidelberg.

28. **Mountassir, A., Benbrahim, H., & Berrada, I. (2012).** *"Some methods to address the problem of unbalanced sentiment classification in an arabic context"*. In Information Science and Technology (CIST), 2012 Colloquium in (pp. 43-48). IEEE.

29. **Colas, F., & Brazdil, P. (2006).** *"Comparison of SVM and some older classification algorithms in text classification tasks"*. Artificial Intelligence in Theory and Practice, 169-178.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Omar Ahmed Nassif AL Jassem

Nationality: Iraqi

Date and Place of Birth: 28 .6.1980, Diyala, Iraq

Marital status: Married

Phone: +9647903869998

E-mail: amoory.amoory@yahoo.com

EDUCATION

Degree	Institution	Year of Graduation
M.Sc.	Çankaya University The Department of Computer Engineering Information Technology Program	2017
B.Sc.	AL Yarmouk University College	2006
High School	Hittin School	2001

WORK EXPERIENCE

Year	Place	Enrollment
2008	Ministry of Health	Programmer

FOREIGN LANGUAGES

English