International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

EFFECT OF SAUDI DIALECT PREPROCESSING ON ARABIC SENTIMENT ANALYSIS

Waad A. Al-Harbi[1] and Ahmed Emam Ph.D [1, 2]
[1]Information System Department, King Saud University, KSA
[2] Menoufia University, Menoufia, Egypt

Abstract

Over the recent years there was a rapid increase of social networking services such as Facebook and Twitter and social media websites such as YouTube. These sites and services allowed people from all over the world to express and share their opinions, likes, and dislikes, about a certain issue or a product freely and openly. Therefore, a new field known as sentiment analysis (SA) / opinion mining (OM) emerged with the goal of extracting people's sentiment from written texts. This emerging field has attracted a large research interest, but most of the existing work focuses on English text. With the availability of huge volume of Arabic opinionated posts on different social media forums, comes an increased demand for Arabic sentiment analysis tools and resources. These social media posts are usually written using dialectal Arabic (DA) and include a lot of slang. Even though, there was some work done to build modern standard Arabic (MSA) sentiment lexicons, there has been a limitation in building a dialectal Arabic lexicon especially for Saudi dialect. The main goal of this paper is studying the effect of preprocessing on Arabic sentiment analysis using Rapidminer. The collected dataset for this project will be online collected from Twitter using Twitter API.

Introduction

With the emerge of Web 2.0 technologies, Internet users started to contribute more to the Internet's contents by adding comments or opinions to the webpages using social networks. According to the Statistics Portal, Statista [4], social network sites such as: Facebook were ranked first globally at the time of conducting this work. Moreover, Google+ is ranked ninth, and Twitter ranked twelfth. Meanwhile, in the Arab countries specifically in Saudi Arabia [5], Facebook is ranked second, Twitter ranked third and Google+ is ranked fourth. Arabic language is spoken by many people in many countries. Arabs constitute around 5.3% of World population and around 4.8% of Internet users [6]. Therefore, a huge amount of data became available through social media. Thus, extracting opinions from these kinds of data is very important because people need to know other people's opinion about the service or the product they provide to enhance and gain competitive edge among their peers. A new field called Sentiment Analysis (SA) or Opinion Mining (OM) has emerged. Although these two terms (Sentiment analysis and Opinion mining) are not exactly the same, but they used interchangeably by a number of authors, where the meaning of term opinion is broader than the meaning of the term sentiment [9]. Opinion mining or Sentiment analysis can be viewed as a classification process that aims to identifying whether a certain document or text is written to express a positive or a negative opinion about a certain object (e.g., a topic, product, or person). In general, opinion mining aims to determine the attitude of a writer with respect to some topic or the overall tonality of a document. It includes several tasks, such as subjectivity detection, polarity classification, review summarization, humor detection, emotion classification, and sentiment transfer. Currently, most of the systems built for sentiment analysis are tailored for the English language but there has been some work on other languages [10]. The reason for this limitation on Arabic Sentiment analysis studies is that Arabic is a challenging language for a number of reasons [1, 11]:

1) Arabic is one of the Semitic languages, which belong to the Afro-Asiatic language family an ancient language, spoken in the Middle East and North Africa, and it is written from right to left in a cursive way.
2) Arabic language has 28 consonants, and has no upper and lower case consonants as in English.
3) Arabic has a very complex morphology relative to the morphology of other languages such as: English.
4) Arabic language is a highly inflectional and derivational language which makes monophonically analysis a very complex and difficult task.
5) Arabic opinions are highly subjective to context domains, where you may face words that have different polarity categories in different contexts.
6) Arabic Internet users mostly used dialectal Arabic rather than using MSA, where dialectal Arabic resources are scarce. The percentage of spelling mistakes within these Arabic opinions is high, and this represents an additional challenge.

Sentiment analysis is a very difficult process by itself and this process gets even harder when dealing with social media text, which is unstructured, full of spelling mistakes, and has many peculiarities and conventions. Moreover, social media text is usually short, abstract, and in many cases it is related to another text, such as a reply to or an elaboration of someone else's post. Unfortunately, the problem becomes even harder when conducting sentiment analysis on Arabic social media text. This is mainly due to the limitation in the existing natural language processing tools and resources available

ISSN:2319-7900

for Arabic language which is developed to deal with MSA only [10]. Another challenge for Sentiment analysis of Arabic is that the dialectal terms and idioms used in social media has been shown to be of a highly dynamic and evolving nature. Creative expressions that imply subjectivity are often created instantly by the social networks users and then quickly propagated and widely employed by other users they then become strong subjective clauses. Another problem is the wide use of transliterated English to reflect sentiment [2]. According to [13], subjectivity and sentiment analysis studies are classified based on:

1) Predicted class, in which the text is subjective or objective.
2) Predicted polarity, in which the text is positive, negative, or neutral.
3) Level of classification, in which (SA) for a word, phrase, sentence or a whole document.
4) Applied approach, in which supervised or unsupervised.

In the supervised approach, also known as the corpus-based or machine learning approach, machine learning classifiers such as: Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (D-Tree), K-Nearest Neighbor (KNN) are applied to a manually annotated dataset. This dataset then is split into a training set and a testing set. The classifier learns from the training dataset then can builds a model which is later used to classify the documents of the testing dataset. The system's accuracy is determined by measuring the different types of errors made by the classifier. This approach generally achieves a higher accuracy than the unsupervised approach for sentiment analysis however, it requires a large dataset that is manually labeled by a human expert. This process of manual annotation can be very difficult and time consuming even for native speakers due to different variants such as sarcasm and cultural references.

Meanwhile, the unsupervised approach, also known as the lexicon-based approach, determines the semantic orientation or the polarity of a word or a sentence based on a lexicon. In the lexicon, each word is associated with a simple polarity value (+1 for positive, -1 for negative or 0 for neutral) or with a polarity strength that can be used for positive polarities, where for an example (a word with a polarity of +5 is a much more positive word than one with a polarity of +1). This lexicon is constructed either manually or automatically. For automatic construction an initial list of seed words is given. Then, the lexicon size is increased by employing some similarity techniques. The total polarity of the sentence or the document is calculated by extracting the polarity score of each word in the text from the lexicon, and summing-up their polarities scores into one score that represents the sentiment of the whole text. The lexicon-based approach is very practical though the accuracy is often lower than the corpus-based approach.

Some research studies have combined the two approaches which they call it a weakly-supervised or a semi-supervised approach. In the article by ElSahar [11], the author used a new approach that combined the lexicon based method and machine learning methods. It passes the document from lexicon based method to two classifiers, maximum entropy and K-nearest. The justification for this is using only one approach produces a poor performance. In addition, after applying lexicon based method, the classified documents are used as training set for machine learning methods. The results show that semi-supervised approach may outperform unsupervised one. Yet, it requires adequate corpus-size to function well.

In this paper, we adapted a machine learning (supervised) approach for Arabic sentiment analysis specifically to Saudi dialect by extracting data from Twitter and studying the effect of preprocessing on the collected data using Rapidminer software.

Literature Review

Many studies have been conducted in the opinion mining / sentiment analysis field. Researchers have proposed interesting approaches and developed various systems to deal with this problem. This section present an overview of several key research papers and methods used in this field. In addition to a summary of the main work related to corpus collection and approaches used for subjectivity and sentiment analysis of Arabic language.

In [13], [14], [15], [16], the authors were creating an Arabic corpus for Subjectivity and Sentiment Analysis. Meanwhile, in [1], [3], [8], [9], [10], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], the authors were using different techniques for Subjectivity and Sentiment Analysis of Arabic as shown in details in the Table 1.

Table 1: Summary of Subjectivity and Sentiment Analysis Techniques

Paper	Copus/Dataset Size	Field	Language	SSA level	Analysis Technique
[1]	1,080 Arabic reviews	72 social media websites	MSA and DA	WL/PL	Lexicon + KNN NB
[3]	1350 comments	News websites comments	Egyptian dialect	•	SVM
[8]	2591 tweets/comments	Twitter Facebook	DA	WL	NB SVM KNN
[9]	1000 tweets	Twitter	Egyptian dialect	SL	NB SVM
[10]	1143 posts contain 8793 Arabic statements	Education, Sports, and Politics forums.	MSA	DL	Lexicon/ Maximum entropy/ KNN
[17]	1,080 Arabic reviews and comments	70 social and news sites	MSA, Egyptian, Iraqi, Jordanian,	SL	Lexicon + KNN
[18]	2855 PATB	News	MSA	SL	SVM

	sentences				
[19]	Dardasha: 2798 chat, 3015 tweets. ArabSenti lexicon3982	Chat, Twitter, Web forums and Wikipe-dia Talk Pages.	MSA and Egyptian dialect	SL	SVM light
[20]	Arabic MPQA: 9700 sentences, Microblogs: 2300 tweets	News and Twitter	MSA and DA	SL	NB
[21]	340,000 tweets	Twitter	Kuwaiti dialect	SL	SVM
[22]	GST 3,031, EBT 184,013, LBT 134,069	Twitter	MSA and DA		DS
[23]	Restaurant reviews	Reviews from qaym.co m	Saudi dialect	DL	Human Computa-tion
[24]	TAD: 28,760 Tweets , ACT: 3,448 words and phrases YT: 29,991 words	Online newswire, chat turns, Twitter tweets, and YouTube comments	MSA, Egyptian dialect and Levan-tine dialect		Lexicon/ PMI
[25]	Twitter: 10249 , Facebook: 7470, Tripadvisor: 3258	Social Netwoks and other sources	Tunisian dialect	WL	k-modes clustering to create a lexicon
[26]	ESWN, Arabic WordNet, and SAMA	WordNets	MSA and DA	WL	Lexicon based
[27]	Twitter: 2000 , Yahoo Maktoob: 2000	Social Netwoks	DA	WL	Lexicon based

WL : Word Level **PL :** Phrase Level **SL:**Sentence Level **DL :** Document Level
GST : Gold Standard Training **EBT :** Emoticon-Based Training **LBT:**Lexicon-Based Training

As shown in Table 1 most of the researches works were conducted on Modern Standard Arabic (MSA) or dialectal Arabic in general. Meanwhile, some of them were focusing on a specific slang or dialect such as: [3] and [9] in Egyptian dialect, [21] in Kuwaiti dialect, [23] Saudi dialect and [25] in Tunisian dialect. With the limited research on Saudi dialect and the fact that text from social networks such as Twitter are filled with new and constantly evolving sentiment words and idioms. There is a need to extract and analysis these sentiment words and idioms. Therefore, this project focus on Saudi Arabic text collected from Twitter

Current Techniques and Tools

This section provides a brief description of several sentiment analysis methods. These methods are the most popular in the literature (i.e., the most cited and widely used) and they cover diverse techniques such as the use of Natural Language Processing (NLP) in assigning polarity, the use of supervised machine learning techniques and unsupervised lexicon based techniques.

SentiWordNet [31] is a tool that is widely used in opinion mining, and is based on an English lexical dictionary called WordNet. These lexical dictionary groups adjectives, nouns, verbs and other grammatical classes into synonym sets called synsets. SentiWordNet associates three scores with synset from the WordNet dictionary to indicate the sentiment of the text: positive, negative, and objective (neutral). The scores, which are in the values of [0, 1] and add up to 1, are obtained using a semi-supervised machine learning method.

CrowdFlower [32] is a crowd-sourcing website that provides a data enrichment using data mining and crowdsourcing software. This software as a service platform allows users to access an online workforce of millions of people to clean, label and enrich data. CrowdFlower is typically used by data scientists at academic institutions, start-ups and large enterprises.

RapidMiner [33] is a software platform that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including results visualization, validation and optimization.

OMA project [34] Opinion Mining in Arabic is a project to develop a robust and accurate opinion mining system with special support for Arabic. The system would automatically extract people's opinions in response to a query on select topics. The project covers several aspects of Natural Language Processing (NLP), Data Mining and Analytics, namely in: Data gathering, Data Processing, and Data Visualization. An optimal approach will be introduced for the selection and modeling of features for sentiment in Arabic text. Several Arabic sentiment lexicons and corpora will be developed to support the derivation of the semantic models. The system will support different types of text sources, including social media, news, and blogs. The system will include advanced multi-dimensional visualization of opinion results. The project will also explore best visualization methods for all the gathered data and resources.

Kalimat Game [35] a sentiment analysis system that is conducted over Arabic text with evaluative content. This system uses an unsupervised technique that conducts fine-grained sentiment analysis which operates on the sentence level. Fine-grained is well known to be more informative than coarse-grained sentiment analysis which analyzes whole documents at a time. The system has two components. The first component is a game that enables users to annotate large corpuses of text in a fun manner. The game produces necessary linguistic resources that will be used by the second component which is the sentimental analyzer.

AWATIF [14] is a multi-genre corpus of Modern Standard Arabic (MSA) labeled for subjectivity and sentiment analysis (SSA) at the sentence level. This corpus is labeled using

ISSN:2319-7900

both regular as well as crowd sourcing methods under three different conditions with two types of annotation guidelines.
SAMAR [19] is a system for Subjectivity and Sentiment Analysis (SSA) for Arabic social media genres. This system use a machine learning approach based SVM light classifier. The system follows a two-stage classification approach, in the first stage Subjectivity a binary classifier to separate objective from subjective cases was built. For the second stage Sentiment a binary classification that distinguishes S-POS from S-NEG cases was applied.
SANA [24] is a large-scale, multi-genre, multi-dialect multilingual lexicon for the subjectivity and sentiment analysis of the Arabic language and dialects. This lexicon is developed manually and automatically. For the manual step, words were extracted and labelled from two different genres: SIFAAT (Arabic for "adjectives"), which is composed of 3,325 Arabic adjectives extracted from the first four parts of the Penn Arabic Treebank (PATB), and HUDA, a lexicon extracted from an Egyptian Arabic chat data set. The automatic step is articulated using two main methods, a statistical method based on pointwise mutual information (PMI), and another method based on simple machine translation.
TunDiaWN [25] Tunisian dialect Wordnet is a lexical resource for the dialect language spoken in Tunisia. This lexicon construction approach is founded, in one hand, on a corpus based method to analyze and extract Tunisian dialect words. A clustering technique is adapted and applied to mine the possible relations existing between the Tunisian dialect extracted words and to group them into meaningful groups.

Data Collection

The dataset was generated by collecting tweets from Twitter. These tweets address the topic of sports in girls' education in Saudi Arabia. Using Rapidminer tweets were pulled using Search Twitter operator. This operator allows creating search queries exactly as when using the Twitter search API. A search query was created that searches for the keywords "الرياضة+للبنات+مدارس+البنات". In addition to setting the language parameter to Arabic using "ar" to make sure the search only pulls in Arabic tweets. Initially, 2000 Tweets were collected therefore; keywords were changed to different trending Saudi hashtags which expanded the topic into more general areas such as education, sports, and political news. Almost 5,500 tweets were collected and manually annotated. When the collected tweets were examined for annotation, the tweets were suffering from several problems. They include high number of duplicate tweets which may be the result of re-tweeting, also some of the collected tweets were empty and contain the address of the sender only. Such tweets were removed from our dataset. Table 2 shows the number of tweets that remained with their sentiment orientation.

Table 2:Number of positive,negative and neutral Tweets in the dataset

	Negative	Positive	Neutral	Total
Number of Tweets	2,434	1,415	1,634	5,500

Preprocessing

After the collection and annotation of the dataset, a preprocessing step must be done before sentiments classification. In the preprocessing step, multiple tasks must be done to ensure the cleaning of data and removing noise that may affect the classification accuracy. Here, the Rapidminer Tokenize, Stem (Arabic), Stem (light, Arabic), Filter Tokens (by Length), Filter Stopwords (Arabic), and Generate-n-Grams (Terms) operators were used.
The Tokenize operator is responsible for splitting the text into tokens or words which will lead to removing non letters characters such as *, #, {}, …. etc. The Filter Tokens by Length operator is responsible of removing short words such as single letters. The Filter Stopword (Arabic) operator removes noise Arabic words that do not affect the classification task such as ((في ,من ,التي ,الذي etc. The problem when using this operator is that negation words are considered stopwords and for that they are removed. This would cause a huge problem for sentiment analysis as the negation words can reverse the sentiment from positive to negative and vice versa.
The Generate-n-grams operator can slightly help in the negation words removal by generating sequences of n-words and each sequence is considered one token. N, specifies the number of words or terms in a sequence.
The Stem(Arabic) operator is responsible for reducing an Arabic token to its stem or root. RapidMiner Arabic Stemming Algorithm Steps: (Motaz K. Saad & Wesam Ashour, 2010)
1.Remove diacritics
2.Remove stopwords, punctuation, and numbers.
3.Remove definite article (ال)
4.Remove inseparable conjunction (و)
5.Remove suffixes
6.Remove prefixes
7.Match result against a list of patterns. If a match is found, extract the characters in the pattern representing the root.
8.Match the extracted root against a list known "valid"‖ roots
9.Replace weak letters ي ا و with و
10. Replace all occurrences of Hamza ئ ء ؤ with ا
11. Two letter roots are checked to see if they should contain a double character. If so, the character is added to the root.

Rapidminer also has another operator called Stem (Arabic, Light). This operator does not reduce a word to its proper root but it removes common prefixes and suffices from

words or tokens. RapidMiner Arabic Light Stemming Algorithm Steps: (Motaz K. Saad & Wesam Ashour, 2010)

1.Normalize word:

- Remove diacritics
- Replace أ إ آ with ا
- Replace ة with ه
- Replace ى with ي

2.Stem prefixes:

- Remove Prefixes: ال، وال، بال، كال، فال، لل، و

3.Stem suffixes:

- Remove Suffixes: ها، ان، ات، ون، ين، ية، ه، ي

Sentiment Classification

After data preprocessing the data will be classified using three supervised machine learning models for sentiment classification. These models are Naïve Bayes classifier (NB), Support Vector Machine classifier (SVM) and K-Nearest Neighbor classifier (KNN). In this work, a 10-fold cross validation was used. This means that, in the 10-fold cross validation method, the data is divided into 10 divisions or parts; one is used for testing and 9 for training in the first run. In the second run, a different part is used for testing and 9 parts, including the one that was used for testing in run one, are used for training. The runs continue until each part or division is given the chance to be part of the training data and the testing data. The final accuracy is the average of the accuracies obtained in the 10 runs.

Performance Metrics

To calculate the performance measures we need to calculate the following metrics:

- TP (True Positive): the number of tweets that were correctly classified by the classifier to belong to the current class.
- TN (True Negative): the number of tweets that were correctly classified by the classifier not to belong to the current class.
- FP (False Positive): the number of tweets that were mistakenly classified by the classifier to belong to the current class.
- FN (False Negative): the number of tweets that were mistakenly classified by the classifier not to belong to the current class.
- Classification Accuracy: is the number of correct predictions made divided by the total number of predictions made, multiplied by 100 to turn it into a percentage.
- Precision: is the number of True Positives divided by the number of True Positives and False Positives. In another way, it is the number of positive predictions divided by the total number of positive class values predicted. It is also called the Positive Predictive Value (PPV). Precision = TP/(TP+FP).
- Recall: is the number of True Positives divided by the number of True Positives and the number of False Negatives. In another way it is the number of positive predictions divided by the number of positive class values in the test data. It is also called Sensitivity or the True Positive Rate (TPR). Recall = TP/ (TP + FN).
- F1 Score: The F1 Score is the 2*((precision*recall)/ (precision + recall)). It is also called the F Score or the F Measure. In another way, the F1 score conveys the balance between the precision and the recall.

Sensitivity: Also called the true positive rate (TPR), or the recall, it's measures the proportion of positives that are correctly identified as such TPR = TP/ (TP + FN).

Specificity: Also called the true negative rate, it's measures the proportion of negatives that are correctly identified as such. SPC = TN/ (TN + FP)

Experiments Analysis

The collected and annotated data set (5,484 tweets) were applied to the experiments. 66% of the data (3600 tweets) were used in the learning process while, the remaining 33% (1,810 tweets) were used in the testing process. In the learning process the dataset was trained using the three classifiers (NB, SVM and KNN) each under three conditions. Once with No Stemming operator, second with Light Stemming operator and third with Stemming operator. Meanwhile, in the testing process the unlabeled dataset is preprocessed under the same conditions, and the learned classifiers were applied. In addition, the performance measures (accuracy, precision and recall) were calculated from both training and testing processes. Below, are the results of the conducted experiments.

Table 3: Summary of Training Accuracy

	No-Stem	Light-Stem	Stem
NB	54.62%	56.25%	52.44%
KNN	56.98%	58.10%	59.22%
SVM	56.55%	58.15%	57.45%

Table 4: Summary of Testing Accuracy

	No-Stem	Light-Stem	Stem
NB	29.45%	29.06%	29.56%
KNN	37.07%	36.24%	34.81%
SVM	36.30%	35.91%	36.96%

Results Enhancement

In this paper several attempts to improve the performance of the learning and the testing process were conducted. This attempts started by replacing the Saudi dialect specific expressions or terms that was manually extracted from the dataset with its meaning in Modern Standard Arabic. This was done using Replace Tokens operator. This operator Re-

ISSN:2319-7900

places all occurrences of all specified regular expression within each token by its specified replacement.

Table 5: Extracted Saudi word list

Saudi Word	Arabic	English
مافيه	لا يوجد	Nothing
تتحسف	تندم	Regret
سطر - طق	ضرب	Hit
اش معنا – وراه - وشو له - وشهوله	لماذا	Why
ماسوو	لم يفعلوا	Did not
ودنا	نريد	We want
كفوو	أحسنت	Well done
معليش	اعذرني	Excuse me
موب - مب	ليس	Not
فارق - انقلع	اذهب	Leave
يهرجك	يكلمك	Talks to you
بينقون	يزعجون	Bother
تفوه	بصق	Spit
بفل امها	استمتع	Enjoy
الزبدة	الخلاصة	Conclusion
الخرط	الكذب	Lie
هرج	الكلام	Talk
سيدا	الى الامام	Forward
بغينا	كنا	We were
نفطس	نموت	Die
طفت	اغلقت	Closed
حستو	حاول أن يعرف	Try to figure out
معد في - مافي	لا يوجد	No more
تنكه	غبي	Stupid
يشحتون	يتسولون	Beg for money
اخصريه	لا تهتم	Don't bother
يخسى ويعقب - معصي	مستحيل	No way
من جدك	هل انت جاد	Are you serious
تبين – تبيه - تبغا	تريد	You want
ابي	اريد	I want
طفت	اغلقت	Closed
حستو	حاول أن يعرف	Try to figure out
بعزقت	صرفت	Spent
الخياس	رائحة كريهه	Stinky smell
اشلون- شلون	كيف	How
وهقني	وضعني في مأزق	Dilemma
شحاتين	متسولين	Beggars
البزر - العيّل	الطفل	The kid
هياطه – الهياط – هياطكم	تضخيم الامور	Magnifies things
خل عنك	اتركه	Leave it
دوبهم	الان	Just now
ملينا	نشعر بالملل	We feel bored
ذولا - هذول	هؤلاء	Those
صح لسانك - ونعم	مدح	Endearment expression
بعَدي	عزيزي	My dear
وش جاك	ماذا حدث لك	What happened to you
فالحين	تقدر على	Can do
حنا	نحن	We
زين	جيد	Good
مروق	مزاج جيد	Good mood
زي	مثل	Such as
معطيها وجه	اعطاها اكثر مما تستحق	Give it more than deserve
ولا فيني	ليس بي	Not me
فزعه	النجدة	Help
الحالي	وحيد	Alone
ترا	انظر	I see
يالنشاما	الناس الطيبين	Good people
حبيت	اعجبني	I liked
كرشوهم	طردتهم	Kick them out
طز فيكم	شتم	Cursing
وش دخل	لا يتضمن	Does not include
مفتح	مدرك	Aware
يحط	يضع	Put
شيك	انظر	Look
حسينا بوه	شعرنا به	We felt it
الخواجات	الاجانب	Foreigners
ينبسط	يفرح	Rejoice
يتزبط	يتزين	Perk
يشلع	يسلب	Steal
لجل	لكي	In order to
توه طالع	خرج للتو	Just went out
افلقني	اضربني	Hit me
واجد - بالحيل	كثير	Many
زود	زيادة	Increase
جاك	أتى اليك	Come to you
انحش	اهرب	Run
يع	قرف	Ugh
يستعبطون	سخرية	Ridicule
مغير	فقط	Just
يالبله	غبي	Dumb
وش زين	ما اجمله	How nice
ياساتر	يا الهي	Oh my god
مانبي	لا نريد	Don’t want
شويه	قليل	Little
ياحليلهم	لطفاء	How nice

Results and discussion

In these experiments three supervised machine learning models of Naive Bayes, SVM and KNN have been applied to a set of collected tweets. In addition, each of these models was tested under three preprocessing conditions of no stemming, Rapidminer Arabic light stemmer and Rapidminer Arabic stemmer. The observation shows that the machine learning models of KNN with no stemming used and SVM using Rapidminer Arabic stemmer achieved the highest accuracy in 37.07% and 36.96% respectively as shown in Table 4. The best recall and precision was achieved by KNN classifier using no stemming at 84.25% and 44.00% respectively. When the training data set was as small as 150 or 300 tweets, the SVM classifier achieved the highest accuracy followed by the Naïve Bayes classifier who outperformed KNN.

One of the most powerful techniques for building highly accurate classifiers is using ensemble learning and combining the results of different classifiers. This means that using several different classifiers that focus on different areas can help us build strong high-accuracy classifiers. Unfortunately, in text analysis this is not as effective. Therefore, the accuracy of grouping the three classifiers was low at 29.59%.

Furthermore, there was an attempt to improve the accuracy of the machine learning models in extracting the domain specific terms which are the Saudi dialect words as shown in Table 5 and replacing them with Modern Arabic Standard. When applying this on a large data set of 1800 tweets the results weren't so much noticeable. The SVM classifier score a 38.45% accuracy while, the accuracy of KNN classifier was 37.25%. Meanwhile, when the training data set was small as 100 tweets, it was first tested without replacing the Saudi dialect words. The resulted accuracy of both SVM classifier and KNN classifier were 35% and 42% respectively. Then, the training data set was tested again this time replacing the Saudi dialect words. The resulted accuracy of both SVM classifier and KNN classifier were 42% and 48% respectively. This shows a significant improvement in both classifiers after replacing the dialect words. the best recall and precision was achieved by KNN classifier at 73.33% and 57.89%.

Conclusion and Future work

In this research the general goal is to study the effect of preprocessing on Arabic Sentiment analysis specifically on the dialect of Saudi Arabic. Twitter was used as the source of the collected data because of the short nature of tweets and its richness with slang language. All experiments were conducted using Rapidminer and Text processing extension. In these experiments, three supervised machine learning models of SVM, Naive Bayes and KNN were compared for sentiment classification of tweets that contains 2,434 negative, 1,415 positive and 1,634 neutral processed tweets. Each of these classifiers were tested under three preprocessing conditions of using no stemming, using Rapidminer Arabic light stemmer and using Rapidminer Arabic Stemmer. The experimental results show that the SVM using stemming classifier and KNN with no stemming classifier outperformed the Naive Bayes classifier. In addition, to improve the classifiers accuracy a Replace tokens approach was made. This approach is concerned with extracting the Saudi dialect terms or words and replacing them with Modern Standard Arabic words. The result of this approach proved valid in improving the accuracy of the classifiers when tested on a small data set, this was due to the limited extracted words there were only 114 words extracted.

There were some limitations in this project such as limited hardware processors available and manually annotated data. In the future, this work should be done with high processing computers that could accommodate a large scale dataset. In addition, it will be helpful using crowdsourcing for manual annotation or using a tool for semi-annotation then deploying and using Saudi dialect lexicons.

References

[1] Al-Kabi, M., Gigieh, A., Alsmadi, I., Wahsheh, H., & Haidar, M. (2014). Opinion Mining and Analysis for Arabic Language. International Journal of Advanced Computer Science and Applications (IJACSA), 5(5, 2014), pp. 181-195.

[2] Abdulla, N. A., Ahmed, N. A., Shehab, M. A., & Al-Ayyoub, M. (2013, December). Arabic sentiment analysis: Lexicon-based and corpus-based. In Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE (pp. 1-6).

[3] A. Soliman, T., Elmasry, M., Hedar, A., & Doss, M. (2013). Mining Social Networks' Arabic Slang Comments. In Proceedings of IADIS European Conference on Data Mining (ECDM), pp. 328–333.

[4] Global Social Networks Ranking. http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. (2015, accessed April 2015).

[5] Saudi Arabia Social Networks Penetration. http://www.statista.com/statistics/284451/saudi-arabia-social-network-penetration/. (2014, accessed April 2015)

[6] Arabic Speaking Internet Users and Population Statistics. http://www.internetworldstats.com/stats5.htm. (2014, accessed April 2015).

[7] Varieties of Arabic. http://en.wikipedia.org/wiki/Varieties_of_Arabic. (Accessed April 2015).

[8] Duwairi, R., & Qarqaz, I. (2014). Arabic Sentiment Analysis using Supervised Classification. The 1st International Workshop on Social Networks Anal-

ysis, Management and Security (SNAMS), pp.1-10.

[9] Shoukry, A., & Rafea, A. (2012). Sentence-level Arabic sentiment analysis. International Conference on Collaboration Technologies and Systems (CTS), pp. 546-550.

[10] El-Halees, A. (2011). Arabic Opinion Mining Using Combined Classification Approach. International Arab Conference on Information Technology (ACIT), pp. 264-271.

[11] ElSahar, H., & El-Beltagy, S. R. (2014). A fully automated approach for Arabic slang lexicon extraction from microblogs. In Computational Linguistics and Intelligent Text Processing (pp. 79-91). Springer Berlin Heidelberg.

[12] Azmi, A. M., & Alzanin, S. M. (2014). Aara'–a system for mining the polarity of Saudi public opinion through e-newspaper comments. Journal of Information Science, 40(3), pp. 398-410.

[13] Mubarak, H., & Darwish, K. (2014). Using Twitter to Collect a Multi-Dialectal Corpus of Arabic. Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pp. 1–7.

[14] Abdul-Mageed, M., & Diab, M. (2012). AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), pp. 3907-3914.

[15] Jarrar, M., Habash, N., Akra, D., & Zalmout, N. (2014). Building a Corpus for Palestinian Arabic: A Preliminary Study. Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pp. 18–27.

[16] Refaee, E., & Rieser, V. (2014). An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), pp. 2268-2273.

[17] Al-Kabi, M., Gigieh, A., Alsmadi, I., Wahsheh, H., & Haidar, M. (2013). An Opinion Analysis Tool for Colloquial and Standard Arabic. In Proceedings of the Fourth International Conference on Information and Communication Systems (ICICS), pp. 1-5.

[18] Abdul-Mageed, M., Diab, M., & Korayem, M. (2011). Subjectivity and Sentiment Analysis of Modern Standard Arabic. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 587–591.

[19] Abdul-Mageed, M., Kuebler, S., & Diab, M. (2012). SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social Media. Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, pp. 19–28.

[20] Mourad, A., & Darwish, K. (2013). Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs. Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 55–64.

[21] Ben Salamah, J., & Elkhlifi, A. (2014). Microblogging Opinion Mining Approach for Kuwaiti Dialect. Proceedings of the International Conference on Computing Technology and Information Management, pp. 388-396.

[22] Refaee, E., & Rieser, V. (2014). Evaluating Distant Supervision for Subjectivity and Sentiment Analysis on Arabic Twitter Feeds. Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pp. 174–179.

[23] Al-Subaihin, A., & Al-Khalifa, H. (2014). A System for Sentiment Analysis of Colloquial Arabic Using Human Computation. The Scientific World Journal, pp. 1-8.

[24] Abdul-Mageed, M., & Diab, M. (2014). SANA: A Large Scale Multi-Genre, Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), pp. 1162-1169.

[25] Bouchlaghem, R., Elkhlifi, A., & Faiz, R. (2014). Tunisian dialect Wordnet creation and enrichment using web resources and other Wordnets. Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pp. 104–113.

[26] Badaro, G., Baly, R., Hajj, H., Habash, N., & El-Hajj, W. (2014). A Large Scale Arabic Sentiment Lexicon for Arabic Opinion Mining. Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pp. 165–173.

[27] Abdulla, N., Ahmed, N., Shehab, M., Al-Ayyoub, M., Al-Kabi, M., & Al-rifai, S. (2014). Towards Improving the Lexicon-Based Approach for Arabic Sentiment Analysis. International Journal of Information Technology and Web Engineering, pp. 55-71.

[28] Ahmed, S. G. K. (2014). Sentiment Mining of Arabic Twitter Data (Doctoral dissertation, American University of Sharjah).

[29] Abdulla, N., Majdalawi, R., Mohammed, S., Al-Ayyoub, M., & Al-Kabi, M. (2014, August). Automatic lexicon construction for arabic sentiment analysis. In Proceedings of the 2014 International

ISSN:2319-7900

Conference on Future Internet of Things and Cloud (pp. 547-552). IEEE Computer Society.

[30] Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., & Perea-Ortega, J. M. (2011). OCA: Opinion corpus for Arabic. Journal of the American Society for Information Science and Technology, 62(10), pp. 2045-2054.

[31] SentiWordNet. http://sentiwordnet.isti.cnr.it/. (Accessed May 2015).

[32] People-powered Data Enrichment Platform. http://www.crowdflower.com/. (2007, Accessed May 2015).

[33] Predictive Analytics, Data Mining, Self-service, Open source - RapidMiner. https://rapidminer.com/. (2001, Accessed May 2015)

[34] Opinion Mining in Arabic project. http://oma-project.com/. (Accessed May 2015).

[35] لعبة كلمات وبالونات. http://kalimat.afnan.ws/. (2011, Accessed May 2015).

[36] Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. International Journal, 2(6).

[37] Kalaivani, P., & Shunmuganathan, K. L. (2013). "Sentiment Classification of Movie Reviews by Supervised Machine Learning Approaches. Indian Journal of Computer Science and Engineering (IJCSE), 4(4), 285-292.

[38] Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. Expert Systems with Applications, 34(4), 2622-2629.

[39] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.

[40] Machine Learning Blog & Software Development News.http://blog.datumbox.com/10-tips-for-sentiment-analysis-projects/. (2013, accessed December 2015).

[41] Saad, M. K., & Ashour, W. (2010). Arabic morphological tools for text mining. Corpora, 18, 19.

Biographies

WAAD A. AL-HARBI received the B.S. degree in Information Technology from the University of King Saud, Saudi Arabia, Riyadh, in 2010, the M.S. degree in Information Systems from the University of King Saud, Saudi Arabia, Riyadh, in 2016. Author may be reached at waad.abdulkareem@hotmail.com.

Dr. AHMED EMAM, Associate Professor of Information Systems at the College of Computer Science and Information Systems in King Saud University has more than 13 years of teaching and research experience. Emam is an Associate Professor of Computer Science at Menoufia University, Egypt. His area of interest and research is Database System, Data Mining, ERP, Artificial Intelligence, and Big data Analytics. Emam received his B.S. from Ain Shames University in Cairo, Egypt. Emam received his M.S. and Ph.D. from Computer Science and Computer Engineering department-Speed School at University of Louisville, Kentucky-USA. He has published and presented papers at several academic journals and international conferences.