An Opinion Analysis Tool for Colloquial and Standard Arabic

Mohammed Al-Kabi
Faculty of Sciences & IT
Zarqa University
malkabi@zu.edu.jo

Amal Gigieh
Faculty of Sciences & IT
Ajloun- College
AL-Balqa' Applied University
amalqiqi@yahoo.com

Izzat Alsmadi
CIS Department
IT & CS Faculty
Yarmouk University
ialsmadi@yu.edu.jo

Heider Wahsheh
CIS Department
IT & CS Faculty
Yarmouk University
heiderwahsheh@yahoo.com

Mohamad Haidar
Web development Department
Brandtologie company
mohammad_haidar@rocketmail.com

ABSTRACT
Social networks and users' interactions are distinct features for the current Web. They constitute a fundamental part of Web 2.0, where people produce, disseminate, and consume information in new interactive forms where users are not only passive information receivers. Social media succeed to attract a large portion of online users, which explains the explosive growth of social media in terms of comments, reviews, blogs, microblogs, Twitters, and postings in social network sites. In this scope, sentiment analysis research field refers to the analysis of people's sentiments, opinions, attitudes, and emotions towards events, products, companies, individuals, issues, sport teams ...etc. Facebook, and YouTube are within the top 3 sites used in many Middle Eastern (ME) countries, and the world. Therefore a huge volume of Arabic comments and reviews are generated daily about different aspects of life in this part of the world. Modern Standard Arabic (MSA) is used mainly in media (Newspapers, Journals, TV and Radio), academic institution, and to some extent in social media. While colloquial Arabic is used by the public in their conversations, chatting, etc.. Analysis of social networks in ME countries shows that both MSA and colloquial or slang languages are used. The aim of this study is to build a novel sentiment analysis tool called colloquial Non-Standard Arabic - Modern Standard Arabic-Sentiment Analysis Tool (CNSA-MSA-SAT) dedicated to both colloquial Arabic and MSA. A large number of Arabic collected comments and reviews from social media were tokenized and analyzed to build polarity lexicons which constitute an essential part of CNSA-MSA-SAT. Each Arabic collected comment and review is manually assigned to one of the three polarity values: (positive, negative, and neutral). Further, each collected review or comment is added to CNSA-MSA-SAT and is then assigned to one of the three polarities values based on algorithms developed for this purpose.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *Information filtering*. I.2.7 [**Natural Language Processing**] – *Text analysis*

ICICS'13, April 23–25, 2013, Irbid, Jordan.

General Terms
Algorithms, Experimentation.

Keywords
Opinion mining, sentiment analysis, Arabic Sentiment Analysis, Arabic Information retrieval.

1. INTRODUCTION
During mid-nineties of the twentieth century the Arab world witnessed the emergence of Internet era, and the Internet started to be used in countries such as: Jordan, Gulf States, Syria, Lebanon, and North Africa. Internet users at that time witnessed the dot com boom, and then the collapse and bankruptcy of many of the dot com companies. This era is followed by the emergence of Web 2.0 which represents the second generation of World Wide Web (WWW). Web 2.0 enables its users to generate and share information, and interact and collaborate with each other in virtual societies.

The emergence of the second generation of WWW, leads to the accumulation of a huge number of reviews and comments on a daily basis. This huge number of reviews needs to be analyzed and mined automatically through a new field of research called Sentiment analysis (Opinion mining). Sentiment analysis is essential to a number of purposes. It can be used by companies to discover the opinions of different customers about different products as part of customer service or marketing purposes. This is essential to many companies to stay competitive and survive bankruptcy. It can also be used to predict sales performance and election results. In addition, social media reviews can be analyzed to rank different products and merchants, and to know public opinions about different policies.

Arabs constitute these days around 5 % of world population, and Arabic speaking Internet users constitute these days around 3.8 % of Internet users worldwide, where around 40% of those Internet users use Facebook alone according to [1]. Therefore the Arab users of social media generate a huge number of comments and reviews using both colloquial Arabic and Modern Standard Arabic (MSA). These Arabic comments and reviews can be analyzed and mined to discover the knowledge within it. To analyze and mine those Arabic comments and

reviews, sentiment analysis is used to extract opinions and determine their polarity (Positive, Negative, Neutral, or Spam). During the last 10 years the world witnessed an increasing interest in this field and there are many studies in this field, but there are very few studies related to analysis and mining of reviews and comments written in colloquial Arabic and MSA.

Several sentiment analysis studies adopt the techniques used in information retrieval (IR), text mining, and natural language processing (NLP). There are keyword-based sentiment analysis methods, and dictionary-based sentiment analysis methods [2, 3], to determine the polarity of each comment or review. Arabic language (MSA) is most widely used Semitic language worldwide. It's written from right to left and highly inflectional language, so its grammar differs from for example the English language. Colloquial Arabic lacks standardization and has no grammar, or references to show the semantics and syntax of this style of Arabic language. This makes sentimental analysis a complex process in Arabic.

In this study we designed and implemented a new tool called CNSA-MSA-SAT to determine the polarity of each inputted review and comment written in colloquial Arabic and MSA. In addition, 18 specialized polarity lexicons for both colloquial Arabic and MSA are built.

The remainder of this study is organized as follow: Section 2 presents an overview to the related work to sentiment analysis, while section 3 presents the proposed algorithm. Section 4 presents the experiments, including the collected dataset and results of the tests conducted on our new algorithm, while section 5 presents conclusion remarks and future work.

2. RELATED WORK

In this section, we will present studies related to the subject of this paper with focus on social networks especially in Middle East countries.

The study of [4] proposed a number of sentiment analysis methodologies to classify Web forum opinions written in different natural languages. Tests on these sentiment analysis methodologies were conducted on three datasets; the first one contains reviews about movies, while the two other datasets are related to two hate/extremist-groups forum postings. Special feature extraction components were integrated to these methods to compute the linguistic characteristics of Arabic, besides developing an entropy weighted genetic algorithm (EWGA) for feature selection. Test Results showed the effectiveness of those methodologies to be used for sentiment analysis in multiple languages. Support vector machines (*SVM*) algorithm yields a high level of accuracy to classify (identify polarity of) different sentiments.

A hybrid approach of three methods is used by [5] to extract opinions automatically from Arabic text. In the first step a manually built lexicon is used to classify those opinions. The classified opinions are used as a training set for maximum entropy method which subsequently classifies some other documents. In the final stage the researcher used *K*-Nearest Neighbor (*KNN*) method to classify documents which contain opinions. Using the three methods lead to enhancing the effectiveness of classification to 80%.

In their study [6], the authors studied Subjectivity and Sentiment Analysis (SSA) and developed an automatic sentence-level SSA tagging system, and a polarity Arabic lexicon, besides manually annotating a corpus of Modern Standard Arabic (MSA). Their corpus contains a collection of newswire documents, annotated on sentence level. They showed that the Stem lemmatization setting outperforms both Surface and Lemma settings for the SSA. The paper [7] conducted another study where they explore using a number of techniques related to subjectivity and sentiment analysis (SSA) for Arabic language.

The study of [8] collected a collection of Arabic reviews about movies. Then they used machine translation to convert the Arabic text into English. Afterward a number of classification algorithms were used to determine the polarity of the translated (English) reviews. They discover that the translation process leads to a deterioration of the effectiveness of determining the polarity of each review. T hey also collected an Arabic corpus for sentiment analysis in the study of [9]. The collected corpus contains 500 movie reviews, where the polarity of these Arabic reviews is divided equally between positive and negative. To identify the polarity of each review two classification algorithms (*SVM*) and Naïve Bayes were tested. Those classifiers yield satisfactory results.
The study of [10] presents a system to extract business Arabic reviews, and then it analyzed these collected reviews to identify their polarity (positive, negative or neutral), and exhibits the general opinion of the Arab public about different products and services.

A lexicon-based sentiment analysis tool dedicated to colloquial Arabic text used in chatting and daily conversation and within social media is designed and implemented by [11]. Those researchers propose that their tool should rely partially on human judgment to overcome the problem arise from using non-standardized colloquial Arabic text. An independent component of the proposed tool is game-based lexicons which are based on human expertise; to overcome problems arise from using non-standardized colloquial Arabic text.

3. ALGORITHM

This section exhibits the main steps followed to build our opinion analysis tool (CNSA-MSA-SAT) to determine the polarity of each review whether it used colloquial Arabic or MSA.

Figure 1 shows the pseudo-code of CNSA-MSA-SAT algorithm. CNSA-MSA-SAT tool is capable to identify the polarities of different colloquial Arabic and MSA reviews and comments, and tag them with one of R_1 set three polarity values (Positive/Negative/Neutral). In addition, CNSA-MSA-SAT output R_2 and R_3 to determine percentage of positive and negative for each input respectively.

Algorithm: CNSA-MSA-SAT**.**
Input:
***R*:** Review/Document Text
***T*:** the set of the Opinion tokens
***PD*:** the set of Positive Sentiment Dictionary
***ND*:** the set of Negative Sentiment Dictionary
Output:
$R_1=\{P, N, NT\}$, *where P*: *Positive, N*: *Negative, NT*: *Neutral*

$$R_2 = \frac{\sum_{i=1}^{n} Pos - TF_i}{\left(\sum_{i=1}^{n} Pos - TF_i\right) + \left(\sum_{i=1}^{n} Neg - TF_i\right)} \times 100$$

$$R_3 = \frac{\sum_{i=1}^{n} Neg - TF_i}{\left(\sum_{i=1}^{n} Pos - TF_i\right) + \left(\sum_{i=1}^{n} Neg - TF_i\right)} \times 100$$

Initialization:

$Pos - TF = 0$, *where* $Pos - TF$ *is the term frequency for positive sentiments*

$Neg - TF = 0$, *where* $Neg - TF$ *is the term frequency for negative sentiments*

Begin

1: *For each* t_i Č *T do*
2: *Search for* t_i *in PD where* t_i Č *T*
3: *If* t_i Č *PD then*
4: *Pos-TF* ← *Pos-TF +1*
5: *Else*
6: *Search for* t_i *in ND where* t_i Č *T*
7: *If* t_i Č *ND then*
8: *Neg-TF* ← *Neg-TF +1*
9: *End For*
10: *If* (*Pos-TF* ≥ 2) && (*Pos-TF* > *Neg-TF*) *then*
11: R_1 = *P*
12: *Return* R_2
13: *End If*
14: *If* (*Neg-TF* ≥ 2) && (*Pos-TF* < *Neg-TF*) *then*
15: R_1 = *N*
16: *Return* R_3
17: *End If*
18: *If* (*Pos-TF* == *Neg-TF*) && (*Pos-TF* != 0)*then*
19: R_1=*NT*
20: *Return* R_1
21: *End If*

End

Figure 1. CNSA-MSA-SAT Algorithm.

4. EXPERIMENTS

This section presents the size and contents of the collected dataset. Further, preprocessing steps on the collected comments and reviews are shown in this section. Collected comments and reviews have to be preprocessed before inputting them to CNSA-MSA-SAT tool. This section also exhibits an example showing the construction of lexicons used in this study. The evaluation of the CNSA-MSA-SAT tool is shown at the end of this section.

4.1 Datasets and Preprocessing

This research is based on 1,080 Arabic reviews and comments. The collected Arabic reviews and comments use Egyptian, Iraqi, Jordanian, Lebanese, Saudi, and Syrian dialects only collected from around 70 social and news sites. Comments and reviews having Latin letters [(English, French...etc) or transliterated Arabic words like momtaz which means excellent in English] were deleted. Also comments used Arabizi (Arabic chat alphabet) were deleted. That means our dataset (1,080 Arabic reviews) contains reviews using only Arabic alphabets.

Afterward preprocessing steps began with the removal of punctuations and non-alphabets, besides normalizing some Arabic alphabets. As shown below:

1. Remove digits, punctuations, symbols, marks, and non-letters.
2. Normalizing of (أ،ا،آ،إ،ٱ) to (Alif, ا).
3. Normalizing of (Yaa ي، a'ىء) to (Yaa, ي).
4. Normalizing of (ة Taa' MarbuuTa, Final haa ه) to (Final haa', ه).
5. Normalizing of ("و Waaw , ؤ O'") to (Waaw,و).
6. Filtering the main text from non-Arabic text.
7. Tokenization.

Afterward these Arabic reviews and comments are classified manually into eight domains: Books, Movies, Places, Politics, Products, Social, Technological, and Educational. Then a feature extraction is started by extracting domain features, sentiment (Positive/Negative) features.

As an example, consider the following sample in the next excerpt, of the collected Arabic reviews with its English translation:

Arabic Comment	فيلم رائع جدا وممتع استمتعوا بمشاهدة هذا الفيلم و هو يعتبر نقلة **في صناعة** هوليود الفيلم خليط من أفلام زي ماتريكس و سبيدرمان
English Translation	"A wonderful and enjoyable movie. Enjoy watching this movie, it is considered as a shift in the Hollywood industry. It mixed between movies like Matrix and Spiderman "

Table 1 exhibits the domain features of above Arabic excerpt. This sample is taken from the domain of movies. Domain features are used to determine the domain of each Arabic review within the dataset. To be more specific, a Naive Bayes classifier is used in this study to determine the domain of each Arabic review and Arabic comment.

Table 1. Domain Features

أفلام	الفيلم	فيلم	هوليود
Movies	The **movie**	**Movie**	Hollywood

Table 2 exhibits a sample of positive polarity features extracted from the above Arabic sample review. These features are stored in the polarity lexicon to be used later by the tool to determine the polarity of different Arabic reviews and comments.

Table 2. Positive Polarity Features

ممتع	رائع
Enjoyable	Wonderful

Around 1,080 Arabic reviews and comments were used to create two general lexicons (that contain 8,000 terms or sentiments). First general lexicon is dedicated to positive polarity, and the second general lexicon is dedicated to negative polarity. In addition, for each of the eight domains two lexicons were created one for positive polarity and another for negative polarity. Therefore this study is based on 18 lexicons, where 2 are general and the other 16 are domain -based lexicons.

4.2 Results

The results of the conducted tests using the proposed tool on the collected dataset show that the proposed tool (CNSA-MSA-SAT) yields more accurate results on the eight domains (Books, Movies, Places, Politics, Products, Social, Technological, and Educational) of the reviews, relative to the results yield by CNSA-MSA-SAT on general reviews and comments.

To evaluate the polarity identification accuracy of our sentiment analysis tool (CNSA-MSA-SAT), we adopted IBK (*KNN*) Classifier. The evaluation was on a sentence (review) based level. The evaluation results of the accuracy of determining the polarity was 90% yields when *K*=1, with a 10% error rate. We used the following four metrics to measure the quality:

Accuracy: is the degree of closeness that a measured value represents the correct value.

The Accuracy is defined by the formula (4.1):

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{(4.1)}$$

Where *TP* is a true positive rate, *FP* is a false positive rate, *TN* is a true negative rate, and *FN* is a false negative rate [12].

Error: is the degree of closeness that a measured value represents the incorrect value [12].

The formulas of the other two performance metrics (Recall and Precision) are shown next.

The Recall is defined by formula (4.2) [13]:

$$Recall_i = \frac{TP}{TP + FN} \quad \text{(4.2)}$$

The Precision is defined by the formula (4.3) [13]:

$$Precision_i = \frac{TP}{TP + FP} \quad \text{(4.3)}$$

where *TP* is the number of documents correctly classified as belonging to a class *i* (“true positive”), *FP* is the number of documents falsely classified as belonging to a class *i* (“false positive”) and *FN* is the number of documents falsely classified as not belonging to a class *i* (“false negative”) [13].

Table 3 presents a summary of the four metrics, and in this case the metrics measure the quality of determining polarity of each review under test.

Table 3. CNSA-MSA-SAT Polarity Accuracy Results

Class	Accuracy	Error	Precision	Recall
Positive	-	-	80%	10%
Negative	-	-	10%	30%
Neutral	-	-	10%	10%
Dataset	90%	10%	90%	90%

The results shown in Table 3 indicate that CNSA-MSA-SAT is trustworthy to be used to automatically identify the polarity (Positive/Negative/Neutral) of reviews and comments written in colloquial Arabic and MSA.

Careful analysis of the test results leads us to identify some of the reasons that may show limitations in the tool:

1- The polarity of some of the phrases depends mainly on the domain they were used into. Consider the following two Arabic comments with their English translation, where the Arabic keyword (high, "عاليه") in these two comments lead to two different polarity values:

Table 4. Polarity of products/places reviews.

Domain	English Translation	Main Arabic Review	Polarity
Products	This is a high cost product.	هذا المنتج تكلفته عاليه	Negative
Places	High-quality service	خدمه عالية الجوده	Positive

The above two Arabic reviews (in Table 4) show clearly that the Arabic word (high, "عاليه") within the first Arabic comment leads to consider the polarity of the first comment negative, while using the same Arabic word (high, "عاليه") within the second Arabic comment leads to consider the polarity of the second comment as positive.

These two Arabic sample comments demonstrate why the effectiveness of our CNSA-MSA-SAT tool to determine the polarity of each Arabic comment within a domain is better than its effectiveness when it is applied on a general dataset. Mixing these Arabic comments in one dataset force the designed tool to consider the polarity of each comment that includes the keyword (high, "عاليه") as either positive or negative. Using key-phrases will help to solve this problem in future studies. Notice that the polarity shown above in the two sample comments is subjective. Therefore some people might consider "This is a high cost product." as positive.

2- Analyzing the results also lead us to discover spam reviews and comments. These are reviews and comments not related to the main topic.

3- The preliminary tests on our tool CNSA-MSA-SAT show its incapability to handle any text that has non-Arabic alphabets, such as those textual reviews and comments which use Arabic and Latin alphabets or those which use the Arabizi (Arabic chat alphabet). Therefore as mentioned in section 4.1 of this study, our collected dataset is filtered from reviews and comments that use non-Arabic alphabets. The capability of the proposed tool in this study is limited to purely Arabic text, and this is one of the limitations of the developed tool.

5. CONCLUSION AND FUTURE WORK

In this paper, we described the development of a specialized sentiment analysis tool for colloquial Arabic and modern standard Arabic (MSA) to evaluate social networks' sentiments in Arabic. The input to the tool is the collected dataset of opinions from social networks. CNSA-MSA-SAT tool is designed to identify the polarity of each inputted Arabic review or comment using a proposed method and formula that evaluates polarity of each word in the review/comment.

Results of the quality of the automatic polarity judgments of the tool are evaluated in comparison with manual judgment for each opinion where results showed 90% accuracy with general and domain based reviews and comments.

The main challenge in this research was the lack of publically tested and verified dataset in sentiment analysis specially in Arabic language where tool accuracy can be compared with as an alternative for self manual opinion evaluation to increase the level of confidence in the results. This is true particularly for Arabic where very few research papers are published in this area or they did not provide public evaluated datasets. We intend in future to publicize collected dataset to help others bypass such issue.

6. REFERENCES

[1] Arabic Speaking Internet Users and Population Statistics, http://www.internetworldstats.com/stats19.htm, accessed on January 8, 2013.

[2] Thelwall, M., Wilkinson, D., and Uppal, S. 2010. Data Mining Emotion in Social Network Communication: Gender Differences in MySpace. *Journal of the American Society for Information Science and Technology*. 61, 1, (Jan. 2010), 190–199.

[3] Thelwall, M., Buckley, K., and Paltoglou, G. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*. 63, 1, (Jan. 2012), 163-173.

[4] Abbasi, A., Chen, H., and Salem, A. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems*. 26, 3, (Jun. 2008), Article 12.

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

[6] Abdul-Mageed, M., Diab, M., and 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* (Portland, Oregon, USA, (June 19-24, 2011). 2, 587-591.

[7] Korayem, M., Crandall, D., and Abdul-Mageed, M. 2012. Subjectivity and Sentiment Analysis of Arabic: A Survey. *Advanced Machine Learning Technologies and Applications, Communications in Computer and Information Science*. 322, 128-139.

[8] Rushdi-Saleh, M., Martín-Valdivia, M., Ureña López, L., and Perea-Ortega, J. 2011. Bilingual Experiments with an Arabic-English Corpus for Opinion Mining. In *Proceedings of Recent Advances in Natural Language Processing* (Hissar, Bulgaria, September 12-14, 2011). 740-745.

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

[10] Elhawary, M., Elfeky, M. 2010. Mining Arabic Business Reviews. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*. 1108-1113.

[11] Al-Subaihin, A., Al-Khalifa, H., and Al-Salman, A. 2011. A proposed sentiment analysis tool for modern Arabic using human-based computing. In *Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services (iiWAS '11)*, ACM, New York, NY, USA, 543-546.

[12] Witten I. H. and Frank, E. 2005. *Data Mining: Practica Machine Learning Tools and Techniques*, Morgan Kaufmann Series in Data Management Systems, second edition, Morgan Kaufmann (MK).

[13] Paltoglou, G., and Thelwall, M. 2012. Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 3, 4, Article 66, 1-19.