Evaluating Social Context in Arabic Opinion Mining

Mohammed Al-Kabi[1], Izzat Alsmadi[2], Rawan Khasawneh[3], and Heider Wahsheh[4]
[1]Computer Science Department, Zarqa University, Jordan
[2]Computer Science Department, University of New Haven, USA
[3]Computer Information Systems Department, Jordan University of Science and Technology, Jordan
[4]Computer Science Department, King Khaled University, Saudi Arabia

Abstract: *This study is based on a benchmark corpora consisting of* 3,015 *textual Arabic opinions collected from Facebook. These collected Arabic opinions are distributed equally among three domains* (*Food, Sport, and Weather*), *to create a balanced benchmark corpus. To accomplish this study ten Arabic lexicons were constructed manually, and a new tool called Arabic Opinions Polarity Identification* (*AOPI*) *is designed and implemented to identify the polarity of the collected Arabic opinions using the constructed lexicons. Furthermore, this study includes a comparison between the constructed tool and two free online sentiment analysis tools* (*SocialMention and SentiStrength*) *that support the Arabic language. The effect of stemming on the accuracy of these tools is tested in this study. The evaluation results using machine learning classifiers show that AOPI is more effective than the other two free online sentiment analysis tools using a stemmed dataset.*

Keywords: *Big data, social networks, sentiment analysis, Arabic text classification, and analysis, opinion mining.*

Received November 20, 2015; accepted March 30, 2016

1. Introduction

In 2001 Laney defined the growth, opportunities and challenges in the data as 3D (volume, velocity, and variety), and called it the "3Vs" model [25], and in 2012 this term was renamed to be big data. The term "Big data" is cast in 2001 a bit of a misnomer, as no actual pre-existing data is somehow small. Big data refers to any collection of large and complex datasets, which could not be processed by traditional data processing applications.

Sentiment Analysis (SA) and Opinion Mining (OM) were cast for the first time by Nasukawa and Yi [28], and it is the field of study interested in the automatic analysis of people's opinions, sentiments, evaluations, attitudes, and emotions [26]. This field of study is one of the most active research areas, and overlaps with Natural Language Processing (NLP), Data Mining (DM), text mining, and Web mining. Also, studies related to management and social sciences use SA and OM. So this field is not restricted anymore to computer science and Information technology [9].

The rise and expansion of the Islamic empire on seventh to twelfth centuries AD, leads to make this language the official language of the entire empire, beside every Muslim should use this language in pray and when he/she reads the holy book of Quran [32].

Nowadays the Arabic language is the descendant of the classical Arabic language of the pre-Islamic era. The two main types of Arabic language used these days are Modern Standard Arabic (MSA) and different Arabic dialects. This language is the native language of more than 300 million people mostly living in the Middle East and North and East Africa. Furthermore, Arabic language is the liturgical language for over a billion Muslims around the World [32].

Around 5% of the world population is Arabs, and 3.8% of the Internet users worldwide are Arabs. The Arabic Web content constitutes less than 1% of the Internet content. Around one third (33%) of the Arabic Web content is a low-quality content generated by the users of Social media. The volume of information generated by Social Web English language users does not exceed 10% of the total English Web content. A study by [10] shows that Arab social media users use MSA, and other Arabic vernaculars (dialects) like Egyptian, Levantine, Khaliji, English, French, Arabizi, and a mixture of MSA and English. The use of vernaculars in social media leads to a low-quality content, and in general when the percentage of vernaculars in a content increased the quality of this content is decreased. Also, social media content in this part of the world has emoticons.

We collected more than 3,000 Arabic opinions collected from Facebook to construct our benchmark corpora. These Arabic opinions are textual comments that use MSA and Arabic dialects (Egyptian, Mesopotamian, Levantine, and Arabian Peninsula). Arabic dialect groups are discussed extensively by [28]. The authors of this study collect these Arabic opinions deliberately so that they are distributed equally among three domains (Food, Sport, and Weather), in order to create a balanced benchmark corpus. The use of three different domains (Food, Sport, and Weather) aims to study the effects of varied characteristics of these domains on the accuracy of determining the polarity of each inputted Arabic opinion.

The rest of this study is organized as follows: section 2 presents the related work to OM with special emphasize on Arabic OM, section 3 presents the methodology followed to accomplish this study. Section 4 presents benchmarking test results. Finally, section 5 outlines the main findings of this study as well as the planned future work.

2. Related Work

The first study which coined the term SA for the first time was conducted by [28].

In their study which conducted before the emergence of Web 2.0 they presented how to automatically extract sentiments associated with polarities from documents instead of classifying a whole document into either positive or negative. Nasukawa, & Yi study indicates that is essential to identify how textual sentiments are expressed in SA, and this will help to identify automatically whether a sentiment is positive (favorable) or negative (unfavorable) or neutral [28].

Previous SA studies have categorized the extracted features into four main feature categories: syntactic, semantic, link-based, and stylistic features. Semantic features used in this study were extracted manually to build the lexicons. This human involvement makes semantic features a powerful approach for SA [1].

There are many free outsource online SA tools, so to conduct this study a large number of them were tested to discover whether they support Arabic language or not. We found only two free outsource online SA tools (SocialMention and SentiStrength) support Arabic [33, 34].

This study aims to evaluate the effectiveness of in-house tool developed by the fourth author, and other two free outsource online SA tools capable to analyzing Arabic opinions. The in-house tool used in this study is called Arabic Opinion Polarity Identification (AOPI).

The first outsource online SA tool used in this study and support the Arabic language is called Social Mention (SM). SM is a Web service and one of the leading social media search engines which allows its users to search for different posts in different social networks, and it is an analysis platform which produces a number of essential statistics in the field of SA [34].

The algorithms adopted by SM are unknown, but the authors of [22] deduce from their observations to search results that SM uses an exact keyword matching approach to identify the social media posts that contain the given keywords.

Last and not least SM supports Arabic language comments and analysis. The second outsource online SA tool that support the Arabic language is called SentiStrength which is proposed for the first time by [36] and was used first to score informal English sentiments. It is similar to SM as it is a Web service and a social media search engine that allows its users to search in different social networks. This tool has been developed at the University of Wolverhampton [33].

SentiStrength is capable of numerically measuring the strengths of positive and negative textual sentiments. SentiStrength is lexicon-based and can handle emoticons and correct spelling mistakes. It is capable of measuring numerically the strengths of positive and negative textual sentiments, where integer scores (positive and negative) are assigned to each sentiment in the range (-5, 5). e.g., SentiStrength can assign (-4, 3) to the sentiment "I hate devil but like God". This is called dual positive and negative scoring and one of SentiStrength merits. It can also handle informal expressions of sentiments such as “I'm saaaaaad!!!!!!!!!”. An improved version of Senti-Strength called SentiStrength 2 was proposed by [35].

The field of Arabic SA has been receiving a lot of attention since its rather shy start a decade ago [6, 7, 15]. Recently, many teams have been making significant contributions to this field such as the ones at Columbia University [4, 5], University of Jaén [30, 31], Egyptian universities [19, 21, 27] and Jordanian universities [2, 3, 8, 9, 10, 14, 16, 17, 29]. The interested reader is referred to the following surveys [18, 20, 24] to learn more about these papers and other important papers in this field.

The authors of [23] present a comparison of the effectiveness of two free outsource online SA tools (SocialMention and SentiStrength) that support Arabic. They concluded that SentiStrength tool is more effective than SocialMention tool.

The closest studies to this one are [11, 12, 13]. In [12], the authors compared the effectiveness of two free online SA tools. They tested SocialMention tool that is used in this study, and another tool called Twendz (http://twendz.waggeneredstrom.com/). They built three manually constructed polarity dictionaries.

The first polarity dictionary is dedicated to Arabic reviews, and the second polarity dictionary is dedicated to English reviews, and the third one is dedicated to emoticons. They concluded that SocialMention is more effective than its counterpart (Twendz). In [13] the authors build new SA system called Colloquial Non-Standard Arabic-Modern Standard Arabic-Sentiment Analysis Tool (CNSAMSA-SAT), and as the system name indicates it is capable to identify colloquial Arabic and MSA reviews and comments.

3. Methodology

To evaluate the effectiveness of each of the three SA tools used in this study a number of essential steps (processes) are presented to show the details of each step in the proposed framework. We designed and implemented a series of experiments to identify the

best tool.

The following steps describe the methodology of this study:

1. Collect 3,015 Arabic opinions that are distributed equally among three main domains: Food, Sport, and Weather. The collected Arabic opinions used MSA and colloquial Arabic. Furthermore, the collected textual dataset includes emoticons. This represents a balanced benchmark corpus.
2. Build ten main lexicons: Two lexicons (Positive, and negative) were constructed for each domain. So, we have six domain-based lexicons, beside two general lexicons. Two additional lexicons were built for emoticons (Positive, and negative).
3. Develop an in-house tool to identify the polarity (positive, negative, and neutral) of each inputted Arabic opinion whether it uses MSA or colloquial Arabic. This in-house tool is called Arabic Opinion Polarity Identification (AOPI).
4. Use Arabic light stemmer to extract the Arabic stems of MSA Arabic words of the collected Arabic opinions.
5. Test the effectiveness of our proposed in-house tool AOPI against the effectiveness of the other two free outsource online SA tools (SocialMention and SentiStrength) using the balanced benchmark corpora created in the first step.
6. Evaluate AOPI performance using machine learning classifiers.

3.1. Dataset

A crawler is used to automatically collect this modest dataset that consists of 3,015 Arabic opinions; this crawler targets Facebook pages based on specific keywords related to each domain. One may ask why this study is based on a relatively small dataset collected by a crawler? The reason of collecting a limited dataset is due to the manual filtration and the manual annotation of the collected opinions. The researchers exclude those opinions that have Latin letters, or those used Arabic chat alphabets (Arabizi).

Also, we exclude duplicated, noisy and spammed opinions during the construction of our dataset.

Analysis of the collected dataset shows the percentage of slang words (colloquial words) is between 55% and 65%. On average 60% of the words included in the opinions were slang.

The collected opinions are classified manually into three domains: Food, Sport, and Weather. Ten Arabic lexicons were built, so that there are two polarity domain-based lexicons for each domain, beside two general polarity lexicons. Authors have manually evaluated the dataset and extracted positive, and negative words/phrases from it based on the common use of each word/phrase by ordinary people. Two additional lexicons were built for emoticons (Positive, and negative). Table 1 shows a sample of Arabic comments as well as their corresponding English translation.

Table 1. Sample of arabic comments with their english translation.

Positive Comments		**Negative Comments**	
Arabic Comment	English Translation	Arabic Comment	English Translation
من ايدك كل شي طيب وافضلها كلها	Everything you made is delicious; I prefer them all	يسبب خطر علي القولون و يسبب القرح	Cause risk to colon and cause ulcers
الدون افضل لاعب في العالم	AlDoon is the best player in the world	اتلتيكو نفس بنفيكا الموسم الماضي ينافس على جميع البطولات وخسر الجميع!!	Atletico behaves like Benfica last season competing at all championships and lost everyone!!
الحمد لله رب العالمين على نعمة الامطار	Praise be to God for the blessing of rain	حساسيه وامراض تنفس):	Allergic and breathing diseases :)

3.2. SocialMention and SentiStrength Tools

SocialMention [34] is a free and real-time social media search and analysis platform, where user generated contents across the universe are aggregated into a single stream of information. SocialMention provides several tools that allow you to track, measure, and analyze what people are saying about your company, a new product/brand or any other topic across the web's social media landscape including: Facebook, Twitter, YouTube, Digg, and many others.

SentiStrength [33] is a SA program that estimates the strength of positive and negative sentiment, and its capabilities are not limited to English short texts (including informal texts) but also other languages such as: Arabic, German, Spanish, Russian, Turkish, French, Italian, Dutch, Greek, etc. It is free for academic research and can be configured to support any other languages by changing its input files.

In order to classify a text, SentiStrength algorithm uses two scales: from 1 (not positive) to 5 (very strong positive sentiment) and from -1 (not negative) to -5 (very strong negative sentiment). It evaluates the contribution of positive and negative sentiments separately and makes a decision based on their values. In addition to a sentiment word dictionary, it takes into account the most common spelling styles in social networks [33].

AOPI is a novel in-house tool SA tool that is used to identify the Arabic Opinions Polarity. AOPI depends on the Term Frequency (TF) which is the numerical measure presenting the weight of different words in a document. Tf counts the numbers of times each word appears in the document. AOPI is capable to read the Arabic opinions in the benchmark dataset, and identify the polarity of each of them using the ten lexicons already built by the authors, and create classification file which assigns (0 for negative), (1 for positive), (2 for neutral). The pseudocode of AOPI algorithm is presented in Figure 1.

Algorithm 1: Arabic opinions polarity identification algorithm

```
Input:
TO: Set of Arabic Textual opinions.
PL: Set of Positive Sentiment Lexicons.
NL: Set of Negative Sentiment Lexicons.
UDS: User Domain Selection.
Output:
OP: Opinion polarity.
OPF: Opinion polarity classification file.
Initialization:
P_TF = 0, where P_TF is the TF for positive sentiments.
N_TF = 0, where N_TF is the TF for negative sentiments.
Neut_TF = 0, where Neut_TF is the TF for neutral sentiments.
Begin
  1:  Inputting UDS by the users [Optional]
  2:  The users insert TO
  3:  For every TO:
  4:    Remove stop words.
  5:    Divide TO into w words (tokens)
  6:    Normalize the similar Arabic alphabets.
  7:    For each w,
  8:      Search for similar w in PL, NL.
  9:      If win PL then
 10:        P_TF = P_TF + 1
 11:        OP = Positive
 12:      Else If w in NL then
 13:            N_TF = N_TF + 1
 14:            OP = Negative
 15:          Else
 16:              Neut_TF = Neut_TF + 1
 17:              OP = Neutral
 18:          End If
 19:      End If
 20:    End For // w
 21:    If (P_TF > N_TF) then
 22:      OP = Positive
 23:    Else If (N_TF > P_TF) then
 24:          OP = Negative
 25:        Else
 26:          OP = Neutral
 27:        End If
 28:    End If
 29:    Write OP to OPF (final result file)
 30: End For
End.
```

The pseudocode presented in Algorithm 1 shows the main steps that describe the proposed AOPI Algorithm. Tokenizing the opinions into a number of tokens (words) is essential to be conducted before pre-processing. Figure 1 includes a description of the pre-processing steps of the collected Arabic text starting with a removal of Arabic stop words that usually represent the most frequent tokens (words) in the collection of the collected opinions. Furthermore, pre-processing steps include a normalization of some Arabic alphabets like Alif by converting different forms of this letter (Alif with Madda above, "آ"), (Alif with hamza above, "أ"), (Alif with Hamza below, "إ") to one unified form (Bare Alif, "ا"). Also the Arabic letter (Waaw seat, "ؤ") is converted to (Waaw, "و"), (Alif maqSuura, "ى") is converted to (FinalYaa, "ي") due to the use of many people to (Alif maqSuura, "ى") instead of final Yaa, (TaamarbuuTa, "ة") is converted to (Final haa, "ه").

The normalization includes the removal of consonant doubling (gemination) which is known as tashdiid by removing the (Shadda, " ّ") character, all other diacritical symbols are removed like Sukuun ('silence'), FatHa, Kasra, Damma characters, beside the removal of Indefinite marker (Nunation (tanwiin)) characters. AOPI tool depends on the TF search for similar words in positive and negative sentiment lexicons. The AOPI Algorithm identifies the opinion under consideration as a positive opinion when P_TF is greater than the N_TF, and similarly the algorithm identifies an opinion as negative when N_TF is greater than P_TF. The opinion is considered as a neutral opinion by the above algorithm when P_TF is equal to N_TF.

4. Results and Evaluation

This section is divided into two subsections; the first is dedicated to present the results of two free online SA tools (SocialMention and SentiStrength), while the second section to present an evaluation of AOPI tool using original and stemmed datasets.

4.1. Online Sentiment Analysis Tools Evaluation

We run the two SA tools under consideration on our dataset to evaluate the effectiveness of these two tools to identify the polarity of each inputted Arabic opinion. The results of the accuracy are shown in Table 2.

Table 2. Accuracy results for the two online tools.

Domain	Accuracy of SentiStrength	Accuracy of SocialMention
Food	49.949%	49.349%
Sport	71.193%	63.433%
Weather	60.101%	53.896%

The results of SentiStrength tool tests provide a strong evidence that this tool is unable to identify the correct polarity of opinions written in one of the Arabic vernaculars. The Arabic opinions are translated by SentiStrength tool to English. Its translation capabilities are limited to MSA, and it is unable to translate any of the Arabic vernaculars. Therefore, all polarities of opinions written in one of the Arabic vernaculars are identified by SentiStrength as neutral due to incapability of this tool to translate Arabic vernaculars to English. This leads to its failure to identify the correct polarity for each inputted Arabic vernacular opinion, while the second tool SocialMention failed to assign any polarity to more than 50% of the inputted opinions, and this is due to the percentage of opinions written in one of the Arabic vernaculars that exceeds 55%. Therefore, the effectiveness of SocialMention is lower than 50%.

The limited effectiveness of these two free tools to identify the polarity of different Arabic opinions motivate us to build our tool AOPI to deal with

opinions written in MSA or in one of the Arabic vernaculars.

4.2. Evaluation of Arabic Opinion Polarity Identification (AOPI)

It is known that the determination of the domain of each opinion improves the accuracy of identifying the polarity of that opinion, because it is normal to face an English/Arabic word such as (quit, "هادئ") that has two different polarities, it is considered a positive polarity word when somebody talk about "quiet cars" within automobiles domain, while in opinions like "Quiet phones" and "Quiet Alarm clock" this word is considered a negative polarity word. Also it is normal within a certain domain like computer domain to find a word that has two polarities, let us consider (This laptop has a long life battery, " هذا الحاسوب المحمول يحتوي على بطارية تدوم طويلا") review where the polarity of the word (long, "طويل") is considered positive, and let us consider (This laptop has a long start-up time, " هذا الحاسوب المحمول يستغرق وقتا طويلا ليبدأ بالإشتغال") review where the polarity of the word (long, "طويل") is considered negative.

To evaluate the AOPI tool we run this tool twice on every collected Arabic opinion: First without user's specification of the domain of the review, and in the second run the domain of the opinion should be specified by the user. Two datasets were used in this study to explore the effect of light stemming on the AOPI tool accuracy:

- First dataset presents the original collected dataset.
- An Arabic light stemmer is used to stem the original collected dataset (First Dataset), and generate the second Dataset.

This study is based on three classification algorithms: (K-Nearest Neighbour (k-NN), Naïve Bayes (NB), and Support Vector Machine (SVM)) to evaluate the AOPI tool. We used 66% of the dataset as a training set and the rest (34%) as a test set.

The first classification algorithm used is the k-NN algorithm, which is a non parametric lazy learning algorithm for classifying objects based on closest training examples in the feature space. The value of k is equal to 1 in all k-NN experiments that were conducted in this study. The k-NN yields 35.7073% accuracy without domain specification by the user of each inputted opinion using the first original dataset, and an accuracy of 37.3896% for the stemmed dataset without domain specification by the user of each inputted opinion.

As expected the determination of domains of different opinions leads to more accurate results. Therefore, when the domains of opinions are set by the user, the *k-NN* yields an accuracy of 66.1765% for food domain, an accuracy of 45.8689% for sports domain, and 62.6866% for weather domain using the original dataset. While *k-NN* yields an accuracy of 67.0352% for food domain, an accuracy of 49.7126 % for sports domain, and 61.9388% for weather domain for the stemmed dataset.

Table 3 presents the results of k-NN effectiveness when the domains of different opinions are not identified, where we used the Receiver Operating Characteristic (ROC) prediction quality metrics: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Precision, Recall, and F-Measure (F-M). Receiver Operating Characteristic (ROC) represents a graphical plot that illustrates the performance of a binary classifier system. Formulas of 1, 2, 3 and 4 represent the Accuracy, Recall, Precision, and F-measure respectively [37].

$$Accuracy_i = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Recall_i = \frac{TP}{TP + FN} \quad (2)$$

$$Precision_i = \frac{TP}{TP + FP} \quad (3)$$

$$F - measure = \frac{(2 \times TP)}{(2 \times TP) + FP + FN} \quad (4)$$

Table 3. Stemmed and non-stemmed results of *k-NN* to identify opinion's polarities (Unspecified Domains).

Dataset	TP	FP	Precision	Recall	*F*-Measure	ROC
Original Dataset						
Weighted AVG	0.357	0.099	0.399	0.357	0.349	0.696
Stemmed Dataset						
Weighted AVG	0.374	0.117	0.408	0.374	0.354	0.733

Different effectiveness results showed in Table 3 reveal that light Arabic stemming has a slight positive effect on the accuracy of results. Afterward, we explore the effect of specifying the domain of each collected Arabic opinion on the accuracy of polarity determination of k-NN algorithm using the two datasets (non-stemmed and stemmed). The effectiveness results of these tests are presented in Table 4.

Table 4 results are better than the results presented in Table 3, and this is a certification of the conclusion that specifying opinion's domains lead to improve the effectiveness of opinion polarity determination as justified in the first paragraph of this section.

SVM is a popular supervised learning model that analyzes data and recognizes patterns based on a discriminative classifier. SVM yields 53.561% accuracy to identify a polarity of each inputted Arabic opinion when the domains of these opinions are not specified using the original dataset, and SVM yields an accuracy of 53.9745% for the stemmed dataset under the same conditions. AOPI tool allows the user to specify the domain of each inputted opinion; so as expected domains specification leads to better results.

Therefore, SVM yields an accuracy of 75.8824% for food domain, an accuracy of 52.1368% for sports domain, and 65.9701% for weather domain using the original dataset. While SVM yields an accuracy of 76.3314% for food domain, an accuracy of 53.4483% for sports domain, and 70.8709% for weather domain using the second stemmed dataset.

Table 4. Stemmed and non-stemmed results of k-NN within specified domains.

Domain	TP	FP	Precision	Recall	*F*-Measure	ROC
Original Dataset						
Food Positive	0.545	0.113	0.871	0.545	0.671	0.78
Food Negative	0.444	0.045	0.211	0.444	0.286	0.679
Food Neutral	0.85	0.406	0.574	0.85	0.685	0.802
Sport Positive	0.319	0.296	0.274	0.319	0.294	0.492
Sport Negative	0.138	0.041	0.5	0.138	0.216	0.527
Sport Neutral	0.672	0.596	0.543	0.672	0.6	0.522
Weather Positive	0.57	0.159	0.67	0.57	0.616	0.771
Weather Negative	0.048	0.027	0.2	0.048	0.077	0.549
Weather Neutral	0.808	0.509	0.626	0.808	0.706	0.718
Stemmed Dataset						
Food Positive	0.507	0.045	0.933	0.507	0.657	0.771
Food Negative	0.152	0.001	0.875	0.152	0.259	0.602
Food Neutral	0.955	0.515	0.554	0.955	0.701	0.762
Sport Positive	0.215	0.149	0.298	0.215	0.25	0.596
Sport Negative	0.133	0.049	0.458	0.133	0.206	0.53
Sport Neutral	0.78	0.753	0.543	0.78	0.64	0.548
Weather Positive	0.699	0.389	0.672	0.699	0.685	0.728
Weather Negative	0.672	0.285	0.562	0.672	0.612	0.803
Weather Neutral	0.088	0.016	0.417	0.088	0.146	0.655

Although it is known that the accurate detection of opinion's polarity is domain-dependant due to domain-specific features, we conducted tests to identify opinion polarities regardless of their domain. Table 5 shows the effect of stemming on the results of SVM to identify opinion polarities without specifying the domains of the opinions under consideration.

Table 5. Stemmed and non-stemmed results of SVM (Unspecified Domains).

Dataset	TP	FP	Precision	Recall	*F*-Measure	ROC
Original Dataset						
Weighted AVG	0.536	0.077	0.53	0.536	0.526	0.83
Stemmed Dataset						
Weighted AVG	0.54	0.078	0.53	0.54	0.528	0.843

The above results presented in table 5 showed that stemming has a slight positive effect on accuracy results of SVM.

Table 6. Stemmed and non-stemmed results of SVM (Domains are specified).

Domain	TP	FP	Precision	Recall	*F*-Measure	ROC
Original Dataset						
Food Positive	0.758	0.197	0.843	0.758	0.798	0.789
Food Negative	0.222	0.012	0.333	0.222	0.267	0.696
Food Neutral	0.797	0.242	0.679	0.797	0.734	0.778
Sport Positive	0.407	0.165	0.463	0.407	0.433	0.639
Sport Negative	0.263	0.137	0.362	0.263	0.304	0.562
Sport Neutral	0.694	0.515	0.587	0.694	0.636	0.585
Weather Positive	0.686	0.121	0.761	0.686	0.722	0.813
Weather Negative	0.143	0.078	0.207	0.143	0.169	0.58
Weather Neutral	0.767	0.399	0.67	0.767	0.715	0.674
Stemmed Dataset						
Food Positive	0.859	0.294	0.779	0.859	0.817	0.787
Food Negative	0.071	0.009	0.25	0.071	0.111	0.522
Food Neutral	0.705	0.161	0.754	0.705	0.729	0.762
Sport Positive	0.468	0.234	0.37	0.468	0.413	0.655
Sport Negative	0.241	0.091	0.455	0.241	0.315	0.617
Sport Neutral	0.694	0.463	0.632	0.694	0.662	0.612
Weather Positive	0.831	0.406	0.743	0.831	0.785	0.702
Weather Negative	0.644	0.135	0.684	0.644	0.663	0.762
Weather Neutral	0.206	0.033	0.412	0.206	0.275	0.665

Next we explore the effect of specifying the domain of each collected Arabic opinion on the accuracy of polarity determination of the SVM algorithm using the two datasets (non-stemmed and stemmed). The SVM effectiveness results of these tests are presented in table 6, where the domain of each inputted opinion is specified.

The results presented in Table 6 show clearly that specifying domains of different opinions yield better results relative to the results of Table 5, when the SVM is used to identify polarity of the inputted Arabic opinions relative to results of identifying the polarity of the inputted Arabic opinions when the domains are not identified. Furthermore, Tables 5 and 6 shows that light stemming has a slight positive effect on the polarity's accuracy for the food and weather domains.

Naïve Bayes (NB) presents a supervised learning method as well as a statistical method for classification. It is based on applying Bayes' theorem with strong independence assumptions. Table 7 presents the detailed results of *NB* to identify the polarity of the inputted Arabic opinions when the domain of each inputted opinion is not specified by the user.

Table 7. Stemmed and non-stemmed results of NB (Unspecified Domains).

Dataset	TP	FP	Precision	Recall	*F*-Measure	ROC
Original Dataset						
Weighted AVG	0.497	0.085	0.494	0.497	0.486	0.833
Stemmed Dataset						
Weighted AVG	0.507	0.082	0.513	0.507	0.499	0.852

Table 7 presents the accuracy results of NB classifier to identify the opinion's polarity, when the domain of each inputted Arabic opinion is not specified by the user. This classifier is run twice on the original non-stemmed dataset and on the stemmed dataset, with 49.6585% polarity's accuracy for original non-stemmed opinions and 50.736% polarity's accuracy for stemmed opinions. As in other classification algorithms, the ignorance of the opinion's domains leads to lower the effectiveness of the classification algorithm.

Last and not least Table 8 shows the effectiveness of *NB* classifier to identify the opinion's polarity, when the domain of each inputted Arabic opinion is specified by the user. The accuracies of *NB* classifier to identify opinion's polarity for food, sport, and weather domains using the original non-stemmed dataset are 70.8824%, 55.5556%, and 66.5672% respectively. While the accuracies of *NB* classifier to identify opinion's polarity for food, sport, and weather domains using the second stemmed dataset are 71.0059%, 53.0792%, and 67.8679% respectively. Table 8 showed a clearly negative effect of light stemming within the sports domain.

The results presented in Table 8 show clearly that the selection of specific domains yield better polarity's accuracy relative to the polarity's accuracy of *NB* when no domain is selected as expected.

Results in Tables 2-8 show that SVM is the most accurate classifier to evaluate AOPI tool, when stemmed dataset is used and the domain is specified by the user.

Table 8. Stemmed and non-stemmed results of NB (Domains are specified).

Domain	TP	FP	Precision	Recall	*F*-Measure	ROC
Original Dataset						
Food Positive	0.662	0.148	0.862	0.662	0.749	0.817
Food Negative	0.333	0.039	0.188	0.333	0.24	0.786
Food Neutral	0.805	0.314	0.622	0.805	0.702	0.822
Sport Positive	0.319	0.096	0.537	0.319	0.4	0.68
Sport Negative	0.375	0.14	0.441	0.375	0.405	0.671
Sport Neutral	0.756	0.544	0.594	0.756	0.665	0.654
Weather Positive	0.678	0.098	0.796	0.678	0.732	0.882
Weather Negative	0.262	0.082	0.314	0.262	0.286	0.773
Weather Neutral	0.756	0.411	0.66	0.756	0.705	0.77
Stemmed Dataset						
Food Positive	0.735	0.275	0.764	0.735	0.749	0.814
Food Negative	0.357	0.015	0.5	0.357	0.417	0.803
Food Neutral	0.712	0.256	0.66	0.712	0.685	0.795
Sport Positive	0.25	0.087	0.496	0.25	0.332	0.674
Sport Negative	0.352	0.161	0.392	0.352	0.371	0.677
Sport Neutral	0.747	0.582	0.58	0.747	0.653	0.628
Weather Positive	0.79	0.391	0.74	0.79	0.764	0.763
Weather Negative	0.635	0.083	0.776	0.635	0.698	0.834
Weather Neutral	0.176	0.114	0.15	0.176	0.162	0.694

5. Conclusions and Future Work

In this study, we conducted a study to evaluate three OM tools for Arabic language; two of them are free online tools (SocialMention and SentiStrength), while the third is a novel tool proposed by the authors and called AOPI.

An annotated and stemmed dataset consisting of 3015 Arabic posts was collected from the Facebook.

AOPI tool is developed to conduct this study to judge the polarities of Arabic opinions based on their contents. Three classification algorithms: (k-NN, NB, and SVM) were used to evaluate the effectiveness the three tools (SocialMention, SentiStrength, and AOPI) to identify the polarities of inputted Arabic opinions.

The results showed that stemmed dataset yields better results than original dataset and SVM yields best results to evaluate the AOPI tool. The benchmark tests show that the effectiveness of our AOPI tool is better than the effectiveness of the other two free online tools (SocialMention and SentiStrength).

Furthermore, this tool can deal with the 3015 Arabic opinions regardless of Arabic variety used to express them.

As mentioned before in the methodology the percentage of slang words (colloquial words) is between 55% and 65%. The used stemmer in this study is designed for MSA, and this means that it is unable to extract the stem of all slang words (colloquial words). This indicates that 55% and 65% of the words in our dataset are not stemmed. We plan in our future studies to solve the problem of stemming of different Arabic vernaculars, and treat the problem of spelling mistakes and repetition of letters and characters that found in many Arabic sentiments and opinions.

Acknowledgment

This research is funded by the deanship of scientific research at Zarqa University, Jordan.

References

[1] Abbasi A., Chen H., and Salem A., "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums," *ACM Transactions on Information Systems*, vol. 26, no. 3, 2008.

[2] Abdulla N., Ahmed N., Shehab M., Al-Ayyoub M., Al-Kabi M., and Al-Rifai S., "Towards Improving the Lexicon-Based Approach for Arabic Sentiment Analysis," *International Journal of Information Technology and Web Engineering*, vol. 9, no. 3, pp. 55-71, 2014.

[3] Abdulla N., Majdalawi R., Mohammed S., Al-Ayyoub M., and Al-Kabi M., "Automatic Lexicon Construction for Arabic Sentiment Analysis," *in Proceedings of the 2nd International*

Conference on Future Internet of Things and Cloud, Barcelona, pp. 547-552, 2014.

[4] Abdul-Mageed M. and Diab M., "AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis," *in Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, pp. 3907-3714, 2012.

[5] Abdul-Mageed M., Diab M., and Kübler S., "SAMAR: Subjectivity and Sentiment Analysis for Arabic Social Media," *Computer Speech and Language*, vol. 28, no. 1, pp. 20-37, 2014.

[6] Ahmad K. and Almas Y., "Visualising Sentiments in Financial Texts?," *in Proceedings of 9th International Conference on Information Visualisation*, London, pp. 363-368, 2005.

[7] Ahmad K., Cheng D., and Almas Y., "Multi-Lingual Sentiment Analysis of Financial News Streams," *in Proceedings of International Workshop on Grid Technology for Financial Modeling and Simulation*, Palermo, 2006.

[8] Al Shboul B., Al-Ayyoub M., and Jararweh Y., "Multi-Way Sentiment Classification of Arabic Reviews," *in Proceedings of the 6th International Conference on Information and Communication Systems*, Amman, pp. 206-211, 2015.

[9] Al-Ayyoub M., Bani-Essa S., and Alsmadi I., "Lexicon-Based Sentiment Analysis of Arabic Tweets," *International Journal of Social Network Mining*, vol. 2, no. 2, pp. 101-114, 2015.

[10] Al-Kabi N., Abdulla N., and Al-Ayyoub M., "An Analytical Study of Arabic Sentiments: Maktoob Case Study," *in Proceedings of 8th International Conference for Internet Technology and Secured Transactions*, London, pp. 121-126, 2013.

[11] Al-Kabi M., Gigieh A., Alsmadi I., Wahsheh H., and Haidar M., "Opinion Mining and Analysis for Arabic Language," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 5, pp. 181-195, 2014.

[12] Al-Kabi M., Al-Qudah N., Alsmadi I., Dabour M., and Wahsheh H., "Arabic/English Sentiment Analysis: An Empirical Study," *in Proceedings of 4th International Conference on Information and Communication Systems*, Irbid, pp. 1-6, 2013.

[13] Al-Kabi M., Gigieh A., Alsmadi I., Wahsheh H., and Haidar M., "An Opinion Analysis Tool for Colloquial and Standard Arabic," *in Proceedings of 4th International Conference on Information and Communication Systems*, Irbid, 2013.

[14] Al-Kabi M., Al-Ayyoub M., Alsmadi I., and Wahsheh H., "A Prototype for a Standard Arabic Sentiment Analysis Corpus," *The International Arab Journal of Information Technology*, vol. 13, no. 1A, pp. 163-170, 2016.

[15] Almas Y. and Ahmad K., "A Note on Extracting 'Sentiments' in Financial News in English, Arabic and Urdu," *in Proceedings of 2nd Workshop on Computation, al Approaches to Arabic Script-based Languages*, California, pp. 1-12, 2007.

[16] Al-Smadi M., Al-Sarhan H., Al-Ayyoub M., and Jararweh Y., "Using Aspect-Based Sentiment Analysis to Evaluate Arabic News Affect on Readers," *in Proceedings of 8th International Conference on Utility and Cloud Computing*, Limassol, pp. 436-441, 2015.

[17] AL-Smadi M., Qawasmeh O., Talafha B., and Quwaider M., "Human Annotated Arabic Dataset of Book Reviews for Aspect Based Sentiment Analysis," *in Proceedings of 3rd International Conference on Future Internet of Things and Cloud*, Rome, pp. 726-730, 2015.

[18] Al-Twairesh N., Al-Khalifa H., and Al-Salman A., "Subjectivity and Sentiment Analysis of Arabic: Trends and Challenges," *in Proceedings of 11th International Conference on Computer Systems and Applications*, Doha, pp. 148-155, 2014.

[19] Aly M. and Atiya A., "LABR: A Large Scale Arabic Book Reviews Dataset," *in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, pp. 494-498, 2013.

[20] Biltawi M., Etaiwi W., Tedmori S., Hudaib A., and Awajan A., "Sentiment Classification Techniques for Arabic Language: A Survey," *in Proceedings of the 7th International Conference on Information and Communication Systems*, Irbid, pp. 339-346, 2016.

[21] ElSahar H. and El-Beltagy S., "Building Large Arabic Multi-Domain Resources for Sentiment Analysis," *in Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics*, Cairo, pp. 23-34, 2015.

[22] Kajanan S., Shariff A., Dutta K., and Datta A., "Resolving Name Conflicts for Mobile Apps in Twitter Posts," *International Federation for Information Processing*, pp. 3-17, 2012.

[23] Khasawneh R., Wahsheh H., Al-Kabi M., and Alsmadi I., "Sentiment Analysis of Arabic Social Media Content: A Comparative Study," *in Proceedings of 8th International Conference for Internet Technology and Secured Transactions*, London, pp. 101-106, 2013.

[24] Korayem M., Crandall D., and Abdul-Mageed M., "Subjectivity and Sentiment Analysis of Arabic: A survey," *in Proceedings of Advanced Machine Learning Technologies and Applications*, Cairo, pp. 128-139, 2012.

[25] Laney D., "3D Data Management: Controlling Data Volume, Velocity, and Variety," [Gartner Group, 2001]. Available at: http://blogs.gartner.com/doug-

laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf, Last Visited, 2015.
[26] Liu B., *Sentiment Analysis and Opinion Mining*, Morgan and Claypool Publishers, 2012.
[27] Nabil M., Aly M., and Atiya A., "ASTD: Arabic Sentiment Tweets Dataset," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Lisbon, pp. 2515-2519, 2015.
[28] Nasukawa T. and Yi J., "Sentiment Analysis: Capturing Favorability using Natural Language Processing," *in Proceedings of the 2nd International Conference on Knowledge Capture*, Sanibel Island, pp. 70-77, 2003.
[29] Obaidat I., Mohawesh R., Al-Ayyoub M., Al-Smadi M., and Jararweh Y., "Enhancing the Determination of Aspect Categories and Their Polarities in Arabic Reviews Using Lexicon-Based Approaches," *in Proceedings of Jordan Conference on Applied Electrical Engineering and Computing Technologies*, Amman, pp. 1-6, 2015.
[30] Rushdi-Saleh M., Martín-Valdivia M., Ureña-López L., and Perea-Ortega J., "Bilingual Experiments with an Arabic-English Corpus for Opinion Mining," *in Proceedings of Recent Advances in Natural Language Processing*, Hissar, pp. 740-745, 2011.
[31] Rushdi-Saleh M., Martín-Valdivia M., Ureña-López L., and Perea-Ortega J., "OCA: Opinion Corpus for Arabic," *Journal of the Association for Information Science and Technology*, vol. 62, no. 10, pp. 2045-2054, 2011.
[32] Ryding K., "A Reference Grammar of Modern Standard Arabic," ISBN-13 978-0-521-77151-1, Cambridge University Press, 2005.
[33] SentiStrength. Available at: http://sentistrength.wlv.ac.uk/, Last Visited, 2015.
[34] SocialMention. Available at: http://socialmention.com, Last Visited, 2015.
[35] Thelwall M., Buckley K., and Paltoglou G., "Sentiment Strength Detection for the Social Web," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 1, pp. 163-173, 2012.
[36] Thelwall M., Buckley K., Paltoglou G., Cai D., and Kappas A., "Sentiment Strength Detection in Short Informal Text," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544-2558, 2010.
[37] Witten I. And Frank E., *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Series in Data Management Systems, 2005.

Mohammed Al-Kabi obtained his PhD degree in Mathematics from the University of Lodz/Poland 2001, his master's degree in Computer Science from the University of Baghdad/Iraq 1989, and his bachelor degree in statistics from the University of Baghdad/Iraq 1981. He is an assistant Professor in the Computer Science Department, Faculty of IT, at Zarqa University. He is the author of more than 88 peer-reviewed articles in IR, Big Data, Sentiment Analysis, NLP, Data mining and software Engineering. His teaching interests focus on Information retrieval, big data, web programming, data mining, DBMS (ORACLE and MS Access).

Izzat Alsmadi is an associate professor in the Department Of Computer Science at University of New Haven. He obtained his PhD degree in software engineering from NDSU (USA), his second master in software engineering from NDSU (USA) and his first master in CIS from University of Phoenix (USA). He has several published books, Journals and Conference articles largely in software engineering, data mining, IR and NLP.

Rawan Khasawneh is a full time lecturer in the department of Computer Information Systems at Jordan University of Science and Technology. She obtained her master degree in Management Information Systems from Yarmouk University in Jordan (2013), and her bachelor degree in Management Information Systems from Yarmouk University in Jordan (2011). Khasawneh's research interests include: e-government, social media and sentiment analysis, E-marketing and E-CRM, knowledge management systems and group decision support systems.

Heider Wahsheh obtained his Master degree in Computer Information Systems from Yarmouk University, Jordan, 2012. Between 2013-2015 he worked as a lecturer in the college of Computer Science at King Khalid University, Saudi Arabia. His research interests include information retrieval, sentiment analysis, NLP, data mining, and mobile agent systems