IJCSI International Journal of Computer Science Issues, Volume 16, Issue 2, March 2019
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org https://doi.org/10.5281/zenodo.3234112 

# Towards Detecting Arabic Opinions Polarity on Social Networks

**Norma Bataina** [1]

[1] **Aqaba University of Technology**
**Aqaba, Jordan**

**Abstract**

The way from just read, to read & write web leads social networks users to be excited about distribution their feeling and opinions by several social media that support textual, audio and videos content. Few years ago, these opinions include various purposes such as: education, economic, political, and sport. With the increasing importance of these opinions, it becomes necessary to understand the meaning of them through the automatic analysis of user opinions. Opinion mining or sentiment analysis aims to fetch the attributes i.e. words from the opinions to decide the polarity of the users posts and comments in various topics depending on several linguistic techniques. The classical polarity classes consist of positive, negative and neutral. Since we have also unwanted comments such as advertising and misleading posts, some research studies added spam to the polarity classes. There are several studies focused on enhancing Arabic sentiment analysis but spam detection techniques are still limited. In this paper we used well-known Arabic sentiment analysis corpus and we automatically used term frequency–inverse document frequency (TFIDF) technique. We compared and evaluated three machine learning classifiers for detecting sentiment polarity using prediction evaluation measurements.

***Keywords:*** *Big data, Social network, sentiment analysis, Arabic text classification*

## 1. Introduction

Social networking sites allow people to communicate in a virtual community structure that brings together a common interest, communicating through messages, accessing profiles, and information that an individual can offer. The first appearance of social networking sites in the late twentieth century, and opened a number of sites that spread widely in the world and gathered millions of users. The most popular social network is "Facebook", which appeared in late 2003, and the number of users 2.32 billion in the fourth quarter of 2018 [1]. While the social network "Twitter" which appeared in 2006, the number of active users per month in the fourth quarter of 2018 about 321 million users [2]. Through the profiles displayed by the users of those social network can identify the person's name and basic information such as gender, work, birth, country and interests. Public pages can also be created, enabling access to information or books and references, as well as helping to publicize an issue and mobilize public opinion around it.

Figure 1 shows Middle East social media usage trends revealed which indicates to the increasing reputation of visually social media in the region [3].

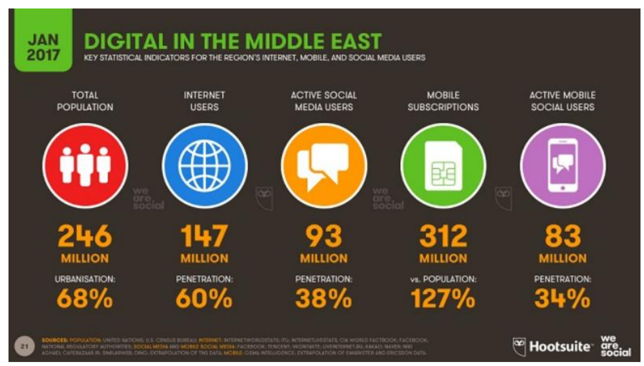


Fig. 1 Middle East social media usage trends revealed [3].

Social networks have many advantages, and they also have many drawbacks. Some educational and social studies have shown that their advantages are rapid in the exchange of information, they help students and researchers, help to create jobs and help to distribute public opinions. They also support economic and social roles. They have contributed to the emergence of the citizen's press, and have strengthened the citizen's ability to express his views and ideas. There are many drawbacks of social networks, which have been warned by educators and sociologists. These are: wasting time, spreading destructive ideas and lies, violating the privacy of others, information chaos, Internet addiction and spam propagation [4]. The shared posts and comments need to understand by automatic analysis. Sentiment analysis refers to mine the language or emoticons in order to extract attributes, compute weights and ranks using particular technique to decide the opinions polarity of the user's posts and comments in various topics. The classical polarity classes consist of positive, negative and neutral [5]. Since that there are unwanted comments such as advertising and misleading posts, some research studies added spam to polarity classes [4]. There are various studies dedicated on improving Arabic polarity detection, but spam detection approaches are still limited. In this paper we use well-known Arabic sentiment analysis corpus [6] that clearly presents the opinion with their topics not like other available datasets. We automatically compute term frequency–inverse document frequency (TFIDF) for each comment in the corpus and then we compare three machine learning classifiers in order to assess the best classifier for detecting Arabic sentiment polarity using prediction evaluation measurements.

The rest of this study is organized as the following: section 2 explores the literature review. Section 3 discusses the Arabic spam polarity. Section 4 presents the proposed methodology. Section 5 presents the experiments and results. Section 6 shows the conclusion and the future plans.

## 2. Literature Review

This section shows various Arabic research papers that are related to our study.

The study of [5] proposed a hybrid technique that uses semantic and machine learning classifiers. The proposed solution focused on Saudi Arabia dialects and used lexicon-based classifier to build the training corpus. The experiments used SVM and the results obtained an accuracy of 84.01%.

The study of [7] proposed a semantic technique to detect user's opinions polarity and business visions from Arabic social networks; MSA and dialects. They presented Arabic Sentiment Ontology (ASO) that consist of several features; i.e. words that discover the user's sentiments. The results showed the usability of the proposed approach in polarity detecting for various tweets on several topics.

The study of [8] developed a tool that detect the Arabic tweets sentiment analysis depending on group of features such as: time of the tweets, stemming, retweets, n-grams and dictionaries. The training corpus contains 8,000 tweets and the results showed that Naïve Bayes obtained the most accurate polarity detection percentages.

The study of [9] presented Lexicon-based opinion mining for Arabic social networks contents. The researchers tested three lexicon methods and generate new lexicon that contains various features. They built Arabic opinion mining system that depends on the enhanced lexicon. The experiment results obtained an accuracy of 74.6%.

The study of [10] dedicated for Arabic tweets opinion mining. They used a corpus of Arabic tweets to detect the polarity as positive or negative sentiments. The experiments employed three supervised machine learning approaches; Support Vector Machine, Naïve Bayes, and Decision Tree. The results of SVM obtained the highest accuracy percentage with 83.16%.

In [11] researchers proposed Modern Standard Arabic and dialects opinion mining tool, called (CNSAMSA-SAT). The proposed tool capable to analyse large number of Arabic posts and comments that pre-collected from several social networks. The developed tool identify the comments polarity as positive, negative, or neutral with an accuracy of 90%.

The study of [12] focused on assessing the political and social content of Arabic opinions in Middle East. Researchers developed a tool for modern standard Arabic (MSA) and dialects. The developed tool detect the comment as subjective or objective and identify the polarity as positive or negative. The conducted test showed that developed tool gain more accurate outputs when it dealing with a particular domain more than general-based Arabic opinions.

The study of [13] assessed the Social Mention API for Arabic language using around 4,000 opinions. They build Arabic and emoticons lexicons, and collect the data collection using them. Naïve Bayes evaluation results for this tool showed an accuracy of 66.2%.

The study of [14] proposed a novel approach for Arabic opinion mining. The new technique employs special arrangement of lexicons features depending on set of rules such as: "contains content" and "equivalent to". The results obtained an accuracy of 89.6% for polarity detection.

Most of the previous studies have focused on identifying opinions polarity based on textual content. The study of [15] proposed a new sentiment analysis tool called DASAP as a detection tool of Arabic Sentiment Analysis Polarity. The main characterizes of [15] are the capability of detecting the polarity for several multimedia opinions such as: textual, audio, images and videos. They conducted several experiments using three machine learning classifiers: K-NN, NB and SVM. Text based detecting method depends on language lexicons while image-based techniques employs content-based image retrieval system and optical character recognition. The results showed 96.1% for the text and 97% for the image polarity detection. While audio and video detecting opinions methods adopting converting their content to the text-based and image-based content.

## 3. Arabic Spam Polarity (ASP)

Digital Arabic content (DAC) [16] is a term that reflects the total of websites and web pages written in Arabic on the Internet. Compared to international content, i.e. content in English, French, German, Japanese, Chinese, etc., Arab content does not exceed three percent of the total global content [17].
There are a lot of Web pages, articles and comments published in Arabic, but the article itself is stagnant and unsaturated to the curiosity of the reader. The most serious problem is that some content contain irrelevant information [4] and [18]. The reason for the inaccuracy of this information may be related to obtaining it from untrustworthy sources or by the fact that the person who translated the news did not do as it should. There is a large percentage of news in Arabic that provides completely false information. The social networks have recently imposed themselves as a platform for news, but it should be confirmed before published as actual information or distributed among users. Useful digital Arabic content may enhance by university sites and E-government portals that used reputation and popularity metrics and publish articles that provide correct and useful information [19], [20] and [21].

Arabic spam is classified into two main parts; content-based [22] and link-based spam techniques [23]. There are few research studies that conducted by the same researchers' team and dedicated to the content Arabic spam [17], [22], [24] and [25]. In these studies, the researchers highlight the Arabic spam methods such as keywords stuffing and employment the attractive words [24]. They build datasets of spam content and evaluate their solutions using supervised machine learning algorithms. The limitations of content-based Arabic spam, that they used binary classification as spam and non-spam [17], [22], [24] and [25].

Arabic spam opinion is considered as a part of Arabic spam content and refers to use spam for illegal largely financial or marketing targets [4]. Spammers may try to counterfeiting polls assessment that is depend on users' opinions, through inserting too many comments. These comments can contain irrelevant and unwanted content [4] and [18]. In [18], authors developed an approach to discover the spam in Arabic opinion by one mining classification technique. The results achieved an accuracy of 99.59% which refers that the performance of the proposed approach was effective.

The study of [4] developed SPAR tool to identify the spam opinions in the Yahoo! - Maktoob social network. SPAR classified comments either spam or non-spam polarity. The spam content consist of two main levels; high levels spam and low level spam. The non-spam are classified as; positive, negative, or neutral based on the language polarity lexicons. SPAR system employs machine learning classification technique to conduct classification and prediction, and yielded high accurate results using SVM.

Arabic Link-based spam refers to link hijacking which is one of the most popular type of link-spam. This type is based on inserting links that point to a spammer controlled page, [17], [23] and [26], which is frequently used in social networking comments. Existing solutions for link-based Arabic spam detection are analyzing link features and using blacklists.

## 4. Methodology

The adopted methodology is summarized in the following main points:

1. Select a benchmark corpus of textual Modern Standard Arabic (MSA), and dialects comments of social networks [6].

2. Remove Arabic stop words.

3. Tokenize the words, normalize them and automatically compute term frequency–inverse document frequency (TFIDF).

4. Use three machine learning classifiers and compare their prediction evaluation results in order to find the appropriate classifier for detecting the sentiment polarity.

IJCSI International Journal of Computer Science Issues, Volume 16, Issue 2, March 2019
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org https://doi.org/10.5281/zenodo.3234112 

### 4.1 Well-known Arabic Opinions Corpus

There are various free Arabic opinions corpuses such as: [27] and [28], but they do not contain Arabic spam opinions. The study of [6] considered as a well-known Arabic social network corpus that consist of positive, negative, neutral and spam opinions. This corpus written in both Modern Standard Arabic (MSA) and different Arabic dialects. It contains 250 topics and 1,442 opinions of several domains. Since that spammers use financial or marketing opinions that could be similar to non-spam opinions in the Economy field. So we will consider Economy domain in this work as a case study. We performed preprocessing techniques for all opinions that includes: stop words removal, tokenize the words, normalize them and automatically compute term frequency–inverse document frequency (TFIDF). Table 1 shows examples of the Arabic features of the used Arabic corpus.

Table 1: Examples of Arabic features in the used Arabic corpus

| *Arabic Feature* | *English Translation* | *Feature Polarity* |
|---|---|---|
| ابداع | Creativity | Positive |
| اعجاب | Like | Positive |
| خسارة | Loss | Negative |
| حزن | Sad | Negative |
| * | * | Neutral |
| ** | ** | Spam |

* Neutral polarity could be when positive and negative features are missed or equals together.
** Spam polarity depends on the comment features

## 5. Experiments and Results

In order to evaluate Arabic opinions polarity detection, we select three well-known machine learning classifiers; Naïve Bayes (NB), Support Vector Machine (SVM) and Decision Tree (DT) [29]. We used; True Positive (TP), False Positive (FP), Precision (P), Recall (R) and F-Measure (F-M) as shown in the following equations [30].

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

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

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

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

The experiments showed that the Naive Bayes (NB) obtained a weighted average result with accuracy of 61.3%. Table 2 shows the prediction evolution results of NB.

Table 2: NB prediction evolution results

| *Class* | *TP* | *FP* | *P* | *R* | *F-M* |
|---|---|---|---|---|---|
| Positive | 0.50 | 0.14 | 0.40 | 0.50 | 0.44 |
| Negative | 0.58 | 0.21 | 0.78 | 0.58 | 0.66 |
| Neutral | 0.60 | 0.16 | 0.35 | 0.60 | 0.44 |
| Spam | 0.90 | 0.03 | 0.90 | 0.85 | 0.83 |
| Avg | 0.61 | 0.17 | 0.66 | 061 | 0.62 |

NB obtained reasonable detection results for positive, negative and neutral classes, and it achieved the highest accuracy results in detecting spam polarity.

Applying SVM enhanced the accuracy with 70.7%. Table 3 presents the prediction evolution results for the four polarity classes.

Table 3: SVM prediction evolution results

| *Class* | *TP* | *FP* | *P* | *R* | *F-M* |
|---|---|---|---|---|---|
| Positive | 0.16 | 0.01 | 0.66 | 0.16 | 0.26 |
| Negative | 0.97 | 0.65 | 0.66 | 0.97 | 0.79 |
| Neutral | 0.10 | 0 | 1 | 0.10 | 0.18 |
| Spam | 0.80 | 0 | 1 | 0.80 | 0.88 |
| Avg | 0.70 | 0.37 | 0.75 | 0.70 | 0.64 |

SVM promoted the weighted average results comparing to NB, but in the details we found that SVM cannot detect the polarity for both positive and neutral classes.

The third evaluated classifier is Decision Tree (DT), which achieved an accuracy of 60%. Table 4 shows the prediction evolution results of DT.

Table 4: DT prediction evolution results

| *Class* | *TP* | *FP* | *P* | *R* | *F-M* |
|---|---|---|---|---|---|
| Positive | 0.08 | 0.04 | 0.25 | 0.08 | 0.12 |
| Negative | 0.86 | 0.71 | 0.61 | 0.86 | 071 |
| Neutral | 0.10 | 0.03 | 0.33 | 0.10 | 0.15 |
| Spam | 0.60 | 0.03 | 0.75 | 0.60 | 0.66 |
| Avg | 0.60 | 0.42 | 0.53 | 0.60 | 0.54 |

DT obtained reasonable accuracy detection percentage for spam detection and high accuracy results for negative opinions. DT cannot predict the correct polarity for both positive and neutral as the same case of SVM. So, we conclude from Table 3 and Table 4 that SVM and DT are not the appropriate classifiers to detect the polarity for all classes of the Economic domain.

## 6. Conclusions and Future Work

There are many research works dedicated for improvement Arabic opinion mining but Arabic spam detection studies are still limited. In this study we used a well-known Arabic sentiment analysis corpus and we applied term frequency–inverse document frequency (TFIDF) technique beside preprocessing methods. We compared and assessed three machine learning classifiers for detecting sentiment polarity using prediction evaluation measurements. The obtained results showed that NB is appropriate to detect all the polarity classes with reasonable accuracy percentages. We plan as a future work to cover more domains in the experiments and employ deep learning algorithms to detect the polarity of the Arabic opinions.

## References

[1] "Facebook users worldwide 2018 | Statista", Statista, 2019. [Online]. Available: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide. [Accessed: 04- Mar- 2019].

[2] "Twitter: number of active users 2010-2018 | Statista", Statista, 2019. [Online]. Available: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/. [Accessed: 04- Mar- 2019].

[3] "Middle East Social Media Usage Trends Revealed - Arabian Gazette", Arabian Gazette, 2019. [Online]. Available: https://arabiangazette.com/middle-east-social-media-usage-trends-revealed. [Accessed: 04- Mar- 2019].

[4] H. A. Wahsheh, M. N. Al-Kabi, and I. M. Alsmadi, "SPAR: A system to detect spam in Arabic opinions" In 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2013, pp. 1-6.

[5] H. Aldayel and A. Azmi, "Arabic tweets sentiment analysis – a hybrid scheme", Journal of Information Science, Vol. 42, No. 6, 2016, 782-797.

[6] M. Al-Kabi, M. Al-Ayyoub, I. Alsmadi, and H. Wahsheh, "A prototype for a standard arabic sentiment analysis corpus". Int. Arab J. Inf. Technol., Vol. 13, No. 1A, 2016, pp.163-170.

[7] S. Tartir and I. Abdul-Nabi, "Semantic sentiment analysis in Arabic social media", Journal of King Saud University-Computer and Information Sciences, Vol. 29, No. 2, 2017, pp. 229-233.

[8] M. El-Masri , N. Altrabsheha , H. Mansourb and A. Ramsay, "A web-based tool for Arabic sentiment analysis", In: 3rd International Conference on Arabic Computational Linguistics, Dubai, United Arab Emirates, 2017, pp. 38–45.

[9] N. Abdulla, R. Majdalawi, S. Mohammed, M. Al-Ayyoub, and M. Al-Kabi, "Automatic Lexicon Construction for Arabic Sentiment Analysis", In: 2014 International Conference on Future Internet of Things and Cloud (FiCloud), 2014, pp. 547-552.

[10] G. AL-Sukkar, I. Aljarah and H, Alsawalqah, "Enhancing the Arabic Sentiment Analysis Using Different Preprocessing Operators", In: Proceedings of the New Trends in Information Technology (NTIT-2017), 2017, pp. 113-117.

[11] M. Al-Kabi, A. Gigieh, I. Alsmadi, H. Wahsheh and M. Haidar, "An opinion analysis tool for colloquial and standard Arabic". In The Fourth International Conference on Information and Communication Systems (ICICS 2013), 2013, pp. 23-25.

[12] M. N. Al-Kabi, A. H. Gigieh, I. M. Alsmadi, H. A. Wahsheh, and M. M. Haidar, "Opinion Mining and Analysis for Arabic Language". International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 5, No. 5, 2014, pp. 181-195.

[13] M. Al-Kabi, N. Al-Qudah, I. Alsmadi, M. Dabour, and H. Wahsheh, "Arabic/English sentiment analysis: an empirical study", In the Fourth International Conference on Information and Communication Systems (ICICS 2013), 2013, pp.23-25.

[14] S. Siddiqui, A. Abdel Monem and K. Shaalan, "Sentiment Analysis in Arabic", In International Conference on Applications of Natural Language to Information Systems, 2016, pp. 409-414.

[15] M. N. Al-Kabi, H. A. Wahsheh and I. M. Alsmadi, "Polarity classification of Arabic sentiments", International Journal of Information Technology and Web Engineering (IJITWE), Vol. 11, No. 3, 2016, pp. 32-49.

[16] "Digital Arabic Content", United Nations Economic and Social Commission for Western Asia, 2019. [Online].

Available: https://www.unescwa.org/sub-site/digital-arabic-content. [Accessed: 04- Mar- 2019].
[17] H. A. Wahsheh, M. N. Al-Kabi and I. M. Alsmadi, "A link and Content Hybrid Approach for Arabic Web Spam Detection", International Journal of Intelligent Systems and Applications, Vol. 5, No. 1, 2012, pp. 30-43.
[18] A. Abu Hammad and A. El-Halees, "An Approach for Detecting Spam in Arabic Opinion Reviews", International Arab Journal of Information Technology (IAJIT), Vol. 12, No. 1, 2015, pp. 9-16.
[19] Y. Tashtoush, A. Zaidan and I. Alsmadi, "Implications for Website Trust and Credibility Assessment", International Journal of E-Entrepreneurship and Innovation, Vol. 3, No. 4, 2012, pp. 17-33.
[20] H. A. Wahsheh I. M. Alsmadi and M. N. Al-Kabi, "The evaluation of trust and credibility metrics: Websites of Jordanian universities and e-government portals as a case study", In 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2013, pp. 1-6.
[21] I. W. Kamal, H. A. Wahsheh, I. M. Alsmadi and M. N. Al-Kabi, "Evaluating Web Accessibility Metrics for Jordanian Universities", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 7, 2016, pp. 113-122.
[22] M. Al-Kabi, H. Wahsheh, I. Alsmadi, E. Al-Shawakfa, A. Wahbeh and A. Al-Hmoud, "Content-based analysis to detect Arabic web spam", Journal of Information Science, Vol. 38, No. 3, 2012, pp. 284-296.
[23] H. A. Wahsheh, M. N. Al-kabi and I. M. Alsmadi, "Evaluating Arabic spam classifiers using link analysis", In Proceedings of the 3rd International Conference on Information and Communication Systems , 2012, pp. 1-5.
[24] H. Wahsheh, I. Alsmadi and M. Al-Kabi, "Analyzing the popular words to evaluate spam in Arabic web pages", IJJ: The Research Bulletin of JORDAN ACM–ISWSA, Vol. 2, No. 2, 2012, pp.22-26.
[25] H. A. Wahsheh, M. N. Al-Kabi and I. M. Alsmadi, "Spam detection methods for Arabic web pages", In First Taibah University International Conference on Computing and Information Technology-Information Systems ICCIT, 2012, pp. 486-490.
[26] H. A. Wahsheh, M. N. Al-Kabi and I. M. Alsmadi, "Detecting Arabic Cloaking Web Pages Using Hybrid Techniques", International Journal of Hybrid Information Technology, Vol. 6, No. 6, 2013, pp. 109-126.
[27] N. Abdulla, M. Mahyoub, M. Shehab and M., Al-Ayyoub, "Arabic Sentiment Analysis: Corpus-based and Lexicon-based", In IEEE conference on Applied Electrical Engineering and Computing Technologies (AEECT 2013), 2013, pp. 1-6.
[28] M. El-Dien Aly, "Datasets -", Mohamedaly.info, 2019. [Online]. Available: http://www.mohamedaly.info/datasets. [Accessed: 04- Mar- 2019].
[29] T. Pranckevičius and V. Marcinkevičius, "Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification", Baltic Journal of Modern Computing, Vol. 5, No. 2, 2017, pp. 221-232.
[30] I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann Series in Data Management Systems, 2005.

**Norma Bataina** obtained her Master degree in Computer Information Systems (CIS) from Yarmouk University, Jordan, 2012. She is working as a lecturer in Aqaba University of Technology (AUT). Her research interests include: NLP and Data Mining.