Survey on Arabic Sentiment Analysis in Twitter

Sarah O. Alhumoud, Mawaheb I. Altuwaijri, Tarfa M. Albuhairi, Wejdan M. Alohaideb

***Abstract*—**Large-scale data stream analysis has become one of the important business and research priorities lately. Social networks like Twitter and other micro-blogging platforms hold an enormous amount of data that is large in volume, velocity and variety. Extracting valuable information and trends out of these data would aid in a better understanding and decision-making. Multiple analysis techniques are deployed for English content. Moreover, one of the languages that produce a large amount of data over social networks and is least analyzed is the Arabic language. The proposed paper is a survey on the research efforts to analyze the Arabic content in Twitter focusing on the tools and methods used to extract the sentiments for the Arabic content on Twitter.

***Keywords*—**Big Data, Social Networks, Sentiment Analysis.

I. Introduction

DATA has become the currency of this era as continually increasing in size and value. The available data online is doubling in size every two years [1]. While the amount of data online that was generated in 2013 was 4.4 Zettabytes (ZB) it is anticipated that in 2020 data created will reach 44 ZB [1]. Individual users are the main source of these data with 75 percent of overall produced data [2].

Big data is characterized by three domains that are called 3'V which are *Variety, Velocity,* and *Volume*. Variety means the variation of data available online that include both structured and unstructured data such as emails, videos, audios, images, click streams, logs, posts, search queries. Velocity refers to how the processing and storing of these huge and complex data need to be fast in order to accommodate the increasing and continuous requests. Volume indicates the size of the generated data online as presented previously [3].

Social networks such as Twitter and Facebook have become popular means for cyber communication among societies. Since the foundation of Twitter in 2006 it has provided the ability to freely, easily, and instantaneously express, reach, and share opinions and feelings in public in an SMS style. Twitter is one of the largest platforms that is full of sentiment. It is a micro blogging site, which contains tweets; each tweet has "140 characters or less". There are over 1 billion Tweets every 72 hours from more than 140 million active users on Twitter [1]. That make it quintessential example of "big data". The Arabic content on the internet has increased in volume especially after the evolution of social networks. In Twitter, there are more than 6.5 Million Arabic users [4] who produce more than 10.8 million tweets per day [4].

Big data that spreads online every second will continually become bigger, hiding a real value. Using big data analytic techniques this value could be extracted. Big data analytics is the process of analyzing a huge amount of data that includes both *structured data* (such as data inside data bases) and *unstructured data* (such as data on web) to get valuable information that can be used for taking decisions through the use of advanced analysis techniques such as *Natural Language Processing (*NLP), *Machine Learning*, and *Predictive Analytics* [5]. *Sentiment Analysis* (SA) is a one of the NLP concepts [5]. This field is used in order to extract the sentiment out of text giving useful information about the author and his/her tendency towards a specific topic. Analyzing Arabic language is much complex compared with English. Next section explains the reasons behind this complexity.

A. The Challenges of Analyzing Arabic Text

Arabic language is complex to analyze because of the properties it has. Following points will explain the properties of Arabic language and their impact on the analysis process.

1. Every country/ part of a county has its own version or dialect of Arabic. That means there are different dialects of Arabic text available online that could hold different meaning. Resulting in high complexities when analyzing sentiments with different dialects.
2. The root for Arabic words could have multiple forms based on the context such as (كلام, كلمات, يتكلم).
3. Arabic have the property of having the same word spelling but with different meaning depending on its punctuation such as (يُعلم) which means teaching and (يَعلم) which means know.
4. The presence of words such as (لكن) can cause a sentence to have two opposite sentiments at the same time.

This paper proposes a survey on some of the works that have been done to analyze the Arabic sentiment in Twitter. The paper will discuss the following: Section II will review Arabic data sources that have been used for sentiment analysis. Within Section III general process flow of sentiment analysis is depicted, describing the process of sentiment analysis in Twitter for Arabic. This section further includes four subsections: Section *A*, the collecting process is described along with the used tools and methods. Section *B*, gives a deep insight on the techniques that are used to preprocess the content to be analyzed. Then, Section *C* shows how the filtering process works. Section *D*, describes all the used classification tools and algorithms. After that, Section IV is a

S. O. Alhumoud is an assistant Professor at the Computer Science Department, College of Information and Computer Science and Deputy Dean of Information Technology Deanship Imam Mohammad Bin Saud University, Saudi Arabia (e-mail: S.alhumoud@ccis.imamu.edu.sa)

M. I. Altuwaijri, T. M. Albuhairi, and W. M. Alohaideb are with the College of Information and Computer Science Deputy Dean of Information Technology Deanship Imam Mohammad Bin Saud University (e-mail: mituwaijri@sm.imamu.edu.sa, tmbuhairi@sm.imamu.edu.sa, wmohaideb@sm.imamu.edu.sa).

literature review on Arabic sentiment analysis in Twitter. Finally, Section V is the conclusion of the survey.

II. Data Sources

Arabic text sources on the web are varied and similar to the English sources. For the purpose of sentiment analysis the sources that can be considered are those full of sentiment and opinions. People normally tend to express their opinion about a subject on the social media such as Facebook, Twitter and Blogs. Sentiment analysis on Arabic did not start from Twitter. It has been done on number of the social media and networks. Here is a description for each data source that was used for Arabic sentiment analysis.

Twitter is a "Social networking and microblogging service utilizing instant messaging, SMS or a web interface" according to the Twitter website [6]. Each message which is called a tweet in Twitter term is composed of 140 letter/ characters long. Users use a hashtags to express their opinion in a specific topic.

Facebook is a social network service where each user has its own profile. Users post their opinions and feelings and share it with their friends and family [7]-[9].

Aljazeera's web site is a news website where users can view the news and comment with their opinion on the topic [10].

Yahoo!-Maktoob was the first Arab mail services. It was sold to Yahoo!. People can share sentiments and opinions about the news that are published on Yahoo!-Maktoob [11].

Blogs are daily logs for their authors. It contains information about a specific topic that their authors interest on. Usually they use it to express their personal opinion about products, political view or other interest they have [12].

III. General Process Flow of Sentiment Analysis

Sentiment analysis of data can be accomplished by passing through four main steps displayed in Fig. 1. It starts by specifying keyword to analyze the people's sentiment towards it. After that is the collecting stage and there are different ways to collect target tweets. These collected tweets will be stored in a dataset. After collecting the tweets, the next step is preprocessing the tweets which will remove all unrelated contents and get the Arabic text only. Then, the filtering step that involves removing all words that do not affect the text meaning. The next stage is to classify the content to positive and negative. The final step is to get the overall sentiment of all the collected tweets. These stages will be described more in detail in the following subsections.

A. Collecting

The first step in the sentiment analysis process is collecting tweets by specifying a keyword to retrieve all tweets that are related to that keyword. Tweets can be collected from different sources. One way is tweet crawler that collects collection of linked tweets by querying the Twitter web service. Another way using Twitter *Application Program Interface* (API) which is provided by Twitter that give developers the ability to use the Twitter's functions such as retrieving tweets with the selected keyword and language. NODEXL tool by Microsoft is another tool that used for collecting tweets. It is a tool that supports multiple social network data providers that import graph data into the Excel spreadsheet. The collected tweets will be stored in a dataset in order to classify it.

B. Preprocessing

Preprocess is a technique used to clean text from unsentimental contents, such as user-names, pictures, hashtags, URLs and all non-Arabic words or sometimes gagging these content with a unified name [13]-[15]. This process referred to as tagging [15]. Tagging is the process of marking unsentimental content in a tweet that does not have any impact on the tweet sentiment. These will differ in type and number. For example, a URL link may be replaced with a URL tag, the username which is a word that appears after the symbol "@" in Twitter will be tagged with *username* and the word that appears after a hash "#" and do not relate to the topic will be that tagged with *hashtag*. Also since Twitter users use symbols such as "(:" and "☺" to express their opinions these emoticons expresses valuable information to the sentiment. Therefore in order to extract the sentiment out of the emotions they are tagged as well. For example an emotion is tagged as *happy* if the used symbols are ":)", ":')" and tagged as *sad* if the used symbols are ":(", ":'(" [15]. Emotions tags will affect the classify process since they hold a sentiment.

C. Filtering

After the preprocessing stage the outcome will be only text. Filtering stage includes other steps that are needed to remove all the words that do not affect or relate to the meaning. Moreover, in this stage misspellings are corrected, and the repeated letters of the text are removed.

1. Misspelling

Users may misspell some words because of their fast typing or even the weakness in spelling skills. In order to overcome this problem misspelling can be corrected manually or by using tools.

2. Repeated Letters

Users express their feeling about something like a product and they may use a word with some repeated letters in it such as "كثييير". Removing repeated letters from the words is important for the classification process to recognize the word. For example, "كثيييير" will be corrected to be "كثير" where the letter "ي" is repeated 4 times. The repeated letters will be replaced by one latter of those repeated letters. Using *naive algorithm* which simply counts the number of letters in each word if the letter repeated then the repeated letters will be removed and keep just one letter [14].

3. Stop Words

Stop Words are a group of words that do not affect the meaning of the text, such as *prepositions*. The problem here is the limitation of the built in stop words lists. One available Arabic stop word list is Khoja stemmer tool [14]. Also if the tweet that need to be analyzed do not use a *Modern Standard Arabic* (MSA) then there will be a need to add extra stop words to the list of stop words from different Arabic dialects. Arabic dialects that have been done so far studied the Egyptian dialect [13] and Jordanian dialect [14].

4. Normalization

Normalization is the process of replacing similar letters that are used interchangeably by one of them and removing none letters from the words, the normalization conditions are as follows:

1. Remove any punctuation from the string such as (. “” ; ').
2. Remove any diacritics (short vowels) such as (⊂ َ ُ ◌ً ◌ٌ).
3. Remove non-letters from the word such as (+ = ~ $).
4. Replace “أ“,”إ”, and “ آ” with bare alif "ا" regardless of position in the word.
5. Remove Tatweel “ـ” (for example, using Tatweel the word “كتب” may look like "كتـــب".
6. Replace final “ى” with "ي".
7. Replace final “ة” with “ه” .
8. If a word starts with “ء” then replace it with “ا”.
9. Replace “ؤ” and “ؤ” with “و”.
10. Replace “ئ” and “ىء” with“ي” [15].

D. Classifying

This stage represents the final stage where each tweet will be classified as positive and negative by the classifier. These tweets will be annotated manually to be compared with classifier results in order to examine the accuracy of the classifier.

The classifiers in general are categorized under two approaches *supervised* and *unsupervised*. In the supervised or corpus-based, the machine learning classifiers are used such as *Support Vector Machine* (SVM), *Naïve Bayes* (NB*), Decision Tree* (D-Tree), *K-Nearest Neighbor* (KNN). This type works by first training the classifier using a training dataset. This dataset contains tweets with positive and negative labels. It teaches the algorithms which tweet is positive and which is negative. After training the classifier it will be able to build a general model to use in classifying a new dataset. Classification using feature vector [13] is considered under supervised approach. For this method, features must be extracted first and this can be done using *unigram*, *bigram*, or *trigram*. Unigram technique means dealing with each word as a single unite without considering what is surrounding [14]. The extracted feature must has been exceeded a threshold value to be extracted. The supervised approach tends to have more accuracy results when it tested on a dataset that have the same domain of the training dataset. It accuracy can be increased by training with huge dataset and extracting multiple features.

The second approach of classification is the unsupervised approach. Unsupervised or lexicon-based uses a lexicon or dictionary is the second type of classification where will be no training step. The classifier will classify a dataset directly using a dictionary of word. Each word has a polarity (+1, -1 or 0 for positive, negative or neutral, respectively). The dictionaries could be built manually or it can be built beforehand. Since there is a lack in the Arabic word dictionaries the researchers often built their own dictionary. The dictionary (lexicon) may contains more than the polarity of the word according to the classifier wither to consider the *Part Of Speech* (POS) of the word within the tweet such as verb, noun, adjective, adverb, and others [16], or just use the polarity inside the dictionary (lexicon). Table I shows the main differences between the performing sentiment analysis using supervised and unsupervised approaches.

IV. Arabic Sentiment Analysis in Twitter

The work that has been done regarding Arabic sentiment analysis in Twitter is limited. Papers could be categorized based on the used classification approach. Each effort that was done in researches [13], [17], [15], [18] used a supervised approach.

Reference [13] has used a supervised approach. They examined two algorithms: SVM, and NB. The dataset has been collected using Twitter API. The used feature extraction are unigram and bigram. They have used a tool that uses Twitter APIs to collect tweets.

Paper [17] proposed a system for sentiment analysis using a machine learning supervised approach. They used SVM algorithms. Their 3015 Arabic tweets where collected from TAGREED corpus.

Paper [15] used five algorithms of supervised approach which are NB, SVM, Maximum Entropy, Bayes Net, J48 decision tree. With the help of Twitter filter stream API they were able to collect Arabic tweets.

Authors of [18] have built KDOEST system (Kuwaiti-Dialect Opinion Extraction System from Twitter). They used a supervised approach using SVM and decision Tree algorithms to classify the Tweets. They extracted features for the machine by dividing the Kuwaiti terms into classes such as happiness class. These classes are categorized under positive or negative.

Another supervised approach was used in [19] where they used RapidMiner to classify their collected tweets using both NB and D-Tree algorithms. They studied the impact of considering the emotion faces (emoji) that are used widely by Twitter’s users. Their approach showed that classification with emotion faces has raised the accuracy from 58.28% to 63.79%.

Another work has been done for the Jordanian dialect was done by [20]. They used RapidMiner for preprocessing and filtering stages. To annotate their tweet they hired the CrowdSource website that displays the tweets on users and they do the annotation process. For the classification process, RapidMiner is used to examine three algorithms which are SVM, NB, and KNN.

While in [14] they examined both approaches the supervised and the unsupervised. In the unsupervised they built their lexicon manually using SentiStrength website. And enhance it through adding the synonyms of the word. The supervised examination has been done using RapidMiner software. Their dataset consist of both MSA and Jordanian dialect. They have been collected using tweets crawler.

Authors of [21] have used a lexicon of 200 words. That size has affected their results. They found that the stemmer is useful to reduce the size of the lexicon since multiple words will have the same root at the lexicon. They used a small dataset of 100 tweets. Their accuracy was 73%.

Finally in this literature the effort of [16] they used unsupervised approach the built their lexicon form a seed lexicon of 380 words and starts to extend it. After that they used two algorithms to annotate these words. Finally they hired two methods to calculate the sentiment of tweets collected by Twitter search API. The first one is the sum, which sums the polarity of each word in a tweet, and the second is the double polarity method where each word in the tweet has positive and negative weight. Table I compares between techniques that were used in each work.

TABLE I
THE DIFFERENCES BETWEEN THE SUPERVISED AND UNSUPERVISED APPROACH

Factors	Supervised	Unsupervised
Classification domain	Restricted to a the trained topic	No limitation.
Accuracy	Depends on: -Dataset related to the domain of training dataset. -Training dataset size. -Features variation.	The size of the data dictionary (lexicon).
Types of features	-Part Of Speech (POS). -Word Frequency. -Sentiment Words.	Part Of Speech (POS) tags.

TABLE II
SUMMARY OF WHAT METHODS HAVE BEEN DONE FOR ARABIC SENTIMENT ANALYSIS IN TWITTER

Paper	Collecting	#Tweets	Preprocessing & Filtering	Classifying
[13]	Tool to get tweets using Twitter's APIs.	-1000 tweets. -500 positive. -500 negative. -Egyptian dialect.	Removing: -Usernames. -Pictures. -Hash tags. -URLs. -Arabic words.	**Supervised approach** -Used algorithms: -SVM. -NB. -Trained using the frequency of the unigrams. -Trained using a combination unigrams and bigrams.
[14]	Tweet crawler.	-2000 tweets. -1000 positive. -1000 negative. -MSA and Jordanian dialect.	-MS Word for misspelling. -Naive algorithm for repeated letters. -Normalization. -The Khoja stemmer tool for Stop Words.	**Unsupervised approach** -Building lexicon manually from SentiStrength website. -Add the synonyms of each word to enhance the lexicon. -Use unigram technique for feature extraction. **Supervised approach** -Used algorithms: -SVM -NB -D-Tree -KNN Using RapidMiner software.
[17]	TAGREED (TGRD) is a corpus of 3015 Arabic tweets.	-1466 MSA tweets. -1549 DA tweets. -MSA and DA.	-Tokenize the text automatically. -Part of Speak tagging.	**Supervised approach** -Used algorithms: -SVM.
[15]	Twitter filter stream API.	-2861 Tweets. -612 positive. -513 negative. -848 neutral.	-Tag adding. -Normalization.	**Supervised approach** -Used algorithms: -SVM -NB -Maximum Entropy -Bayes Net -J48 decision tree.
[20]	Twitter API	-1000 Tweets	-Stemming -Tokenizing. -Filtering -Stop words -Using RapidMiner	**Supervised approach** -Used algorithms -NB -k-nearest classifier (k-NN) -SVM
[16]	Twitter Search API	-500 tweets. -155 positive. -310 negative. -35 neutral.		**Unsupervised Approach** -Building lexicon using a seed 380 words manually. -The lexicon words tagged with their POS. -Set the polar sentiment for each word on the lexicon. -Two algorithms used: -Sum method -Double polarity method
[18]	Twitter API	-340,000 tweets. -Kuwaiti dialect	-Tokenization using Stanford Arabic tokenizer.	**Supervised approach** -Used algorithms: -SVM. -D-Tree.
[19]	Twitter API	-3000 tweets.	-Normalization -Convert the emotion to synonym sentiment word. -Removed username, URL, hashtags. -Remove stop words.	**Supervised approach.** -Used algorithms: -NB. -KNN. Using RapidMiner software.
[21]	NODEXL tool (Microsoft)	-100 tweets. -40 positive. -69 negative. -Informal Arabic.	-Tokenization. -Remove stop words. -Stemming using Khoja stemmer.	**Unsupervised approach** -Lexicon of 200 words.

V. CONCLUSION

Arabic content in the web continue to increase especially in social networks. Social networks enclose a lot of information that is valuable for decisions making. To take the advantage of this content and make it valuable, analysis techniques must be applied. This paper surveys techniques that were presented in 9 papers published in the IEEE, Science Direct, and ACM. This paper also shows what have been done in the Arabic sentiment analysis in twitter. As a future work we will build a sentiment analyzer for Arabic tweets with the help of what was presented here.

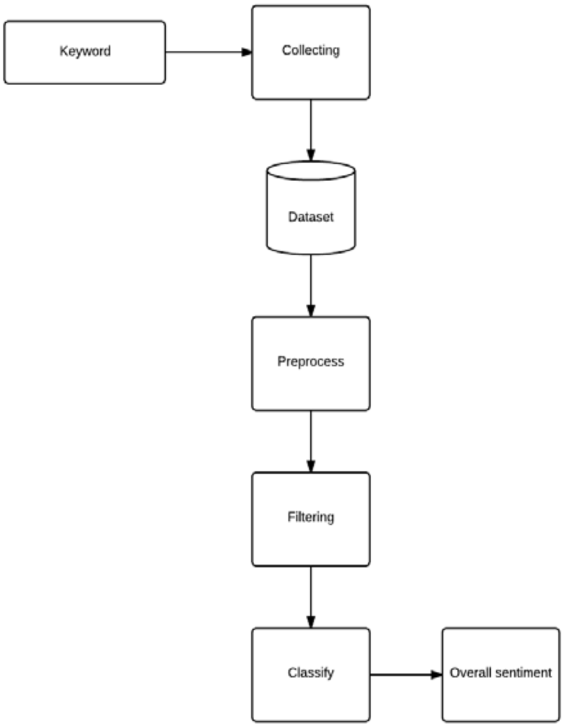

Fig. 1 The general process flow of sentiment analysis

REFERENCES

[1] J. Gantz and D. Reinsel, "Digital Universe Study: Extracting Value from Chaos," EMC2, June 2011. (Online). Available: Internet: http://www.emc.com/leadership/programs/digital-universe.htm (Accessed 6 Nov 2014).

[2] "The 2011 IDC Digital Universe study sponsored by EMC," (Online). Available: http://www.emc.com/collateral/about/news/idc-emc-digital-universe-2011-infographic.pdf (Accessed 6 Nov 2014).

[3] S. Sagiroglu and a. D.Sinanc. "Big data: A review," in Proc. CTS, 2013, pp. 42 – 47.

[4] "Social Media Usage in Middle East – Statistics and Trends (Infographic)," Go-Gulf, 4 Jun 2013. (Online). Available: http://www.go-gulf.com/blog/social-media-middle-east (Accessed 6 Nov 2014).

[5] B. Liu. (2012, Apr 22). Sentiment Analysis and Opinion Mining, (1st edition). (On-line). Available: http://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf (Des 22, 2014]).

[6] "About," Twitter, (Online). Available: https://about.twitter.com/what-is-twitter. (Accessed 6 Nov 2014).

[7] J. Akaichi. "Social Networks' Facebook' Statutes Updates Mining for Sentiment Classification," in Porc. SOCIALCOM, 2013, pp. 886 - 891.

[8] R. Khasawneh, H. Wahsheh, M. Al Kabi and I. Aismadi. "Sentiment analysis of arabic social media content: a comparative study," in Porc. ICITST, 2013, pp. 101 - 106.

[9] M. Itani, B. A. U. B. L. Math. & Comput. Sci. Dept., L. Hamandi, R. Zantout and I. Elkabani. "Classifying sentiment in arabic social networks: Naïve search versus Naïve bayes," in Porc. ACTEA, 2012, pp. 192 - 197.

[10] A. Mountassir, M. 5. U. R. M. ALBIRONI Res. Team, H. Benbrahim and I. Berrada. "Some methods to address the problem of unbalanced sentiment classification in an arabic context," in Porc. CIST, 2012, pp. 43 - 48.

[11] M. Al-Kabi, Z. J. Zarqa Univ., N. Abdulla and M. Al-Ayyoub. "An analytical study of Arabic sentiments: Maktoob case study," in Porc. ICITST, 2013, pp. 89 - 94.

[12] J. Varlack. "What are Blogs?," MedNews Blog, 2 March 2009 . (Online). Available: http://www.mednet-tech.com/newsletter/blogs/what-are-blogs. (Accessed 6 Nov 2014).

[13] A. Shoukry and a. A. Rafea. "Sentence Level Arabic Sentiment Analysis," in Proc. CTS, 2012, pp. 546 – 550.

[14] N. Abdulla, N. Ahmed, M. Shehab and a. M. Al-Ayyoub. "Arabic Sentiment Analysis: Lexicon-Based and Corpus-Based," in Proc. AEECT, 2013, pp. 1 – 6.

[15] S. Ahmed and G. Qadah. "Key Issues in Conducting Sentiment Analysis on Arabic Social Media Text," in Porc. IIT, 2013, pp. 72 – 77.

[16] S. El-Beltagy and A. Ali. "Open Issues in the Sentiment Analysis of Arabic," in Porc. IIT, 2013, pp. 215-220.

[17] M. Abdul-Mageed, S. K¨ubler and a. M. Diab. "SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social Media," in Proc. WASSA, 2012, pp. 19-28.

[18] J. Salamah and a. A. Elkhlifi. "Microblogging Opinion Mining Approach for Kuwaiti Dialect," in Proc. ICCTIM, Dubai, 2014.

[19] S. Al-Osaimi and a. K. Badruddin. "Role of Emotion icons in Sentiment classification of Arabic Tweets," in Porc. MEDES '14, 2014, pp.167-171.

[20] R. Duwairi, R. Marji, N. Sha'ban and S. Rushaidat. "Sentiment Analysis in Arabic Tweets," in Porc. ICICS, 2014, pp. 1 - 6.

[21] L. Albraheem and a. H. Al-Khalifa. "Exploring the problems of Sentiment Analysis in Informal," in Proc. IIWAS '12, 2012, pp. 415-418.