Issues of Dialectal Saudi Twitter Corpus

Meshrif Alruily
College of Computer and Information Sciences, Jouf University, Saud Arabia

Abstract: *Text mining research relies heavily on the availability of a suitable corpus. This paper presents a dialectal Saudi corpus that contains 207452 tweets generated by Saudi Twitter users. In addition, a comparison between the Saudi tweets dataset, Egyptian Twitter corpus and Arabic top news raw corpus (representing Modern Standard Arabic (MSA) in various aspects, such as the differences between formal and colloquial texts was carried out. Moreover, investigation into the issues and phenomena, such as shortening, concatenation, colloquial language, compounding, foreign language, spelling errors and neologisms on this type of dataset was performed.*

Keywords: *Microblogs, tweets, Saudi colloquial, corpus and modern standard Arabic.*

Received January 27, 2018; accepted August 13, 2018
https://doi.org/10.34028/iajit/17/3/10

1. Introduction

Social media and microblogs, such as Twitter and Facebook have become a platform and the first choice for people wishing to write about their daily life, share information and search for real-time news events [11]. Twitter is the most popular social media tool among Internet users worldwide with 500 millions tweets per day [19, 22]. Twitter is a social media tool where users are able to send very short messages, known as tweets [1]. Previously, the total number of characters for every tweet was 140 but the length has recently changed, and the number of characters has expanded to 280.

The importance of social media platforms is clear for business leaders, governments and organizations. Many researches, such as sentiment analysis have been conducted on the data produced by social media users to measure the opinions of people in various aspects, such as a specific product, companies' services and hotels [12, 29]. In addition, many studies for monitoring real-time events utilizing social media data have been performed; for example, using Twitter streams or messages for traffic detection, tracking flu infections, and monitoring food-borne outbreaks [13, 14, 19, 20]. The Arab world has been affected by these recent developments in technology. For instance, the total number of Twitter users in Arabic countries now stands at more than 11 million, with 27.4 million tweets per day. Moreover, the sources of more than half of these tweets are from Saudi Arabia and Egypt, accounting for 30% and 20% of all tweets, respectively. The most active users are from Saudi Arabia followed by Egypt, Algeria and the United Arab Emirates [26]. The Arabic language is a Semitic language and is the native tongue in 22 Arab countries. In addition, it is one of the six official languages of the United Nations [15, 18]. Arabic consists of 29 letters that can be used to form words; other languages, such as Farsi and Urdu, also use Arabic characters [15]. Although Arabic is a widely spoken language, it is a language with a very complex morphology owing to the fact that it is highly inflectional. The Arabic language is classified into two types: Standard Arabic, which includes Classical Arabic (CA) and Modern Standard Arabic (MSA) and represents formal language, and dialectal Arabic that represents informal language [1, 2]. Colloquial or dialectal Arabic is the language used in social media, and reflects the spoken Arabic of daily life. It differs from MSA in many aspects [1, 8]; for example, it is the main informal writing style used for posting messages on social media [25]. According to Abozinadah and Jones [1], the main dialectal languages in the Arab region are: Egyptian, Moroccan, Levantine, Iraqi, Gulf, and Yemeni. However, the Arabian Gulf region is comprised of five states: Saudi Arabia, Kuwait, Bahrain, Qatar and Oman. Although the last four states and the eastern region of Saudi Arabia share similarities in terms of culture and dialectal or colloquial language, the other parts of Saudi Arabia are different. As is well known, text mining research relies heavily on the availability of a suitable corpus. This paper is concerned with building a Saudi Twitter corpus available online for use by other researchers, which is the main contribution of this work. Furthermore, the Saudi Twitter corpus is compared with the Egyptian Twitter corpus [24] and the Arabic top news raw corpus [6], i.e., compared with other dialectal language and formal language. Moreover, analysis of the current built corpus was performed to investigate the language used in this type of social media dataset.

The remainder of the paper is organized as follows. In section 2, a background to the topic is presented. The collected Saudi Twitter corpus is described in section 3. The language issues raised in the collected data are investigated in section 4. Section 5 provides a

discussion on the inferred results. Finally, the conclusion is presented in section 6.

2. Related Works

According to Assiri *et al*. [9], Arabic corpora are relatively few in number and can be difficult to access. The developed Arabic corpora were often collected from newswires, and represent MSA. However, in the past few years, researchers have started to pay attention to the data generated by social media. Sentiment analysis on Twitter data has become an active research field in many languages as well as real-time event extraction. In this section, research performed on Saudi Twitter datasets in the literature are discussed.

Mubarak and Darwish [23] developed a multi-dialectal corpus of Arabic from Twitter. They used Twitter users' geographical information that appeared in their profiles to collect the tweets of interest. Specific normalization process that contains mapping frequent non-Arabic characters and decoration to their mappings and processing elongated and shortened words as well as basic Arabic normalization were applied to clean the corpus. The aim of the research was to identify the country of the users based on their tweets.

Refaee and Rieser [25] built a manually labeled corpus with 8,868 tweets for subjectivity and sentiment analysis. The collected dataset was cleaned by removing the usernames, digits, links and emails. In addition, sets of features were used for the annotations: morphological, syntactic, semantic, stylistic, and social signals.

Assiri *et al*. [9] developed an annotated Saudi dialectal dataset of 4,700 tweets for sentiment analysis purposes. This dataset was cleaned by removing non-Arabic elements, such as hashtags, links and non-Arabic words. Additionally, duplicated tweets were removed from the dataset. With regard to the annotation part, each tweet was manually labeled as positive, neutral or negative, based on specific instructions. The annotations reliability of this dataset achieved Kappa=0.807.

Baly *et al*. [10] used Twitter to create the annotated Multi-Dialect Arabic Sentiment Twitter Dataset (MD-ArSenTD) compiled from various Arab states including the Arabian Gulf, the Levant, Egypt, and North Africa, and Saudi Arabia. The total number of tweets was 14,400. The normalization process was performed, i.e. repeated tweets were removed as well as tweets that contained less than 30 characters. Crowd Flower was used for sentiment and dialect annotations. In their experiments, repeated characters were removed, emoticons were substituted by happy or sad tokens, and different types of parentheses were replaced by square brackets. Implementation was only performed on tweets from Egypt and the UAE. In addition, a comparative study between the two tweets was performed in terms of the type of language used, sentiment and tweet topic. The results of the study showed that the language of the Egyptian tweets was dialectal, whereas the language of the UAE tweets was MSA. Regarding the sentiment, the Egyptian tweets were neutral, whereas the UAE tweets were positive. Lastly, the topics of the Egyptian tweets were mainly related to personal matters whereas the UAE tweets were more concerned with religious matters.

Al-Twairesh *et al*. [3] created a corpus for Arabic sentiment analysis of Saudi tweets. This corpus was manually annotated for sentiment: positive, negative, neutral and mixed. Sentiment keywords, such as سعادة happiness, and مؤسف regrettable, were used for building the corpus. The corpus was cleaned and preprocessed by removing repeated tweets, URLs and users' accounts. Furthermore, for reducing the data sparseness problem in the data, normalization was performed on some Arabic letters.

The presented Saudi tweets datasets were mainly developed for sentiment analysis research. To achieve this, preprocessing was carried out on the datasets. However, some issues in social data that have not been discussed or investigated were found, including shortening, compounded words, misspelled words, abbreviations, dialectal words (slang), neologisms, concatenation, word elongation and idiomatic expressions. Therefore, this study aims to investigate these issues and phenomena in our collected corpus as well as in the other two corpora (Egyptian dataset and Arabic top news raw corpus). Table 1 presents different developed Arabic and Saudi Twitter datasets discussed in this section.

Table 1. Developed Arabic and Saudi twitter datasets.

Corpus	Collected tweets	Final dataset	Year	Main task	Sentiment keywords	Preprocessing
Arabic Twitter [25]	8,868	8,868	2014	Subjectivity and sentiment analysis	No	Yes
Multi-Dialectal Corpus of Arabic [23]	175 M	6.5 M	2014	Classification	No	Yes
Saudi Twitter Corpus [9]	4,700	4,700	2016	Sentiment analysis	No	Yes
Multi-Dialect Arabic Sentiment Twitter Dataset (MD-ArSenTD) [10]	470K	14,400 *Saudi tweets (1200)	2017	Sentiment and dialect	No	Yes
AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets [3]	6.3 M	2.2 M *Saudi tweets (17,573)	2017	Sentiment analysis	yes	Yes

3. Saudi Twitter Corpus

3.1. Corpus Description

Twitter API was used to compile Twitter users' tweets for building the current corpus. All the tweets were

collected in 2017, generated by 101 Saudi users. The corpus contains 101 UTF-8 text files and the total number of the tweets is 207,452. In this work, to avoid text repetition, a user's tweets and replies were collected and any retweet was ignored. This confirms that the content of every single file in the corpus was generated by a single author. The corpus contains raw data in order for researchers to perform their research based on their goals. The names of the 101 files were assigned numbers from 1 to 101 in order to hide users' accounts. Every single tweet contains the tweet's date. Also, a table containing the number of tweets for each file was attached with the corpus. The Corpus can be downloaded by the following link:https://figshare.com/s/c38c845ce72f2c6fb599.

3.2. Frequency Analysis

Table 2 shows a comparison between the current Twitter corpus with a general corpus [6] containing more than 2 million words and the Egyptian colloquial Twitter corpus [24].

Table 2. Frequency distribution of the first 50 most frequent tokens in the Saudi Twitter Corpus, Egyptian corpus and MSA corpus.

No	Saudi Twitter Corpus	Freq	Egyptian Twitter Corpus	Freq	MSA corpus	Freq
1	https	39311	على	1700	في	119549
2	الله	16285	مصر	1108	من	73659
3	على	14716	الله	570	ان	50048
4	اللي	6827	اليوم	386	على	50022
5	هذا	6152	مرسي	345	الى	28289
6	ولا	5151	اللي	338	وقال	25780
7	انا	4299	بعد	338	يوم	21633
8	والله	4250	كلام	308	التي	20752
9	اللهم	3969	ولا	279	عن	16428
10	اذا	3583	كان	268	اسرائيل	14958
11	بعد	3482	عمرو	264	المتحدة	13509
12	فيه	3260	آخر	261	مع	12557
13	كان	3177	شاهد	256	أن	12516
14	الا	3107	محمد	255	بعد	12339
15	حتى	2876	إلى	253	الله	12216
16	لكن	2686	علي	251	لبنان	10643
17	اليوم	2550	السادة	248	رويترز	10492
18	انت	2390	الشعب	248	انه	9538
19	http	2235	عبد	248	حزب	9257
20	قبل	2112	يوسف	244	حماس	9248
21	عليه	2023	المحترمون	239	العراق	8362
22	يعني	2014	الثورة	231	بغداد	8205
23	عشان	1957	الاخوان	230	قال	8181
24	يوم	1940	الإخوان	220	الذي	8085
25	انه	1889	بين	217	لا	7782
26	غير	1842	أبو	211	بين	7456
27	شيء	1830	الفتوح	211	ما	7380
28	خير	1816	هذا	211	لم	6443
29	الناس	1770	الى	209	هذا	6145
30	هذه	1739	يوم	205	الجيش	6128

The intersection between the three corpora shows that the Saudi Twitter corpus is similar to the Egyptian and MSA datasets with 38% and 28%, respectively. The results indicate that there is a difference between the language used in Twitter and the formal language used in newswires. However, it was noticed that the Saudi Twitter corpus contains some tokens that do not appear in the top 50 tokens of the other two corpora:

- Occurrence of 'https' and 'http' 39311 and 2235, respectively.
- Links, such as Sq3MIY1WfP (https://t.co/Sq3MIY1WfP).
- Colloquial words, such as عشان, الي.

The language used by Twitter authors often contains informal and colloquial words. The following are more investigations on phenomena identified from the Saudi Twitter corpus.

4. Corpus Issues

4.1. Shortening

The size of a tweet was previously limited to 140 characters, although it has recently increased to 280. This led Twitter users in the Arabic language to abbreviate words in order to maximize writing their tweets. This phenomenon does not exist in the MSA corpus and the trimmed words style is not used in the formal data. In the Saudi Twitter corpus it was found that users often use abbreviations with Arabic prepositional words. This analysis confirms the findings mentioned in [5]. The following Table 3 lists the most significant words that were identified.

Table 3. Some trimmed words identified from Saudi Twitter corpus.

Short form	Corrected from	Translation	Saudi Twitter Corpus	Egyptian Twitter Corpus	MSA corpus
ع	عن | على	On | At	2450	71	0
ف	في	In	998	88	0
ي	يا	Vocative particle "O"	1342	12	0
م	ما	What	707	19	0
ذا	هذا	This	1106	1	0
ذي	هذه	This	632	0	0

4.2. Compounding

According to Somanova [28], fusing two or more words to form a single lexical word is known as compounding. This phenomenon happens in the English language, such as e-commerce and cheesecake, but does not exist in formal Arabic language. However, today, with the use of social networking applications the Arabic language has been affected. While exploring the current corpus, it was found that many words are now joining together. Table 4 shows some examples.

Table 4. Some compound words in the Saudi Twitter corpus.

Compound words	Corrected form	Translation
أنشهد	أنا أشهد	I certify
هالدعم	هذا الدعم	This support
عالعموم	على العموم	In general
بهالبلد	بهذا البلد	In this country
صوتلي	صوت لي	Vote for me
كلشي	كل شيء	Everything

However, common compounding follows this expression: the "w+هـ". The letter "هـ" is fused with the word as a prefix instead of writing the

demonstrative "هذا". This was found in the Egyptian corpus but the letter "هـ" is used instead of the word ("سوف will"; for example, in the word "يضرب will hit" ,"هنسالك will ask", and "هعمل will do", in the sentences below [24].

- بيهددنا انه هيضرب بكرة ذخيرة حية He threatens us that he will hit by using all live ammunition tomorrow.
- مش هنسالك وانتا واقف جنبى I will not ask you while you stand up next to me.
- ولا هعمل بلوك لحد I will not block anybody.

Moreover, sometimes the letter "ح H" is used instead of the above letter. However, all the above words rarely exist in the Saudi corpus. On the other hand, it was noticed that in Egyptian colloquial language, the letter "ش sh" is often fused or attached as a suffix to words, such as "ما بتستمتعش you did not enjoy", and "ما شاركوش they did not participate", but this is not used in Saudi dialectal language.

4.3. Foreign Language

The Arabic language is similar to other languages in terms of borrowing words from other languages; these are called loanwords [28]. In MSA, journalists or authors are forced to use some Latin and English words, especially when they write about scientific subjects, such as the word "virus". On the other hand, in Twitter language, the use of loanwords has increased. Table 5 lists some loanwords that were used in the collected corpus.

Table 5. Some loanwords in the Saudi Twitter corpus.

Loanwords	Frequency	Meaning
منشن	841	Mention
بلوك	291	Block
بلكني	71	He blocked me
ابلك	82	I block
بلكته	21	I blocked him
هاشتاق	512	Hashtag
هاشتاقات	75	Hashtags
رتويت	311	Retweet
برتوت	51	He retweets
برتوتون	19	They retweet

As can be seen the word "منشن mention" is the most frequent foreign word in the Saudi corpus. On the other hand, in the Egyptian corpus it occurred four times in different forms: "منشن mention" twice, "منشنات mentions" once, and "المنشن the mention" once. Moreover, words, such as "بلوك block" and "رتويت retweet" were also found in the Egyptian corpus but with very low frequency.

4.4. Spelling Errors

The data published by newswires, organizations and governments have very low rates of spelling errors, which is due to their professionalism and editing. However, the opposite is found in the Twitter corpus where many spelling errors occur.

4.5. Concatenation

The Arabic language is similar to English in that they are both classified as segmented languages, i.e., they have some signs that assist in indicating segmentation boundaries; the words in both languages are separated by blank spaces. Concatenation does not exist in formal Arabic language. However, the blank spaces are often exploited by Twitter users to write extra words, which leads to concatenation of words without spaces in between. As a result, many problems can occur, such as the developed Part of Speech taggers being unable to identify the types of concatenated words. Also, these are considered as spelling errors because when concatenated words become one single word, they do not exist in the dictionary. The following Table 6 lists three examples of this phenomenon.

Table 6. Some concatenated words in the Saudi Twitter corpus.

Concatenated words	After adding a space
بلداآمنا	بلدا آمنا
بزيادةالمعروض	بزيادة المعروض
واوامرةونواهيةوليرسم	وأوامره ونواهيه وليرسم

4.6. Word Elongation

It was noticed that repetition of letters in words are frequently used by Twitter users to express their opinions or emotions about a subject. In this corpus, many words were affected by this phenomenon. Table 7 presents some examples.

Table 7. Some words with repetition of letters found in the Saudi Twitter corpus.

Word with repetition letters	Corrected lexical	Word with repetition of letters	Corrected lexical
صصبباااحح	صباح	محظووظ	محظوظ
شششووفيي	شوف	رجججججججال	رجال
تتتزوج	تزوج	حمممممدلله	حمدلله
قشعرررريره	قشعريرة	هندسسسه	هندسه
اامممييييي	أمي	جججميل	جميل
عمرررررررررررري	عمري	مممميبررووووككك	مبروك

4.7. Colloquial Language

Colloquial language is, "used in informal conversation rather than in writing or formal language" [21]. As previously mentioned, the collected data were generated from Saudi Twitter authors. However, it was found that users often use their colloquial languages based on the region where they live in Saudi Arabia. Generally, processing data collected from Twitter and compiling very rich dictionaries of slang to interpret and provide synonyms for the colloquial words or expressions requires considerable effort. Table 8 shows some colloquial words found in the current corpus and their frequency in the other two corpora, i.e., the Egyptian Twitter corpus and the Arabic top news raw corpus.

Table 8. Most frequent colloquial words in the Saudi Twitter corpus.

Colloquial words	Formal Arabic	English	Frequency		
			Saudi Twitter Corpus	Egyptian Twitter Corpus	MSA corpus
وش	لماذا	Why	3729	7	0
ايش	ماذا	What	833	0	0
شلون	كيف	How	811	2	0
مب/ موب	ليس	Not	758	0	0

As can be seen in the Table 8 above, all the colloquial or dialectal Saudi words occurred zero times in the Arabic top news raw corpus. This is because the corpus was collected from newswires that use MSA. On the other hand, the word "*وش*" was found 7 times in the Egyptian Twitter corpus. It was found that this word was used to mean "*face*" in 5 tweets by Egyptian people, which is completely different from the Saudi usage.

4.8. Neologisms

A newly coined word was found in this corpus. For example, the word "داعش ISIS" is a new lexical item that has become in daily use. It is an abbreviation of "الدولة الاسلامية في العراق والشام" the Islamic State in Iraq and Syria". This new word was used 565 times in the corpus. In addition, new meanings for existing words were found in the current corpus, such as the word "جحفلي", which means "the great army", but this word also has a different meaning. Generally, today, it means "lost thing at the last second" because of a football match that occurred in 2016. Figure 1 shows the popularity of this word around the world [16]. As can be seen, it is only used in Saudi Arabia. Figure 2 presents the word cloud analysis for the word "جحفلي" [16]. The analysis approved that the meaning and the context of word "جحفلي" were changed. As can be seen, the word "جحفلي" has association with words related to sport domain, such as "الهلال Alhilal football club", "الجمهور fans", "المعلقين commentators" and " هيئة الرياضة sports authority".

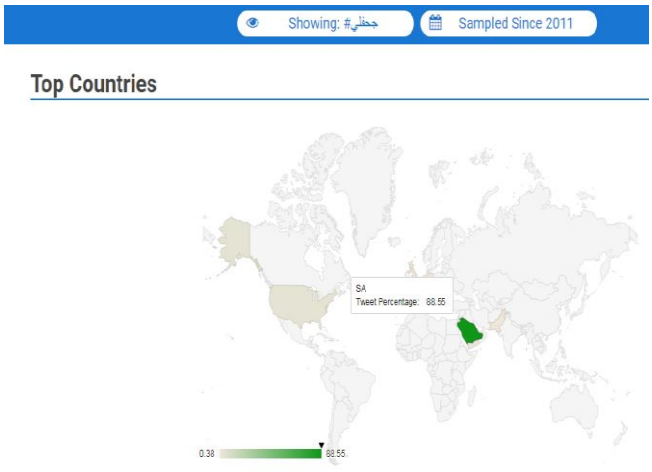

Figure 1. Popularity of the "جحفلي" word around the world (from [16]).

Figure 2. The "جحفلي" word cloud (from [16]).

Moreover, new words have been coined to express the emotion of the user, such as the word "هههههههههههه hahahahahaha", which is a repetition of the letter "هـ h". The length of this word depends on the writer's emotion. The longest repetition of this word was identified in this corpus at 114 letters.

4.9. Respelling

Forming a misspelled new word by changing some letters in a word is called respelling. This phenomenon exists in the English language, such as the word *thanx*, and is mostly used in commercials and slogans [28]. On the other hand, the language of social media contains many respelled words found in the current corpus. Some respelled words identified from the corpus are presented Table 9.

Table 9. Some respelled words.

Respelled words	Corrected Words	Translation
درباوية	درب	Path
جت	جاءت	Come
مزبوط	مضبوط	Exactly
فين	أين	Where

4.10. Twitter Hashtags

Many hashtags were found in the Saudi Twitter corpus. However, Twitter performs normalization on words that constitute hashtags and contain the letter "أ" (with hamza above) "إ" (with hamza below), "آ" (with maad above). These are normalized to "ا". Moreover, remove "ء", if it is at the end of the word, as in this hashtag "اخطا_املايية_شايعة". The word "اخطا" should be written in its correct form "أخطاء". Also, the word that contains "ئـ" changes to letter "ي", as in this "املايية". The word "املايية" should be written "إملائية". However, this normalization leads to spelling errors.

4.11. Idiomatic Expressions

As is well known, idioms or proverbs are phrases that are constituted from collections of words to indicate a meaning. Treating the words of idioms individually can lead to a different sentiment, and hence, the idioms cannot be understood [7, 17]. Ibrahim *et al.*

[17] manually built the AIPSeLEX idioms/proverbs sentiment lexicon for MSAand Egyptian colloquial language to be utilized by sentiment analysis systems for detecting and classifying lexical phrases. In addition, based on n-gram and similarity measurement methods, a classifier to extract idioms and proverbs was developed. However, idioms or proverbs in Saudi colloquial language are different and, to the author's knowledge, have not been investigated to date except in [7], who studied idioms in the Saudi press from the linguistic aspect. In the current corpus, some idioms were identified but did not appear in the compared Egyptian corpus, as shown in Table 10.

Table 10. Some Saudi dialectal idioms found in the Saudi Twitter corpus.

Respelled words	Corrected Words	Translation
درباوية	درب	Path
جت	جاءت	Come
مزبوط	مضبوط	Exactly
فين	أين	Where

5. Discussion

From the above analysis, it is clear that many issues exist in the collected Saudi Twitter data, which can be described as noisy and poor in terms of consistency.Moreover, the comparison between the three corpora highlights that each corpus has its own characteristics, and the corpus that represents MSA differs from the data collected from Twitter. The issues found in the Twitter data (Saudi Twitter data and Egyptian Twitter data), such as shortening, concatenation, colloquial words, compound words, neologisms, word elongation and foreign words do not exist in the written data using MSA. Furthermore, as seen in Table 1, all the developed corpora were cleaned and preprocessed but not all issues were discussed and addressed.

In [4, 5], the Twitter data posted from the Arabian Gulf countries and Egypt were collected for Part of Speech (POS) tagging using the popular Arabic taggers: AMIRA, Morphological Analysis and Disambiguation of Arabic (MADA) and Stanford. However, the performance results yielded were low and unsatisfactory because the previous taggers were developed for MSA and not for colloquial language. Therefore, suggestions were proposed for tackling some of the issues previously explained to increase the POS accuracy. Although the performance of the aforementioned POS systems improved after utilizing the suggested solutions, some issues, such as neologisms, foreign words and compounding have still not been investigated and covered. Moreover, for handling the dialectal (slang) words, some words were mapped to their formal equivalents; for example, the word "وش" (which occurred in our data 3,729 times) was mapped to its equivalent formal word "لماذا why", but this word is used differently by Egyptians, where it means "face" in their dialectal language. Hence, mapping or converting colloquial words directly to their formal equivalent without analyzing the whole context in which the word occurs may lead to an incorrect equivalent. Furthermore, despite their dataset collected from the Arabian Peninsula and Egypt, the developed slang list is almost all from Saudi dialectal language.

On the other hand, Shoukry and Rafea [27] studied how preprocessing is able to improve the accuracy of sentiment analysis in Egyptian dialectal tweets. They proposed three stages for achieving preprocessing: normalization, stemming and stop words removal. For performing normalization, off-the-shelf software was used to handle the spelling variants and diacritics in Arabic. For stemming, 11 rules were developed for performing light stemming on broken plurals in Egyptian dialect. Finally, for stop words removal, a list of stop words for the Egyptian dialect was built. However, many issued identified in this current research regarding the Egyptian Twitter corpus have not been studied.

Although the Saudi dataset and the Egyptian dataset were both collected from Twitter, they are different, as explained above. As a result, handling these issues depends on the type of data and where they are collected from. As previously mentioned, the Arabic region has different dialectal languages, and hence, developing a system to preprocess data collected from Twitter generated from specific users may not able to process the data generated from others.

6. Conclusions

Social media have become significantly important platforms for many natural language processing researches. Therefore, the main aims of this research are to develop a Saudi Twitter corpus and to investigate issues related to this type of data. The number of collected tweets is 207,452. However, unlike MSA, data collected from social media are noisy and informal. Related research shows that some cleaning and preprocessing were performed on their collected data. Nevertheless, this research has identified a number of issues, such as newly coined words (neologisms), borrowing words from other languages, compounding, and respelling that have not been investigated by previous researches in the literature. Moreover, a comparative study between the corpora shows that Saudi data differs from Egyptian data, especially in respect of colloquial words. Consequently, future work will focus on addressing the aforementioned issues pertaining to social media data.

References

[1] Abozinadah A. and Jones J., "Improved Micro-Blog Classification for Detecting Abusive Arabic Twitter Accounts," *International Journal of Data Mining and Knowledge Management Process*, vol. 6, no. 6, pp. 17-28, 2016.

[2] Al-Kabi M., Alsmadi I., Khasawneh R., and Wahsheh H., "Evaluating Social Context In Arabic Opinion Mining," *The International Arab Journal of Information Technology*, vol. 15, no. 6, pp. 974-982, 2017.

[3] Al-Twairesh N., Al-Khalifa H., Al-Salman A., and Al-Ohali Y., "Arasenti-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets," *Procedia Computer Science*, vol. 117, pp. 63-72, 2017.

[4] Albogamy F. and Ramsay A., "Pos Tagging for Arabic Tweets," *in Proceedings of the International Conference Recent Advances in Natural Language Processing*, Hissar, pp. 1-8, 2015.

[5] Albogamy F. and Ramsay A., "Fast and Robust Pos Tagger for Arabic Tweets Using Agreement-Based Bootstrapping," *in Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, pp. 1500-1506, 2016.

[6] Almas Y. and Ahmad K., "Lolo: a System Based On Terminology for Multilingual Extraction," *in Proceedings of the Workshop on Information Extraction Beyond the Document*, Sydney, pp. 56-65, 2006.

[7] Alqahtni H., "The Structure and Context of Idiomatic Expressions in the Saudi Press," Phd Thesis, the University of Leeds, 2014.

[8] Alwakid G., Osman T., and Hughes-Roberts T., "Challenges in Sentiment Analysis for Arabic Social Networks," *Procedia Computer Science*, vol. 117, pp. 89-100, 2017.

[9] Assiri A., Emam A., and Al-Dossari H., "Saudi Twitter Corpus for Sentiment Analysis," *International Journal of Computer and Information Engineering*, vol. 10, no. 2, pp. 272-275, 2016.

[10] Baly R., El-Khoury G., Moukalled R., Aoun R., Hajj H., Shaban K., and El-Hajj W., "Comparative Evaluation of Sentiment Analysis Methods Across Arabic Dialects," *Procedia Computer Science*, vol. 117, pp. 266-273, 2017.

[11] Bani-Hani A., Majdalawieh M., and Obeidat F., "The Creation of an Arabic Emotion Ontology Based on E-Motive," *Procedia Computer Science*, vol. 109, pp. 1053-1059, 2017.

[12] Cherif W., Madani A., and Kissi M., "Towards an Efficient Opinion Measurement in Arabic Comments," *Procedia Computer Science*, vol. 73, pp. 122-129, 2015.

[13] D'Andrea E., Ducange P., Lazzerini B., and Marcelloni F., "Real-Time Detection of Traffic from Twitter Stream Analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2269-2283, 2015.

[14] Elhadad N., Gravano L., Hsu D., Balter S., Reddy V., and Waechter H., "Information Extraction from Social Media for Public Health," *in Proceedings of KDD at Bloomberg Workshop, Data Frameworks Track*, New York, pp. 1-14, 2014.

[15] Elmadany A., Abdou S., and Gheith M., "Towards Understanding Egyptian Arabic Dialogues," *International Journal of Computer Applications*, vol. 120, no. 22, pp. 7-12, 2015.

[16] Hashtagify, "Find and Analyse Top Twitter and Instagram Hashtags [online]," http://hashtagify.me/, Last Visited, 2017.

[17] Ibrahim H., Abdou S., and Gheith M., "Idioms-Proverbs Lexicon for Modern Standard Arabic and Colloquial Sentiment Analysis," *International Journal of Computer Applications*, vol. 118, no. 11, pp. 26-31, 2015.

[18] Ibrahim H., Abdou S., and Gheith M., "Sentiment Analysis for Modern Standard Arabic and Colloquial," *International Journal on Natural Language Computing*, vol. 4, no. 2, pp. 95-109, 2015.

[19] Kulkarni R., Dhanawade S., Raut S., and Lavhakare D., "Twitter Stream Analysis for Traffic Detection in Real Time," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 2, no 5, pp. 1-5, 2016.

[20] Lamb A., Paul M., and Dredze M., "Separating Fact from Fear: Tracking Flu Infections on Twitter," *in Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, pp. 789-795, 2013.

[21] Macmillan Dictionary, 2002, "Colloquial-definition and synonyms [online]," Available from https://www.macmillandictionary.com/dictionary/british/colloquial, Last Visited, 2017.

[22] Mallek F., Belainine B., and Sadat F., "Arabic Social Media Analysis and Translation," *Procedia Computer Science*, vol. 117, pp 298-303, 2017.

[23] Mubarak H. and Darwish K., "Using Twitter To Collect A Multi-Dialectal Corpus of Arabic," *in Proceedings of the EMNLP Workshop on Arabic Natural Language Processing*, Doha, pp. 1-7, 2014.

[24] 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.

[25] Refaee E. and Rieser V., "An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis," *in Proceedings of 9th International Conference on Language Resources and Evaluation*, Reykjavik, pp. 2268-2273, 2014.

[26] Salim F., "Social Media and the Internet of Things towards Data-Driven Policymaking In The Arab World: Potential, Limits And Concerns," Technical Report, Arab Social MediaReport, 2017.

[27] Shoukry A. and Rafea A., "Preprocessing Egyptian Dialect Tweets For Sentiment Mining," *in Proceedings of The 4th Workshop on Computational Approaches to Arabic Script-based Languages*, California, pp. 47-56, 2012.

[28] Somanova L., "Words Recently Coined and Blended: Analysis of New English Lexical Items," Phd Thesis, Masaryk University, 2017.

[29] Tartir S. and Abdul-Nabi I., "Semantic Sentiment Analysis in Arabic Social Media," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 2, pp. 229-233, 2017.

Meshrif Alruily is an Assistant professor, department of Computer and Information Sciences at Jouf University, Saudi Arabia. He received his PhD in Computer Science from the University of De Montfort UK, in 2012. He published many conference papers and journal articles. He has published papers in the European Conference on Artificial Intelligence (ECAI) and Information processing & Management journal. His research interests are related to Arabic text mining field, such as information extraction, summarization, text classification and clustering and data analysis.