A Model for Processing Arabic Text on Twitter

Mohamed Osman Hegazi[1], Yasser Al-Dossari[1], Abdullah Al-Yahya[1], Abdulaziz Al-Sumari1, and Anwer Hilal[2]

m.hjegazi@psau.edu.sa , yasserd99@gmail.com , abdulllahsaad00@gmail.com, az.alsumari@gmail.com , a.hilal@psau.edu.sa

[1]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia

[2]Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia

Summary

This paper proposes a model that can be used as a framework for preprocessing Arabic text on Twitter for data analysis and information extraction. The model provides an online collection of Arabic text on Twitter and stores it in a structured database. The source data are then preprocessed to derive clean, meaningful Arabic text from which information can be extracted. The paper presents new methods and algorithms for preprocessing unstructured Arabic text on social media, and it provides solutions that address the difficulties of working with Arabic text on social media, including uncleaned, informal, and dialect language. The preprocessed Arabic text is stored in structured database tables to provide a useful data set to which information selection and data analysis algorithms can be applied. The implementation of the model yields a useful and full-featured dataset, and the text is presented as the source data, the cleaned data and separate Arabic words with their stems, roots and morphologies, among other forms. In addition, the model shows how information can be selected and extracted from this dataset.

Key words:

Information retrieval; Natural Language Processing; Database; Data Analysis; Text Mining; Arabic Text.

1. Introduction

Social media have increasingly grown beyond personal use. Use of social media has become a necessary daily activity millions of people use social media for different purposes, large amounts of data flow through online networks every minute, and these data contain valuable information that can be extracted if the data are properly analyzed. Twitter is a powerful social media network. In the Arabic world, Twitter has been used for marketing, advertising, business purposes, entertainment, knowledge sharing, news, sports and political communication. On Twitter, there are 310 million users who are active monthly, and 1.3 billion Twitter accounts have been created in total; 83% of world leaders are on Twitter, and 500 M tweets are sent each day, corresponding to an average of 6000 tweets every second [1,2].

On Twitter, a tweet can contain only 280 characters. Thus, users must be very succinct and brief. Twitter also supports the distribution of videos and pictures, in addition to many other features, such as polls and online video streaming. Each user on Twitter follows other users (and can thus see their tweets) and has followers (who can see the original user's tweets), and a reply to a tweet is called a mention. Twitter is clearly very open to the public, and there is no need to have an account to access its information. For developers, Twitter provides a well-designed application program interface (API). Twitter also has a platform for developers that includes communities, customized sets of tools and well-written documentation [3].

While there are many different tools, libraries and applications for making use of English text on social media, the same is not true of Arabic text on social media. Developing and applying computerized tools for extracting knowledge and information from unstructured Arabic text on social media platforms is not an easy task. Many methods must be developed, including NLP (Natural Language Processing), normalization, learning algorithms, and application design tools, for application to Arabic text on social media.

Arabic language is an important language; is the native language for More than 422 million speakers and the language of the holy book for all Muslims (more than 1.62 billion people) [4, 11]. Arabic consists of 28 letters, including three vowel letters, Arabic letter shape changes according to its position in the word. Arabic is written from right to left.

Analyzing and processing Arabic language is one of the challenges in natural language processing and information retrieval. This is due to some of its special characteristics such as abundance of vocabulary, structural templates such (فعل، فاعل ، مفعول), diacritic, and multiple derivations in additional to the special characteristics of Arabic text in social media.

Recently, research works in social media Arabic language processing gained increasing importance, and several research works and applications have been developed, as examples Darwish, Magdy, and Mourad [4] worked in retrieving Arabic text from microblogs on Twitter. They addressed the challenges associated with cleaning microblogs Arabic text. They normalized the Arabic letters, handle the elongated and shortened words, removing stop

Manuscript received May 5, 2020
Manuscript revised May 20, 2020

words, stemming, and considering Arabic dialects. Al-Twairesh, Nora, et al. [5] worked in sentiment analysis using in twitter Arabic text. They applied existing methods for cleaning and processing their data. Eshrag Refaee and Verena Rieser [6] worked in corpus development. They developed Arabic twitter corpus for Subjectivity and Sentiment Analysis. They used the Twitter API to collect tweets in real time, and they cleaned the extracted data by eliminating the Latin characters. Alshutayri and Eric [7] devolved corpora for dialect. They created a corpus of dialect Arabic text from Twitter using common dialect words (seeds). They used Twitter API to connect to Twitter, and they then collected the tweets using the seed words. Mubarak and Darwish [8] devloped a multi-dialect Arabic corpus from Twitter. They used the Twitter API to obtain Arabic tweets and then extracted user locations to classify the text as being in specific dialects according to these locations. They also processed the Arabic characters to convert them to normal forms.

Although interest in preprocessing Arabic text on social media has recently increased, standard or integrated work on preprocessing Arabic text on social media is still lacking. In particular, models, tools and frameworks that can produce full-featured datasets are needed.

This paper proposes a model for preprocessing and analyzing Arabic text on Twitter and producing full-featured datasets

2. The Proposed Model

The structure of the proposed model for preprocessing Arabic text on Twitter has four stages (Fig. 1):

1- In the first stage, the model establishes an online connection to Twitter's API and server.
2- In the second stage, Arabic text is selected directly from the Twitter server and stored in a local database file. This file contains our source data.
3- In the third stage, the source data are preprocessed to result in the final data set.
4- In the last stage, the information needed can be obtained by applying the appropriate data analysis, information extraction or/and mining algorithms.

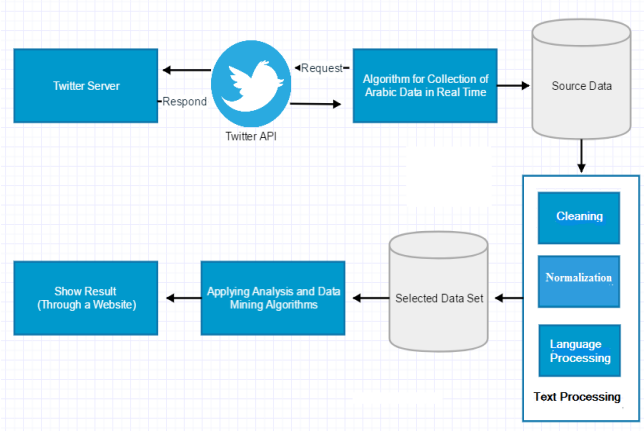

Fig. 1 A model for processing Arabic text on Twitter

2.1 The online connections

The first stage in our model is to connect with Twitter OAuth "An open protocol to allow secure authorization in a simple and standard method from web, mobile and desktop applications" [9]

The connection algorithm sends a query to the Twitter database using Twitter's API through the Tweepy API wrapper. Twitter returns the query as a JSON object. The algorithm then decodes the JSON object and converts it into a string. (Fig. 2).

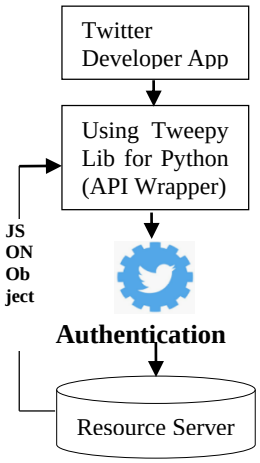

Fig. 2 Connection module

2.2 Collecting source data

In this model, Arabic text on Twitter is collected into a structured dataset (a database table) to provide additional flexibility when analyzing and extracting information from the text. This format can also represent very large amounts of data. The data collection algorithm functions as follows:

1- The data are obtained as a Javascript Object

Notation (JSON) object.
2- The Tweepy interface is used to stream the Twitter data.
3- Check each tweet to determine whether the text is in the Arabic language.
4- If yes, the required information is stored in the database.

These processes can be continued in an infinite loop, as shown in Fig. 3.

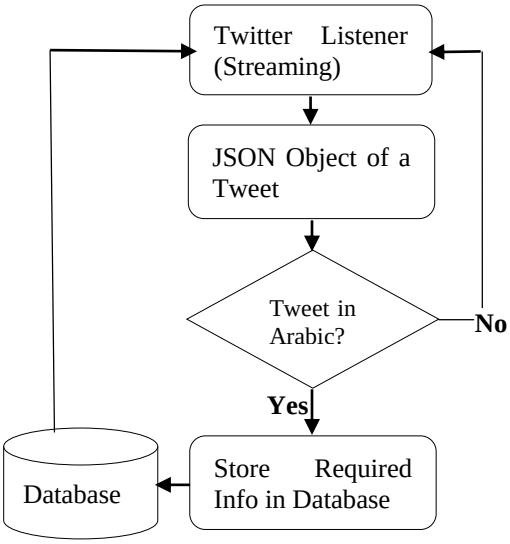

Fig. 3 Data collection module

2.3 Preprocessing Arabic text

Preprocessing formal Arabic text is difficult. Preprocessing Arabic text on social media, which is usually informal, is even more difficult for many reasons, such as the presence of dialect text, common spelling mistakes, extra characters, diacritical marks, and/or elongations. Accordingly, to preprocess this type of Arabic text, we need to carry out additional processing, such as stripping the elongations, diacritical marks, and extra characters. We must also convert the text to its normal Arabic form and then process the language (i.e., carry out natural language processing of Arabic text) to provide more information on the contents of the text of the tweets. The relevant natural language processing techniques include stemming and morphological analysis. Accordingly, this model provides three important preprocessing modules:

1- Text cleaning
2- Text normalization
3- Natural language processing of Arabic text

2.3.1 Text Cleaning

Most Arabic text on social media contains noise, such as elongations, diacritical marks, extra symbols, and/or mixed language. The first step and the most important step in preprocessing Arabic text on Twitter is to clean the tweet text by removing such noise.

Most of the previous algorithms work on the cleaning the noises by trying to expect all possible noises and then, looking inside the text for each noise and clean it, which makes the cleanliness of the texts depends on the degree of anticipation of noises. This way of cleaning cannot provide accurate results because; it is not easy to expect all noises, beside the possibility of arising new noises. Our cleaning algorithm works differently than the previous algorithms, since it does not look at any noises; it only selects the Arabic characters and ignores any non-Arabic characters or symbols.

As it shown in Fig. 4, the cleaning algorithm reads the tweets text character by character and checks whether each character belongs to the Arabic alphabet. If a character belongs to the Arabic alphabet, it is selected; otherwise, the algorithm replaces it with a space.

This method eliminates all noise, such as additional characters, samples, non-Arabic characters and any shapes, and provides cleaned tweets containing Arabic text without affecting the meaning or the content.

The following is a sample that show the appearance of the text after this cleaning algorithm is applied, and the final cleaning result is illustrated in Fig.5

Text before cleaning:

مصر | #الارصاد : طقس اليوم معتدل .. والعظمي في #القاهرة 29 درجة#

Text after cleaning:

مصر الارصاد طقس اليوم معتدل والعظمي في القاهرة درجة

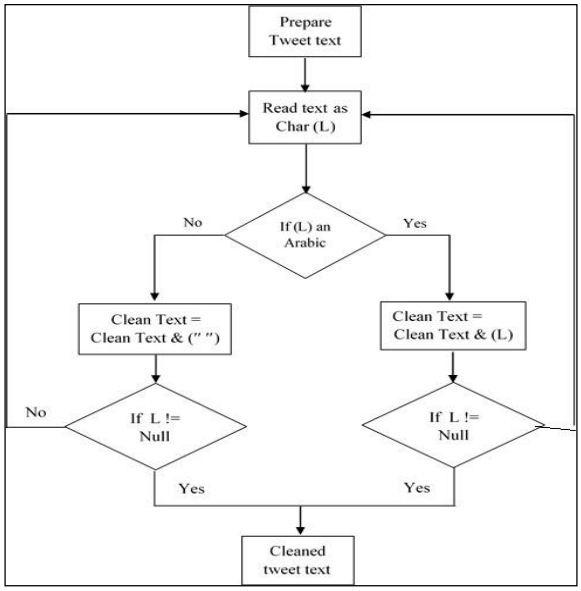

Fig. 4 Cleaning algorithm

عبدالعزيز مروق اتوقع مخلص غسيل ملابسه حياتك	عبدالعزيز مروق اتوقع مخلص غسيل ملابسه
استغفر الله العظيم واتوب اليه	استغفر الله العظيم واتوب اليه
الي مثى كلمات حمد البريدي القاء ص	الي مثى كلمات:حمد البريدي@ad_albraidi
عطاه اشار يمين وراح يسار	عطاه اشارة يمين وراح يسار
ربي بلغنا رمضان مكتملين مجموعين متحابين لا فاقدين	ربي بلغنا رمضان مكتملين مجموعين متحابين
ميسي ينشر غسيل ريال مدريد في البرنابيو	ميسي ينشر غسيل ريال مدريد في البرنابيو.. (:
قروب الشوق للدعم اناق ت	★ ★ ★ ★ قروب_الشوق_للدعم#
لدينا مبلطين هنود باسعار منافسه و جوده عاليه ب	بلطين هنود باسعار منافسه و جوده عاليه ب.
لما تبقى متضايق ابقى قولي ماهو مش ميكانيكي عشا	لما تبقى متضايقة ابقى قولي.. ماهو مش ميكا

Fig. 5 Sample of cleaning algorithm results

2.3.2 Text normalization

This module concerning about the process of putting the text in its normal Arabic language form. Some Arabic words on social media are written in nonstandard ways; for example, some words contain repeated characters, include common spelling mistakes, or are written in dialects of Arabic. The normalization algorithm replaces the un-normal word by normal one using a set of common un-normal words.

In this stage, we consider only words that contain repeated characters and some common nonstandard words (fig. 6). In future work, this set will be expanded and implemented in a database table to act as a dictionary of nonstandard and dedicated Arabic

```
Def Function RepeatedletterNormalization (Tweet) as String
        Tweet = Replace ("آ,إ,أ","ا",Tweet)
        Tweet = Replace ("ى","ي", Tweet)
        Tweet = Replace ("ييي","ي", Tweet)
        Tweet = Replace ("يييـ","ي", Tweet)
        Tweet = Replace ("ؤ","ء", Tweet)
        Tweet = Replace ("ئ","ء", Tweet)
        Tweet = Replace ("و و و و و و","و", Tweet)
        Tweet = Replace ("و و و و و","و", Tweet)
        Tweet = Replace ("و و و و","و", Tweet)
        Tweet = Replace ("و و و","و", Tweet)
End Function
```

Fig. 6 Normalization algorithm

2.3.3 Processing the Arabic language (ANLP)

To provide a useful and full-featured dataset, the model includes a module that carries out natural language processing of Arabic text; this module provides word stemming, root and morphological analysis functionality. The module works as follows:

1- Text tokenization: At first, the cleaned tweet text is fragmented into words and stored in the database table. To do so, we adopt the fragmentation algorithm used in [10], which considers the spaces between words and punctuation, such as stop marks, commas, and semicolons.

2- Stemming: In the second stage, a new light stemmer algorithm is designed to provide the word stems and roots. This algorithm functions as follows (fig. 7):

- The algorithm provides sets of common Arabic names, prepositions, pronouns and special roots (roots that have more than three characters).
- The algorithm checks each word. If it is in one of these sets, then the algorithm retains it unchanged and specifies the lack of change in the appropriate field.
- If the word is not in one of these sets, then the algorithm checks to determine whether the word has three characters, in which case the root and the stem are the same and are equal to the word. If the word has more than three characters, then the algorithm removes the affixes from the word (the vowel letters at prefixes, suffixes and infixes). Otherwise, the algorithm considers this word to be an unrecognized Arabic word or a dialect word.

Morphological analysis: The last module generates word morphology templates, "فاعل", "فعول", "مفاعيل", "مفعول" and "يفعل". To do so, we adopt the morphology generation algorithm used in [11], which inserts or adds letters to root words at certain positions according to a morphology template (Fig. 8)

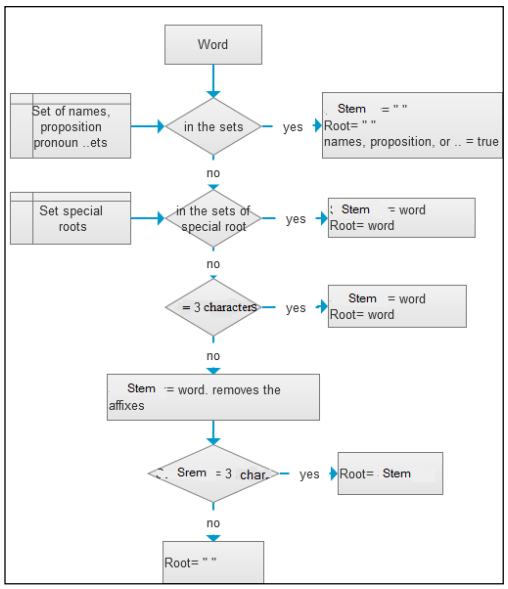

Fig. 7 Stemming and root generating

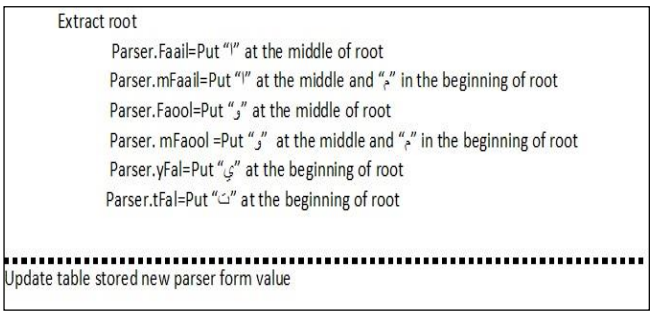

Fig. 8 Morphological algorithm

2.4 The final dataset

The preprocessing of the source data results in the final dataset. The final dataset is designed and structured to occupy a database table with almost all of the features (more than 20 features are included). Fig (9) show the entity relation diagram (ER) of the final dataset.

As it mentioned in the previous sections the final dataset is generated using the following stages:

1- At first data is connected as a stream of JOSN strings, and stored in a database table (see Fig. 3 and 4). This stage is a continuous stage and the collected data are stored temporally in a single database table.
2- At the second stage, data is cleaned and normalized, resulting new database table with a cleaning tweet as a new field. This new database table is used as final storage for twitter data,
3- At the third stage, the new database table is manipulated using Arabic NPL to provide additional features.
4- At the last stage, the final dataset is then presented in many ways and tables according to the requirements of processing algorithms (Fig. 10)

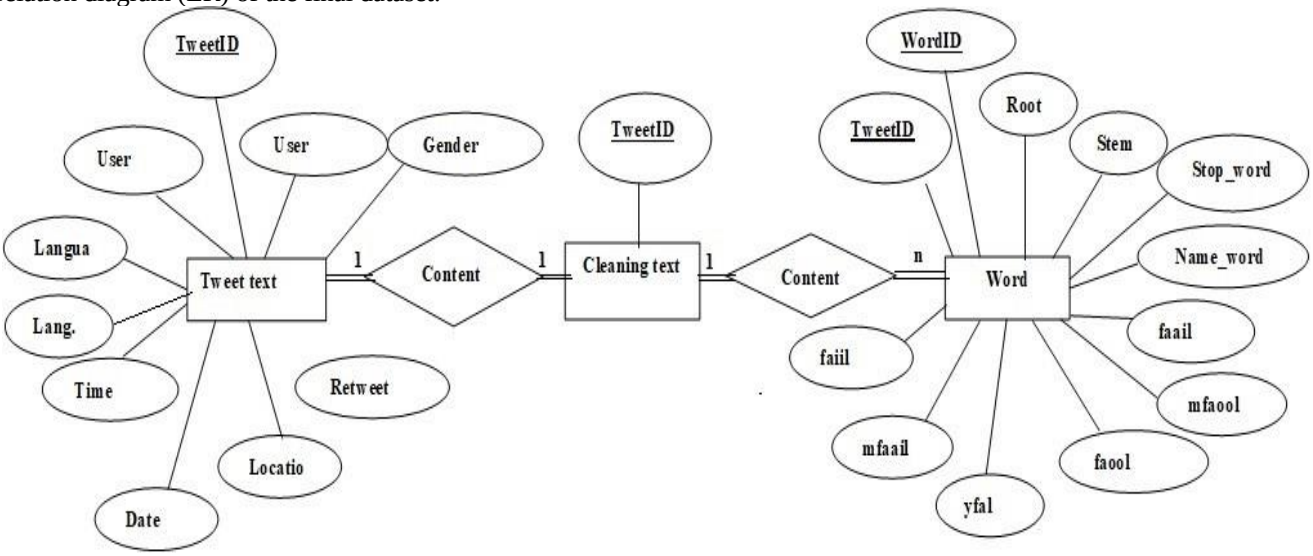

Fig. 9 Entity relation diagram of the model data

Fig. 10 Screen shot collected diagram representing part of the final data

3. Results and Discussions

This model has been developed and tested using VB.net, Python, SQLite, MS-Access, SQL-Server and Excel. The data presented in this paper were collected online in April 2017. The implementation of the model provides a useful and full-featured dataset; Arabic text on Twitter became available in the fields of a database table with other fields, such as the name, location, date, time, etc. The final dataset can be made available in many ways, such as a single database table or broken into three tables (Fig. 10); it can also be made available in any format, such as text, Excel, SQLite, Access SQL Server and/or Oracle files).

The preprocessing of tweet text results in cleaned text with additional useful fields, such as word fragments, word stems, word roots, and word morphologies. Because all of these data are available in databases and other formats, the implementation of information-extracting models or algorithms will be easy. For example, we can implement the following:

- database queries to obtain many types of information, such as search results, selecting and/or providing reports on specific or general information (Fig 11 as example);
- analysis algorithms;
- visualization techniques to provide visual representations of the data, such as plots and graph (Fig 11 and 12 show an example of how the data can be classified and visually presented according to tweets topics) ;
- and/or predictive or descriptive algorithms to extract hidden patterns from these data to obtain more useful results or perform additional studies.

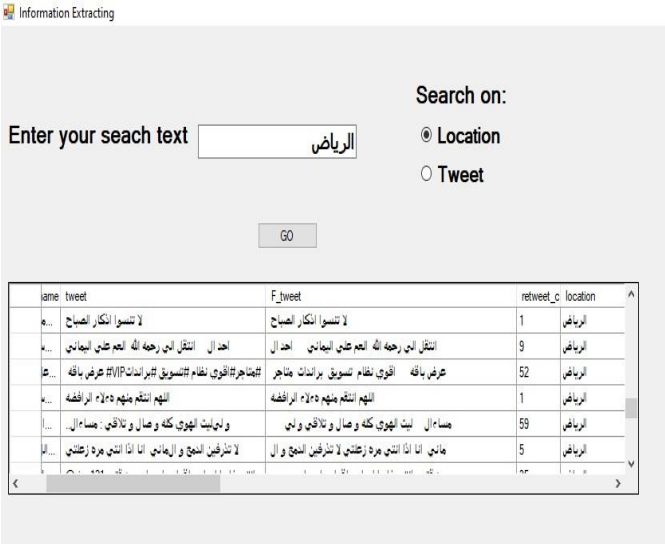

Fig. 11 An example of information retrieval (select the tweets from location = "الرياض")

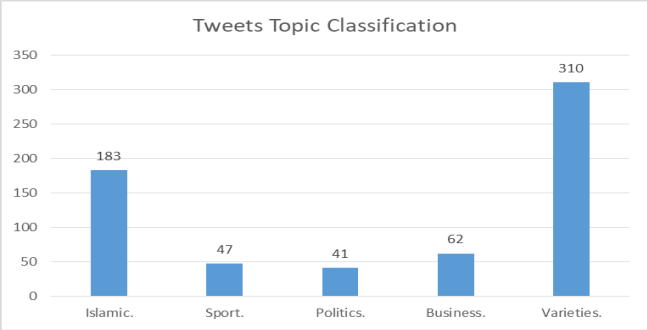

Fig. 12 Virtual representation for tweets topics classification

4. Conclusions

This paper illustrates the first stage of a project that aims to produce an integrated application for processing and analyzing Arabic text on social media. The paper presents the general framework of the project, which is expressed in a model that collects, preprocesses Arabic text on Twitter. The paper shows that the proposed model succeeds in addressing the difficulties associated with preprocessing Arabic text on social media and can provide a considerable Arabic corpus that can be used to extract useful information from Arabic text on Twitter.

The model introduces three new methodologies and algorithms for cleaning, normalization and stemming, in addition to enhancing and adapting the Arabic text fragmenting algorithm and Arabic morphology generating algorithm illustrated in [10] and [11], respectively. These algorithms can act as tool kits and can be reused or embedded in future work. The dataset provided by this model has been tested and implemented in several ways, and it successfully provides useful information.

The model shows additional and integrated aspects of preprocessing Arabic text on social media compared with previous studies. The cleaning process that is presented in this paper has not been included in most previous studies, even in preprocessing the English text on twitter, according to Jianqiang, and Xiaolin [12], six complicated approaches were used to reduce the amount of noise in the tweets. Which are: replacing negative mentions, removing URLs, expanding acronyms, stemming, removing stop words, and removing repeated letters, and according to Han and Baldwin [13] most of the previous work did not provide an accurate solution.

Natural language processors did not find enough spaces in most of the previous studies except [4], in which the authors applied stemming to address some issues in word normalization. In addition, this paper presents a number of methods of enhancing the dataset by providing additional information that is not presented in the previous papers, such as roots and word morphologies.

The model presented in this paper provides a viable solution to most of the obstacles that arise in work that considers Arabic text on social media, such as subject analysis, sentiment analysis, studies of dialect language and/or information retrieval.

Acknowledgments

This project was supported by the deanship of scientific research at Prince Sattam bin Abdulaziz University under research project number 2017/01/7773

References

[1] [Alshutayri and Atwell 2017] Areej Alshutayri & Eric Atwell. "Exploring Twitter as a Source of an Arabic Dialect Corpus." International Journal of Computational Linguistics (IJCL) 8.2 (2017): 37-44. ACM 978-1-4503-1156-4/12/10, 2012.

[2] Twitter Usage Statistics, [Online]. Available: http://www.internetlivestats.com/twitter-statistics/, Feb. 5, 2018

[3] Twitter Usage Statistics, [Online]. Available: https://developer.twitter.com/en/docs, Feb. 5, 2018

[4] Darwish, Kareem, Walid Magdy, and Ahmed Mourad. "Language processing for arabic microblog retrieval." Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012.

[5] Al-Twairesh, Nora, et al. "AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets." Procedia Computer Science 117 (2017): 63-72.

[6] Refaee, Eshrag, and Verena Rieser. "An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis." LREC. 2014.

[7] Alshutayri, A. O. O., and Eric Atwell. "Exploring Twitter as a Source of an Arabic Dialect Corpus." International Journal of Computational Linguistics (IJCL) 8.2 (2017): 37-44. Copyright 2012 ACM 978-1-4503-1156-4/12/10

[8] H. Mubarak, K. Darwish. (2014). "Using Twitter to collect a multi-dialectal corpus of Arabic". In: Proceedings of the EMNLP workshop on natural language processing. Doha, Qatar, 25 October, 2014, pp. 1-7.

[9] Developer Twitter. Authentication [Online]. Available: "https://developer.twitter.com/en/docs/basics/authentication/overview/oauth" Feb. 5, 2018

[10] Hegazi, M.O., Hilal, A. and Alhawarat, M., 2015. Fine-grained Quran Dataset. International Journal of Advanced Computer Science and Applications (IJACSA), 6(12), pp.262-267

[11] Hegazi, M.O., 2016. An Approach for Arabic Root Generating and Lexicon Development. International Journal of Computer Science and Network Security (IJCSNS), 16(1), p.9.

[12] Jianqiang, Z. and Xiaolin, G., 2017. Comparison Research on Text Preprocessing Methods on Twitter Sentiment Analysis. IEEE Access, 5, pp.2870-2879.

[13] Han, B. and Baldwin, T., 2011, June. Lexical normalisation of short text messages: Makn sens a# twitter. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-

Volume 1 (pp. 368-378). Association for Computational Linguistics.

Mohamed Osman Hegazi is a Professor of Computer Science in the Department of Computer Science at Prince Sattam bin Abdulaziz University. He obtained his Ph. D. degree in 2004 from the Sudan University of Science & Technology, Khartoum Sudan, for his thesis entitled An Approach for Heterogeneous Distributed Database Systems Integration. His research interests include databases, data mining, data science, NLP and computer applications.

Yasser Al-Dossari received an honors B. Sc. Degree with distinction in computer science from Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia. Currently, he is a Computer Security Researcher at the Saudi Arabian Monetary Authority.

Abdullah Al-Yahya received a B.Sc. degree in computer science from Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia.

Abdulaziz Al-Sumari received a B.Sc. degree in computer science from Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia.

Anwer Mustafa Hilal is an Assistant Professor of Computer Science in the Department of Computer and Self Development at Prince Sattam bin Abdulaziz University. He obtained his PhD degree in 2017 from Omdurman Islamic University, Khartoum, Sudan, for his thesis entitled A Semantic Data Mining Model for Exploring the Holy Quran. His research interests include data mining, text mining, and mobile and Web development.