# Building a Twitter Social Media Network Corpus for Libyan Dialect

Husien Albashir Alhammi[1], Ramadan Alsayed Alfard[2]
[1] Department of Electrical and Electronic, Polytechnic Institute of Zawia, Zawia, Libya.
[2] Department of Computer, Faculty of Science, University of Zawia, Zawia, Libya.

* Corresponding author. Tel.: (218)911815564; email: ramadan.alfared@zu.edu.ly
Manuscript submitted July 10, 2017; accepted September 23, 2017.
doi: 10.17706/ijcee.2018.10.1.46-52

**Abstract**: We present a new Libyan dialect corpus which contains data set of 5000 statements that written by Libyan Twitter's users as tweets. The corpus is manually classified into fifteen categories. However, the statistical results of corpus data show that there is the significant variety of tweets numbers among categories. Obviously, Libyan Twitter's users tend to express most their tweets in sports whereas the lowest number of tweets was in films category. Yet no corpus of Libyan dialect has emerged. Therefore, building such a corpus is very crucial to be used for the public purposes of not only linguistic but also NLP research for Libyan dialect.

**Keywords**: Libyan dialect, libyan dialect NLP, libyan twitter corpus, twitter corpus.

## 1. Introduction

Libyan dialect is one of many Arabic dialects which are spoken in north Africa. Arabic scripts come in many forms: Classical Arabic (CA), Modern Standard Arabic (MSA) and Dialectal Arabic (DA) have become increasingly used in social media networks such as facebook, twitter, etc. Generally, individuals posting text on social media networks tend to use informal writing style, the majority of Arabic tweets in social media networks are mainly expressed in dialect-written or having a combination of Arabic dialects and standard Arabic.

Nowadays, Twitter social media network has hindered of million active users over the world, about 500 million tweets are posted per day[1]. Consequently, twitter is considered a rich and precious resource of data for researching. Twitter users offer free great benefits from their personal tweets on web, these tweets could help researchers to carry out their research in NLP tasks [1]-[3]. Therefore, our overall aim is to build a new corpus to provide an ideal data source for supporting both linguists and researches in the field of NLP for Libyan dialect.

The modern Libyan dialect has been shaped by some of historical events that occurred in north Africa mainly the Hilalian-Sulaimi migration, and the migration of Arabs from Muslim Spain to north Africa as well as it has also used numerous words originally from colonialist languages such as Italian and Turkish in colonization era. In addition, some indigenous languages such as Berber or Amazigh languages have widely used in the Libyan dialect [4].

The Libyan dialect is spoken by an estimated six million speakers throughout the country and slightly extends beyond the borders to: southeast Tunisia, southeast Algeria, northwest Egypt, northwest Sudan, north Chad and northeast Niger. The Libyan dialect can be classified into three very similar dialects

[1]www.internetlivestats.com/twitter-statistics

depending on their geographic sites: the eastern dialect, the western dialect and the southern dialect being centered in Benghazi, Tripoli and Sabha respectively. Fig. 1 illustrates The Libyan dialects and their geographic areas.

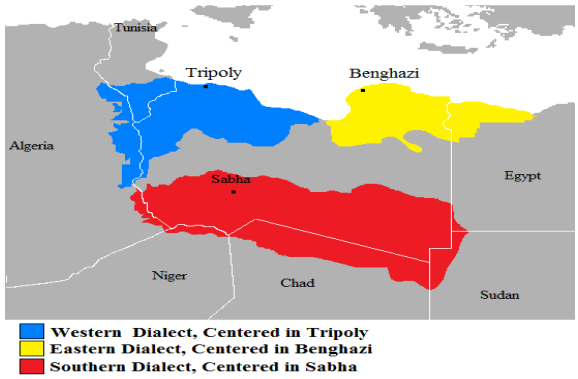


Fig. 1. Libyan dialects and their geographical distribution.

As shown in Fig. 1 Libyan dialect is also spoken in neighboring countries. For example, the Libyan word باهي 'bahi' is one of the most typical Libyan words which means in English “ok”, “agree”, or ‘nice’ is also spoken by Tunisians in the southeast of Tunisia but less often. In this paper, we describe the main steps for constructing a new corpus of Libyan dialect with manually annotation on level of category or domain which has been collected from Twitter social media network over the span of two months.

The rest of the paper is organized as follows. In the following section, we outline the essential steps of collecting data corpus. Related work in Section 3. In Section 4, we describe the process of data pre-processing and corpus format. Section 5 presents the results of statistical analysis of corpus data. Finally, Section 6 concludes the article and future work.

## 2. Related Work

In 2010, Petrovi´c et al, presented the large-scale Edinburgh Twitter Corpus [5], a collection of 97 million tweets for researchers dealing with social media, natural language processing, or big data processing. With regard to tweets collecting there have been several papers, for example, Tjong Kim Sang [6] presented instructions for linguists about how to collect tweets as well as some ideas about how information of tweets can be visualized.

## 3. Corpus Collection

For corpus collection, we use the Twitter Search Application Programming Interface (API)[2] called streaming API, which allows obtaining a stream of real-time tweets and sets of tweets from the past up to last seven days by querying their content. In order to retrieve a collection of relevant tweets matching a specified query for Libyan dialect, we create a set of search queries to increase the accuracy of obtaining tweets that are probably to be Libyan dialect. To this end, and getting more precise querying results, three types of querying were used for this. Firstly, a collection of Libyan dialect keywords were used in the first type of querying while the geocode system was used in the second type of querying, searching users and tweets by location is really useful for our task. Finally, the third type of querying is to combine two previously mentioned querying types in one sophisticated search.

### 3.1. Keywords Querying

Twitter also offers another type of APIs so-called REST APIs beside the streaming APIs for searching tweets,

It provides two main functionalities to request information from web servers in different ways: GET and POST methods. To be able to query the twitter search APIs in GET method, we have to send the URL of API services and other information to web servers, this other information can be added to the end of the URL of API by putting some search parameters. Here is the example (1) that illustrates the basic structure of a Twitter search API query for searching for Libya keyword in GET method.

(1) https://api.twitter.com/1.1/search/tweets.json?q=libya&count=100

**where:**
**?**=the beginning of searching parameters
**q** = the search string
**count** = the number of tweets we want to get

The other method POST is basically used when submitting a form. There is no exception in selection either of them for searching tweets. Normally, the GET method can be used when searching data such as getting tweets, for writing data such as sending tweets can be done by the POST method. In our research, the GET method was used to collect data from Twitter thus we had to identify a set of keywords for each fifteen search categories. Table 1 shows a sample of keywords that used to search for a set of possibly relevant tweets from Twitter.

Table 1. A Set of Query-Keywords Used for Retrieving Data from Twitter Social Media Network

| ID | Libyan Dialect keyword | English meaning | relevant tweet which contains certain keyword | English meaning of tweet |
|---|---|---|---|---|
| 1 | هلبة | Many or much | تحب نوعية البنات الي يتكلموا هلبة ؟؟ - قصدك فيه نوع ثاني ؟ | You like girls who talk a lot, do you mean there are the others who talk less |
| 2 | باهي | Good, ok or agree | باهي اكويس انتي هكي اطعمتي جائع | Good, that means you fed the hungry |
| 3 | شنو | What | خلونه نتمرد على الي يصير من وزارة التربية والإهمال الصاير في مجال التعليم شنو ذنب هالطلاب واهلهم | Let's rebel against what happened on the Ministry of Education and the Neglect in Education and what guilt that the student and their families did |
| 4 | حقيته | See it or find it | العمق والمشاعر اللي حقيته فالسوشال ميديا محقيتش على أرض الواقع. | I saw depth emotion on social media more than reality |
| 5 | باتي | My father | باتي يبيني نعزفله نسم علينا الهوي | My father ask me to play the music of the Nacm 3lyna alhawa |
| 6 | مصبي | Stand, bias | حكم مصبي مع الاهلي شويا | The referee slightly bias towards Alahaly team |
| 7 | زردة | Picnic | ليا هلبا ما طلعتش وسنين ما مشيتش زردة اليوم عوضنا كل شيء | I haven't hanged out for a long time, for 2 years I haven't gone picnic, today ,I recoup everything |
| 8 | صقع | Cold, very difficult | لنقنه صقععع ياسي الشتا خاش بقوة واندفاع | It is very cold, man , the winter is coming soon strongly |
| 9 | ندوي | Talk | انا ندوي في واقع انت تشوف فيه بروحك | I talk about reality that you see |
| 10 | علاش | Why | خالد معندكش خلفية علاش النت يفصل كل يوم العشية ! | Khaled, do you know why the internet shut down every evening |

Since our goal is to identify the search keywords that are very relevant to the targeted categories such as

[2]https://dev.twitter.com

politics, news, sports and so on. Therefore, we made a big effort to collect approximately 120 unique keywords collectively for all categories, These keywords were used for retrieving relevant tweets. In some cases, more than one tweet belong to different categories might be retrieved by certain keyword because the tweets queries probably contain too general keywords. For instance, the keyword "هلبة" which means in English "many" or "much" used for retrieving several tweets that might be classified into many categories. Although some of these keywords that are not suitable to be search keywords for specific category because they are usually too general, there are many keywords that are highly related to the specific category. For example, the Libyan prime minister "السراج" keyword is highly related to the politics category.

## 3.2. Searching Twitter by Location

In this type of querying returns tweets by users located within a given radius of the given latitude and longitude. The Twitter search API has a geocode parameters we can pass to search for tweets. The parameters values are specified by latitude, longitude and a radius, where radius units must be specified as either "mi" (miles) or "km" (kilometers). The parameters must be written in the same order as previously mentioned. In our research, we precisely identified six sites on the map of Libya such "circles" that cover all areas of searching so that we know what tweets are coming from a specific area. Fig. 2 shows the six circles which represent the six areas of Libyan dialects on the map of Libya.

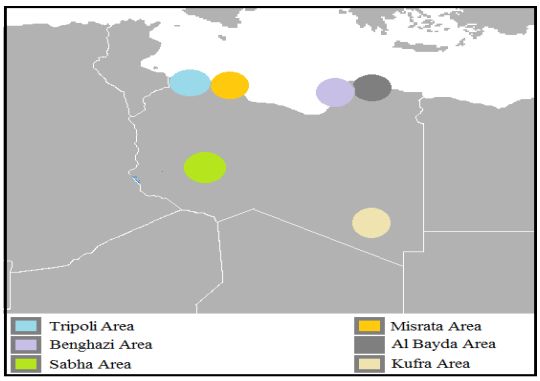


Fig. 2. Illustrates the six areas for collecting Libyan dialects.

Here is the example (2) for showing how to query a specific area and get the top 100 tweets in the JSON format.

(2) https://api.twitter.com/1.1/search/tweets.json?q=&geocode=32.8872090,13.1913380,100mi&count=100

**where:**
latitude=32.8872090
longitude=13.1913380
radius=100 mile

The above example is a very precise way to restrict the query by giving location information. In our example, the geocode parameters are given the values of latitude=32.8872090, longitude=13.1913380 and radius=100 mile. When conducting searching by using geocode, the search API will first attempt to find tweets which have latitude and longitude that match geocode in searching queries. In case of not having success, it will attempt to find tweets created by users who have the same geocode in their profile location. The search string above would provide up to 100 tweets according to search terms within 100 miles of Tripoli area. To find the accurate information for Twitter location search, we used Geocode website[3] to find the Geocodes of our six interested areas. Table 2 shows the six Libyan cities which represent the best areas for collecting three Libyan dialects.

Table 2. Geocode of Six Libyan Cities for Collecting Libyan Dialects.

| District | Latitude | Longitude | Radius in mile |
|---|---|---|---|
| Tripoli | 32.8872090<br>32° 53' 13.95'' N | 13.1913380<br>13° 11' 28.82'' E | 100 |
| Benghazi | 32.0947710<br>32° 5' 41.18'' N | 20.1879110<br>20° 11' 16.48'' E | 100 |
| Sabha | 27.0087130<br>27° 0' 31.37''N | 14.4490400<br>14° 26' 56.54'' E | 100 |
| Misrata | 32.3196830<br>32° 19' 10.86'' N | 15.1025080<br>15° 6' 9.03'' E | 100 |
| Al Bayda | 32.7564170<br>32° 45' 23.10'' N | 21.7376180<br>21° 44' 15.42'' E | 100 |
| Kufra | 23.3112390<br>23° 18' 40.46'' N | 21.8568590<br>21° 51' 24.69'' E | 100 |

### 3.3. Searching Twitter by Location and Keywords

To enhance the search results of tweets, we combined keywords with location parameters in one search. The obtained search is incredibly powerful for refining search results within a certain area. The following example (3) illustrates how the keywords and location parameters are combined in one superb search query.

(3) https://api.twitter.com/1.1/search/tweets.json?q=keyword&geocode=32.8872090,13.1913380,100mi&count=100

**where:**

keyword= contains the Libyan prime minster " السراج "
latitude=32.8872090
longitude=13.1913380
radius=100 mile

By combining fields in previous search query, the search results have become more fruitful and powerful. In former example, the search query was used for retrieving a number of tweets that contained the keyword " السراج " which is the Libyan prime minster of Libya and came from the area of Tripoli. In addition, Twitter search allows us to filter Tweets by removing the noise from Twitter search results via setting different parameters to define what data to request. For example, the parameter LANG is used for filtering tweets and return tweets that are written in a specific language. Once, set the parameter LANG by the AR value, the returned tweets would only contain the Arabic script.

## 4. Data Pre-Processing and Corpus Format

Unimportant text removal is an important step that should be considered during the preprocessing stages. In this stages, we developed a new tool to corpus preprocessing, it was needed for eliminating all unimportant text from tweets. For example, all of the following Twitter @-mentions, emoticons, URLs and #hash-tags were separately removed from each tweet by tool in a preprocessing chain. After that, we had to remove some repeated sentences manually. The remained tweets were organized into sentences. Then we have decided to store sentences and all relevant data in a way to be more easier to handle in terms of data exchange and storage format. To do this, corpus data was stored in two data type format. Firstly, all data of corpus was stored in a relational database (MySQL) while the second type of data format is written in standard XML format.

We properly specified the corpus structure in standard XML format to be readable for both human and machines. The current version of our XML corpus comprises an one major element named <sentence>, the main element has two child elements named <text> and <category> as well as it contains the attribute named

[3]www.mygeoposition.com

ID for representing the sentence number. See example (4) from our corpus in standard XML format.

```
- <Sentence id="1970">
    <Text> ؟؟ جماعة الرناسي صرحوا ولا يستنوا في تغريدة كويلر باهي</Text>
    <Category> News</Category>
  </Sentence>
- <Sentence id="1971">
    <Text> جينتيلوني: هناك ضرورة للوصول إلى اتفاق مع حفتر</Text>
    <Category> News</Category>
  </Sentence>
```

example 4: sample of corpus data is stored in standard XML format

## 5. Statistical Results

The statistical results show some statistical information about the tweets which are written by Libyan Twitter's users during the period from 1st Oct, 2016 to 30th Nov, 2016. Overall, the numbers of tweets in fifteen categories show a considerable diversity. However, there was an upward trend in the number of both sports and news tweets, whereas there was a downward trend in the number of films and industry.

In detail, the most popular topic among Libyan Twitter's users is sport, with participation rate reaching 19.04 per cent and 18.8 per cent of tweets were in news and has the second highest rate of tweets. Obviously, the sports and news topics are clearly the favorite topics for Libyan Twitter's users. Loves topic accounted for a reasonable rate which is just slightly over the number of religion and economics topics with only about 7.64 per cent of tweets. While the rest of tweets topics film, culture and art have lower levels of popularity, and their participation rates were approximately 2.94, 4.16 and 4.8 per cent respectively. See Fig. 3, which shows the fifteen categories and their percentage.

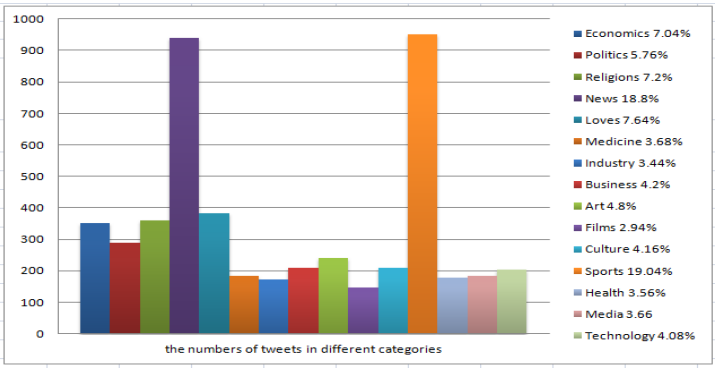


Fig. 3. The bar chart illustrates the percentages of fifteen categories.

On the other side, we also conducted statistical analysis in the entire data of corpus to identify a number of frequency for each keyword which used to collect the corpus. As shown in Table 3, the most frequent keyword in the corpus is "شنو" which means in English "what" while the least frequent keyword in the corpus is "علاش" that means in English "why".

Table 3. The Five Most Frequently Used Keyword in the Corpus.

| ID | keyword | English | Frequency |
|---|---|---|---|
| 1 | شنو | What | 181 |
| 2 | باهي | good, nice, ok or agree | 111 |
| 3 | هلبة | much, many | 106 |
| 4 | علاش | Why | 98 |
| 5 | بلاش | cancel, without money | 115 |

## 6. Conclusion and Future work

In this paper, we studied the main problem that face researchers who study linguistics and do NLP researches in Libyan dialect. The absence of any Libyan dialect corpus has made a big problem to researchers because almost all researches which need social media data to carry out have to use such a corpus. We collect a corpus of 5000 tweets which are manually classified into many categories. To enhance usability and to achieve an easier access to corpus data, the data was stored by using two general format MySQL database and standard XML format. Statistical results obtained from analyzing corpus data showed that 19.04 per cent of tweets was in sports and the smallest proportion was in films with 2.94 per cent while the rest of tweets constitute about 78.02 per cent collectively. Eventually, we believe that this corpus will prove valuable to researchers working in social media and natural language processing.

In our future work, we plan to develop a set of corpuses that are absolutely necessary to study linguistics and NLP researches for Libyan dialect such as building a new Libyan dialect corpus for sentiment analysis.

## References


[1] Assiri, A., & Emam, A., & Al-Dossari, H. (2016). Saudi Twitter corpus for sentiment analysis, world academy of science. *Engineering and Technology International Journal of Cognitive and Language Sciences*, *3(2).*

[2] Derczynski, L., Bontcheva, K., & Roberts, I. (2016). Broad twitter corpus: A diverse named entity recognition resource. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*. Technical Papers.

[3] Sang, E. T. K., & Bosch, A. (2013). Dealing with big data: The case of Twitter. *Computational Linguistics in the Netherlands Journal 3*, 121-134.

[4] Ashour, A. (2014). *Code Switching Between Tamazight And Arabic In The First Libyan Berber News Broadcast: An Application Of Myers-Scotton's Mlf And 4m Models.* Master thesis, Portland State University.

[5] Sa˜sa, P., Osborne, M., & Lavrenko, V. (2010). The Edinburgh Twitter corpus. *Computational Linguistics in a World of Social Media*.

[6] Sang, T. K. (2011). Het gebruik van Twitter voor taalkundig onderzoek. *TABU: Bulletin voor Taalwetenschap, 39 (1/2),* 62–72.


**Husien B. Alhammi** was born in Libya on January 17, 1974. He received the B.Sc. degree in computer science from Zawia University in 1996 and he was awarded the M.Sc. degree in computer science by Libyan Academy in 2008. Since then, he has worked in Department of Electrical and Electronic, Polytechnic Institute of Zawia, Libya as assistant lecturer. Currently, he is planning to do his Ph.D in computer science in Turkey.

**Ramadan Alfared** was born in Libya on November 02, 1973. He received the B.Sc. degree in computer science from Zawia University in 1995 and he was awarded the M.Sc. degree in computer science by University of Nancy (France) in 2008 and Ph.D degree from University of Nantes (France) in 2012. Since then, he has worked in Department of Computer, University of Zawia , Libya as a lecturer.