Annotated Corpus of Mesopotamian-Iraqi Dialect for Sentiment Analysis in Social Media

AL-KHAFAJI ALI J ASKAR[1], NILAM NUR AMIR SJARIF[2]
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia (UTM), Kuala Lumpur, Malaysia

***Abstract*—Research on Sentiment Analysis in social media by using Mesopotamian-Iraqi Dialect (MID) of Arabic language was rarely found, there is no reliable dataset developed in MID neither an annotated corpus for the sentiment analysis of social media in this dialect. Therefore, this gap was the main stumbling block for researchers of sentiment analysis in MID, for this reason, this paper introduced the development of an annotated corpus of Mesopotamian-Iraqi Dialect for sentiment analysis in social media and named it as (ACMID) stands for (the annotated corpus of Mesopotamian-Iraqi Dialect) to help researchers in future for using this corpus for their studies, to the best of our knowledge this is the first annotated corpus that both classify polarity as well as emotion classification in MID. Likewise, Facebook as the most popular social platform among Iraqis was used to extract the data from its popular Iraqi pages. 5000 comments were extracted from these pages classified by its polarity (Positive, Negative, Neutral, Spam) by two Iraqi annotators, these annotators were simultaneously classifying the same comments according to Ekman seven universal emotions (Anger, Fear, Disgust, Happiness, Sadness, Surprise, Contempt) or no emotion. Cohen's kappa coefficient was then used to compare the two annotators' results to find the reliability of these results. The data shows a comparable value among the two annotators for the polarity classification as high as 0.82, while for the emotion classification the result was 0.65.**

Keywords*—Sentiment analysis; Mesopotamian dialect; Iraqi dialect; social media; annotated corpus; emotion classification; Arabic language

I. INTRODUCTION

Mesopotamian-Iraqi Dialect (MID) is a main dialect of Arabic among more than 40 million people in Iraq and its neighbors. Making it the second most popular dialect of Arabic after the Egyptian dialect (which reach around 100 million speakers) in the Arab world. Facebook is the most popular social network among Iraqis, and usually, Iraqi people use their dialect in Facebook comments and posts.

Iraq is an important country in the region of the Middle East and the whole world, it is the cradle of civilization and one of the wealthiest countries in the world in its oil reserves and production that might affect the world economy, Iraq was the main front in so many global events during human history, it's hard to find someone in the world does not hear about Iraq because of the events that keep happening there.

Therefore, MID as a dialect for most residents of this country has an important role to extract the opinion of its people to have full knowledge of their thoughts and thinking better than hear their thoughts from others that cannot be mostly correct and lead to be misleading. Also, understanding people's opinions can be useful in making trading and social decision as well as investing in so many fields of the economy.

Social Media is the main source of getting people's opinions, by extracting data from people's comments and posts useful information can be introduced after classify its polarity and emotion towards certain events and ideas. Facebook as mentioned before is the main platform of social media using by Iraqi people, it has more than 21 million users in Iraq [1], extracting data from Iraqi pages of Facebook can be so useful to get people's thoughts and opinions.

Regardless of the Important of Mesopotamian-Iraqi Dialect (MID) in the world (and Arabic Language in general), studies on Sentiment Analysis in social media using this dialect is so rare and there is no real dataset developed in MID neither an annotated corpus that can be relay on for the sentiment analysis of social media in this dialect [2].

Some Researchers preferred to do their researches on the English version on the original Arabic text instead, because of the complexity of Arabic language in general and the features that facilitates the extracting of the result in the English language to get a more accurate result [3].

Therefore, this gap was the main stumbling block for researchers of sentiment analysis in MID, for this reason, this paper will introduce a new annotated corpus named (ACMID) extracting its data from popular Iraqi Facebook pages to help researchers in the future using this corpus for their studies and researches on sentiment analysis in social media used MID.

To make the new annotated corpus ACMID, Facebook was used to extract the data from its popular Iraqi pages as it is the most popular social platform among Iraqis. 5000 comments were extracted from these pages classified by its polarity (Positive, Negative, Neutral, Spam) by two Iraqi annotators, these annotators were simultaneously classifying the same comments according to Ekman seven universal emotions (Anger, Fear, Disgust, Happiness, Sadness, Surprise, Contempt) or no emotion.

In this paper, related works will be stated in the next section, a brief description for Arabic dialects will be shown in the third section, the fourth section will demonstrate the data collection and pre-processing, the fifth section will state the data annotation and the rules that have to be followed by the annotators, while the sixth section will discuss the results of this work.

II. Related Works

Related works for sentiment analysis in MID are so limited, most of the related works in the Arabic language are available in MSA and some regional dialects of Egypt (Egyptian dialect), Saudi Arabia (Najidi and Gulf Arabic dialects which referred to as Saudi dialect at most) and other dialects of Arabic language (Levanti, Meghribi, etc.).

AWATEF corpus one of the most reliable corpus by researchers of Arabic, AWATEF corpus was extracting its data from different sources in MSA [4]. COLABA (Cross-Lingual Arabic Blog Alerts) is a project in many Arabic dialects including MID was developing Natural Language Processing (NLP) resources for these dialects [5]. On the other hand, DIWAN software was developed to help training annotators to create their tagging corpus, it can capture the morphological characters in a certain text [6]. Itani et al. build Arabic corpora by extracting their data from Arabic Facebook pages (Al-Arabiyya and the voice) [7].

Al-Kabi et al. [8] create an Arabic corpus from reviews written in MSA and in addition to five Arabic dialects (Egypt dialect, Levant dialect, Arab Peninsula dialect, Maghrebi dialect, and Mesopotamian-Iraqi dialect), this corpus has 250 topics and 1442 reviews.

Meanwhile, many researchers were done studying sentiment analysis in Saudi Arabic dialect, Assiri et al. created the first reliable Saudi annotated corpus from Twitter comments [9]. While SDTC [10] was the first Saudi twitter corpus labeled by three annotators.

Alnawas et al. [11] were one of the few researchers who focuses on MID as the dialect of their interest, they used Doc2Vec to represent for binary classifier of machine learning (Decision Tree, Logistic Regression, Naïve Bayes and Support Vector Machine).

III. MSA, CA/QA and MID

Modern Arabic Language (MSA) was derived from the Classic Arabic CA in the late 19th century and the beginning of the 20th century by Arab linguistic scholars as a modern form of the CA. MSA is used widely in the Arab world (Arab Homeland as prefer to call by Arabs) as the main language for learning, writing, the conversation among educated people in the universities, legislation, and other formal speech, and sometimes as a lingua franca among Arabs from different dialects of remote regions that cannot be intelligible understood between their speakers (e.g. Iraqi speaking with Algerian).

Classic Arabic Language (CA) or Quranic Arabic (QA) is the root language of all other Arabic dialects. It is based on the text of the Quran (The holy book of Muslims around the world), Quran was first introduced in the 7th century in the west part of the Arabian Peninsula which used the dialect of Arabic of that time in that region as the dialect of Arabic which eventually became the root of all Arabic dialects since.

Most of the Arab speakers cannot distinguish the differences between MSA and CA and most of them consider it as one dialect. Arab people usually named the two dialects as (Al-Arabiya Al-fusha-العربية الفصحى) [12].

Arabic dialects can be divided into five groups as mention below:

- Mesopotamian Dialects
    - South Mesopotamian Dialect (gelet)
    - North Mesopotamian Dialect (geltu)
- Levantine Dialects
    - North Levantine Arabic
        - Syrian Arabic
        - Lebanese Arabic
        - Çukurova Arabic
    - South Levantine Arabic
        - Jordanian Arabic
        - Palestinian Arabic
- Bedawi Arabic
- Arabian Peninsula Dialects
    - Najdi Arabic
    - Gulf Arabic
    - Bahrani Arabic
    - Hejazi Arabic
    - Yemeni Arabic
    - Omani Arabic
    - Dhofari Arabic
    - Shihhi Arabic
- Egypto-Sudanic Dialects
    - Sudanese Arabic
    - Egyptian Arabic
    - Sa'idi Arabic
    - Chadian Arabic
- Magheribi Dialects
    - Moroccan Arabic
    - Algerian Arabic
    - Tunisian Arabic
    - Libyan Arabic
    - Saharan Arabic
    - Hassaniya Arabic

Mesopotamian-Iraqi Dialect (MID) is a main dialect of Arabic in most of the present-day country of Iraq, some regions in Iraqi neighbors as well as Iraqi people in diaspora around the world. People of this region usually use MID as their mother tongue in their daily conversation while using Modern Standard Arabic MSA in writing, formal conversation, and

media. Using MID in witting was so rare all the time from its development during the last 10 centuries ago until the inventing of the Internet and the phone which was used for texting and chatting at first and then was used when social media came after. South Mesopotamian Dialects (gelet) was used in this work, as it is the main dialect among Iraqis, especially in Baghdad the largest city and the capital of Iraq, Iraqis mostly used this dialect in social media even people from the north part of Iraq [13].

IV. Data Extracting and Pre-processing

Facebook as one of the most popular social media platforms among Iraqi people was used as a source to extract data in Mesopotamian-Iraqi Dialect for sentiment analysis. Three Iraqi Facebook pages was the target to get the data from its comments on different kinds of posts of these pages. The first page called ("دليل مطاعم بغداد", Baghdad Restaurants Directory (which has more than one million followers, the second page called ("برنامج ولاية بطيخ", Melon City show) which belongs to a famous comedian show among Iraqis and has more than three million followers, while the third page as unofficial page of Baghdad university which called ("جامِعة بغداد university of Baghdad") and has around forty thousand followers at the time this paper was written.

Facepager an application for retrieving data from the web was used to extract data from Facebook. At first, getting the address ID of the Facebook page from the Findmyfbid website to specify the page that comments will be retrieved from by Facepager and then extracting these comments to a CSV file.

In the next step pre-processing of the retrieval data will take place by the following procedures:

- Remove empty comments from the corpus.
- Remove comments that contain just a tagged name without a real review.
- Remove redundancies from the corpus.
- Remove Facebook reactions (like, love, haha, wow, sad, angry).
- Remove serious bad words that cannot be acceptable in any way.
- Remove comments that contains just one character or simple (e.g., ".", "م").
- Remove any comment that wasn't written in MID or the Arabic language in general.

V. Data Annotation

To make the new annotated corpus ACMID two Iraqi Arab native speakers (one doctor in his thirties and one engineer 25 years old) will be involved tagging each comment that was extracted from Facebook pages and classifying them according to their polarity, the polarity classification will be either Positive, Negative or Neutral.

Simultaneously, the annotators will classify these comments according to Ekman's seven universal emotions (Anger, Fear, Disgust, Happiness, Sadness, Surprise, Contempt) [14] and if it shows no emotion the annotator will tag it as (no emotion).

The classification of these comments will be done according to the following steps and rules:

- A brief explanation about sentiment analysis will be given to the annotators.
- An example of annotating five comments will be shown to the annotators.
- At first, annotators will be asked to classify ten comments only.
- After that, a short discussion among annotators and their works will take place.
- Annotators will be asked then to complete tagging all the comments separately.
- Annotators will be asked not to discuss their work with each other.
- Annotators will be asked not to influence their personal views about a certain topic in their classification.

VI. Results and Discussion

The 5000 comments will be classified according to their polarity and emotions by two annotators as mentioned in the previous sections. The polarity will be either positive, negative, neutral or spam, these classifications will give a wide range for the annotators to classify the comments according to their polarity, not limit their choices to the positive or negative classification which might be confusing in some comments for the annotator to choose accordingly.

The second classification is about emotion according to Ekman seven universal emotions (Anger, Fear, Disgust, Happiness, Sadness, Surprise, and Contempt) and if the annotator saw there is no emotion to show in a certain comment, he can then choose the eighth choice which it is (no emotion).

The results of classification according to their polarity for the first annotator shows that positive toke 2243 comments out of 5000 with a percentage of 44.86%, while negative toke 1682 comments out of 5000 with a percentage of 33.64%, the neutral recorded 1038 out of the 5000 comments with a percentage of 20.76%, and finally the spam recorded only 37 comments out of 5000 comments with a percentage of 0.74%.

The second annotator has the following results, positive recorded 2179 comments out of 5000 with a percentage of 43.58%, negative 1662 comments out of 5000 with a percentage of 33.24%, the neutral recorded 1080 out of the 5000 comments with a percentage of 21.6%, and the spam recorded the same result of the first annotator of 79 comments out of 5000 comments with a percentage of 1.58%.

Table I shows that the annotators agreed on 88.32% for the comment's classification according to their polarity which is considered as so high.

TABLE I. MATRIX ILLUSTRATION FOR THE CONFUSION BETWEEN FIRST AND SECOND ANNOTATORS FOR THE POLARITY CLASSIFICATION

	Positive	Negative	Neutral	Spam	Total
Positive	2034	65	78	2	2179
Negative	60	1501	101	0	1662
Neutral	115	111	850	4	1080
Spam	34	5	9	31	79
Total	2243	1682	1038	37	5000

To ensure the reliability of the result for the polarity classification Cohen Kappa coefficient was used to compare the results between the two annotators, Cohen Kappa is used to measure inter-rater reliability for qualitative items [15], when κ takes into account the possibility of the agreement by chance (AC).

The following formula will show the Cohen Kappa coefficient for the agreement between the two annotators:

OA:(2034+1501+850+31)/5000=0.8832

AC:0.4358*0.4486+0.3324*0.3364+0.216*0.2076+0.0158*0.0074

AC: 0.1955+0.11182+0.04484+0.00012

AC: 0.35228

$\kappa = (OA-AC) / (1-AC)$

$\kappa = (0.8832-0.35228) / (1-0.35228)$

$\kappa =0.53092/0.64772$

$\kappa =0.8196751682825912$

The final result for polarity classification shows the Kappa coefficient for the agreement between the two annotators as high as (0.82).

The classification of emotions shows the result for the first annotator as the following: (Anger= “256” out of 5000 comments with a percentage equal to “5.12%”, Fear= “38” out of 5000 comments with a percentage equal to “0.76%”, Disgust= “227” out of 5000 comments with a percentage equal to “4.54%”, Happiness= “976” out of 5000 comments with a percentage equal to “19.52%”, Sadness= “346” out of 5000 comments with a percentage equal to “6.92%”, Surprise= “336” out of 5000 comments with a percentage equal to “6.72%”, Contempt= “400” out of 5000 comments with a percentage equal to “8%”, and No emotion= “2421” out of 5000 comments with a percentage equal to “48.42%”).

While the result from the second annotator was as the following: (Anger= “369” out of 5000 comments with a percentage equal to “7.38%”, Fear= “45” out of 5000 comments with a percentage equal to “0.9%”, Disgust= “198” out of 5000 comments with a percentage equal to “3.96%”, Happiness= “803” out of 5000 comments with a percentage equal to “16.06%”, Sadness= “360” out of 5000 comments with a percentage equal to “7.2%”, Surprise= “347” out of 5000 comments with a percentage equal to “6.94%”, Contempt= “422” out of 5000 comments with a percentage equal to “8.44%”, and No emotion= “2456” out of 5000 comments with a percentage equal to “49.12%”).

Table II shows that the annotators agreed on 75.06% for the comment’s classification according to their emotions.

Cohen Kappa coefficient again was used to compare the results between the two annotators for the emotion’s classification, the following formula shows the Cohen Kappa coefficient for the agreement between the two annotators:

OA:(2004+188+280+168+21+610+243+239)/5000=0.7506

AC:0.4912*0.4842+0.0738*0.0512+0.0844*0.08+0.0396*0.0454+0.009*0.0076+0.1606*0.1952+0.072*0.0692+0.00694*0.0672

AC:0.2378+0.0038+0.0068+0.0018+0.0000684+0.0313+0.005+0.0047

AC: 0.2912

$\kappa = (OA-AC) / (1-AC)$

$\kappa = (0.7506-0.2912) / (1-0.2912)$

$\kappa =0.4594/0.7088$

$\kappa =0.64813769751693$

The final result for emotion classification shows the Kappa coefficient for the agreement between the two annotators as (0.65).

TABLE II. MATRIX ILLUSTRATION FOR THE CONFUSION BETWEEN FIRST AND SECOND ANNOTATORS FOR THE EMOTION’S CLASSIFICATION

	No-Emotion	Anger	Contempt	Disgust	Fear	Joy	Sadness	Surprise	Total
No-Emotion	2004	20	27	8	5	327	28	37	2456
Anger	72	188	37	13	3	8	25	23	369
Contempt	63	20	280	24	2	6	11	16	422
Disgust	9	2	12	168	0	2	3	2	198
Fear	8	3	4	0	21	4	4	1	45
Joy	136	6	17	6	3	610	17	8	803
Sadness	67	8	12	7	2	11	243	10	360
Surprise	62	9	11	1	2	8	15	239	347
Total	2421	256	400	227	38	976	346	336	5000

VII. CONCLUSION

Mesopotamian-Iraqi Dialect (MID) is a main dialect of Arabic, Researches that have interested in this dialect were so rare, researchers have difficulties studying sentiment analysis in this dialect because of the lack of reliable annotated corpus in MID as well as a real dataset.

To the best of our knowledge, this paper was introduced the first annotated corpus ACMID that both classify polarity as well as emotion classification in MID. Two annotators were involved to tag the extracted data of comments from three Iraqi famous face pages. The result shows the Kappa coefficient for the agreement between the two annotators for the polarity classification as high as 0.82, while for the emotion classification the result was as 0.65.

Future plan is to applied Machine Learning techniques on the created corpus ACMID (Annotated Corpus of Mesopotamian-Iraqi Dialect).

REFERENCES

[1] World Population Review, "Facebook Users by Country 2021." https://worldpopulationreview.com/country-rankings/facebook-users-by-country.

[2] M. E. M. Abo, R. G. Raj, and A. Qazi, "A Review on Arabic Sentiment Analysis: State-of-the-Art, Taxonomy and Open Research Challenges," IEEE Access, vol. 7, pp. 162008–162024, 2019.

[3] M. A. Ahmed, H. Baharin, and P. N. E. Nohuddin, "Analysis of K-means, DBSCAN and OPTICS Cluster Algorithms on Al-Quran Verses," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 8, 2020, doi: 10.14569/IJACSA.2020.0110832.

[4] M. Abdul-Mageed and M. T. Diab, "AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis.," in LREC, 2012, vol. 515, pp. 3907–3914.

[5] M. Diab, N. Habash, O. Rambow, M. Altantawy, and Y. Benajiba, "COLABA: Arabic dialect annotation and processing," in Lrec workshop on semitic language processing, 2010, pp. 66–74.

[6] F. Al-Shargi and O. Rambow, "Diwan: A dialectal word annotation tool for Arabic," in Proceedings of the Second Workshop on Arabic Natural Language Processing, 2015, pp. 49–58.

[7] M. Itani, C. Roast, and S. Al-Khayatt, "Corpora for sentiment analysis of Arabic text in social media," in 2017 8th international conference on information and communication systems (ICICS), 2017, pp. 64–69.

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

[9] A. Assiri, A. Emam, and H. Al-Dossari, "Saudi twitter corpus for sentiment analysis," Int. J. Comput. Inf. Eng., vol. 10, no. 2, pp. 272–275, 2016.

[10] A. Al-Thubaity, M. Alharbi, S. Alqahtani, and A. Aljandal, "A saudi dialect twitter corpus for sentiment and emotion analysis," in 2018 21st Saudi Computer Society National Computer Conference (NCC), 2018, pp. 1–6.

[11] A. Alnawas and N. Arici, "Sentiment analysis of iraqi Arabic dialect on Facebook based on distributed representations of documents," ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 18, no. 3, pp. 1–17, 2019.

[12] L. A. Al Suwaiyan, "Diglossia in the Arabic Language," Int. J. Lang. Linguist., vol. 5, no. 3, pp. 228–238, 2018.

[13] H. Palva, "From qəltu to gələt: Diachronic notes on linguistic adaptation in Muslim Baghdad Arabic," in Arabic Dialectology, Brill, 2009, pp. 17–40.

[14] P. Ekman, "An argument for basic emotions," Cogn. \& Emot., vol. 6, no. 3–4, pp. 169–200, 1992.

[15] M. L. McHugh, "Interrater reliability: the kappa statistic," Biochem. medica, vol. 22, no. 3, pp. 276–282, 2012.