УНИВЕРСИТЕТ ИТМО

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2022 Том 22 № 2 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2022 Vol. 22 No 2 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2022-22-2-308-316

Constructing twitter corpus of Iraqi Arabic Dialect (CIAD) for sentiment analysis

Mohammed M. Hassoun Al-Jawad[1], Hasanein Alharbi[2]✉, Ahmed F. Almukhtar[3], Anwar Adnan Alnawas[4]

[1,3] University of Kerbala, College of Computer Science and Information Technology, Karbala, 56001, Iraq
[2] University of Babylon, IT College, Babylon, 51002, Iraq
[4] Southern Technical University, Nasiriyah Technical Institute, Nasiriyah, Iraq

[1] mohammedm@uokerbala.edu.iq, https://orcid.org/0000-0001-6750-0294
[2] hasanein.alharbi@uobabylon.edu.iq✉, https://orcid.org/0000-0003-2577-278X
[3] ahmmed.ahmukhtar@uokerbala.edu.iq, https://orcid.org/0000-0002-4585-3977
[4] anwar.alnawas@stu.edu.iq, https://orcid.org/0000-0001-9181-9377

Abstract

The number of Twitter users in Iraq has increased significantly in recent years. Major events, the political situation in the country, had a significant impact on the content of Twitter and affected the tweets of Iraqi users. Creating an Iraqi Arabic Dialect corpus is crucial for sentiment analysis to study such behaviors. Since no such corpus existed, this paper introduces the Corpus of Iraqi Arabic Dialect (CIAD). The corpus has been collected, annotated and made publicly accessible to other researchers for further investigation. Furthermore, the created corpus has been validated using eight different combinations of four feature-selections approaches and two versions of Support Vector Machine (SVM) algorithm. Various performance measures were calculated. The obtained accuracy, 78 %, indicates a promising potential application.

Keywords

sentiment analysis, data mining, support vector machine, user behaviors, social media mining

For citation: Hassoun Al-Jawad M.M., Alharbi H., Almukhtar A.F., Alnawas A.A. Constructing twitter corpus of Iraqi Arabic Dialect (CIAD) for sentiment analysis. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, vol. 22, no. 2, pp. 308–316. doi: 10.17586/2226-1494-2022-22-2-308-316

УДК 004.021

Создание корпуса иракского арабского диалекта в Твиттере для анализа настроений

Мохаммед М. Хассун Аль-Джавад[1], Хасанейн Альхарби[2]✉, Ахмед Ф. Альмухтар[3], Анвар Аднан Алнавас[4]

[1,3] Университет Кербелы, Колледж компьютерных наук и информационных технологий, Кербела, 56001, Ирак
[2] Университет Вавилона, Колледж информационных технологий, Вавилон, 51002, Ирак
[4] Южный технический университет, Технический институт в Насири, Насири, Ирак

[1] mohammedm@uokerbala.edu.iq, https://orcid.org/0000-0001-6750-0294
[2] Hasanein.alharbi@uobabylon.edu.iq✉, https://orcid.org/0000-0003-2577-278X
[3] ahmed.ahmukhtar@uokerbala.edu.iq, https://orcid.org/0000-0002-4585-3977
[4] anwar.alnawas@stu.edu.iq, https://orcid.org/0000-0001-9181-9377

Аннотация

За последние годы количество пользователей Twitter в Ираке значительно увеличилось. Крупные события и политическая обстановка в стране оказывают значительное влияние на информацию в Twitter и затрагивают сообщения иракских пользователей. Создание корпуса иракских арабских диалектов Corpus of Iraqi Arabic Dialect (CIAD) имеет решающее значение для анализа настроений и изучения такой информации. Представлено

исследование CIAD, который был разработан и описан другими исследователями. Корпус проверен с использованием восьми различных комбинаций четырех подходов к выбору признаков и двух версий алгоритма метода опорных векторов. Рассчитаны различные показатели эффективности. Полученная точность составила 78 %, что указывает на многообещающее потенциальное применение.

Ключевые слова
анализ настроений, интеллектуальный анализ данных, метод опорных векторов, поведение пользователей, интеллектуальный анализ социальных сетей

Ссылка для цитирования: Хассун Аль-Джавад М.М., Альхарби Х., Альмухтар А.Ф., Алнавас А.А. Создание корпуса иракского арабского диалекта в Твиттере для анализа настроений // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 2. С. 308–316 (на англ. яз.). doi: 10.17586/2226-1494-2022-22-2-308-316

Introduction

The data generated by social medial platforms is enormous. Videos, audios, text and images are produced as a result of users' interactions. Text is the most common data type used by social media users to express their ideas, feeling and thoughts. Arabic-text is the fourth most-used on the internet and the fifth most-spoken language worldwide [1, 2].

Arabic-text is available online in two forms: Modern Standard Arabic and Arabic dialects. Dialects are categorized according to various regions. The work presented in [3] breaks down the regional dialect into five groups: Egyptian, Levantine, Gulf, Maghrebi, Iraqi and others. The authors of [4] presented the five groups and added Yemenite Arabic as a sixth dialect. Fig. 1 below shows the regional classification of Arabic language dialects.

The internet is very rich with English language dataset which can be used in Sentiment Analysis (SA), whereas the Arabic language has a very small corpora available. The state of the art shows that there are no public access of the Arabic Iraqi Dialect corpus available for researchers [4–7]. Hence, this paper attempts to bridge this gap.

SA maintains different machine learning methods to detect and analyses certain patterns of the studied data. In Kumar and Jaiswal's research [8], a systematic review showed that Support Vector Machine (SVM) is the most dominant method which has been used in SA. Accordingly, the Corpus of Iraqi Arabic Dialect (CIAD) constructed in this research was validated using two versions of SVM.

The main goal of this paper is to present an open access Arabic Iraqi dialect corpus. The constructed corpus establishes a solid base for future studies on SA for Iraqi Arabic Dialect (IAD). The elements of this work are listed below:

— The raw data comprising an elongated words dictionary of Arabic Iraqi Dialect.
— Original tweets before and after preprocessing of Arabic Iraqi Dialect.
— Arabic Iraqi Dialect annotated tweets (Corpus).
— Availability of all mentioned resources for public access.

Related Works

In recent years, SA (also known as opinion mining) has attracted many researchers around the world. Various machine learning approaches are utilized in SA task. The most frequently used methods are SVM and Naïve Bayes. Convolutional Neural Network (CNN) is also considered to be promising method in this field [8]. This section will

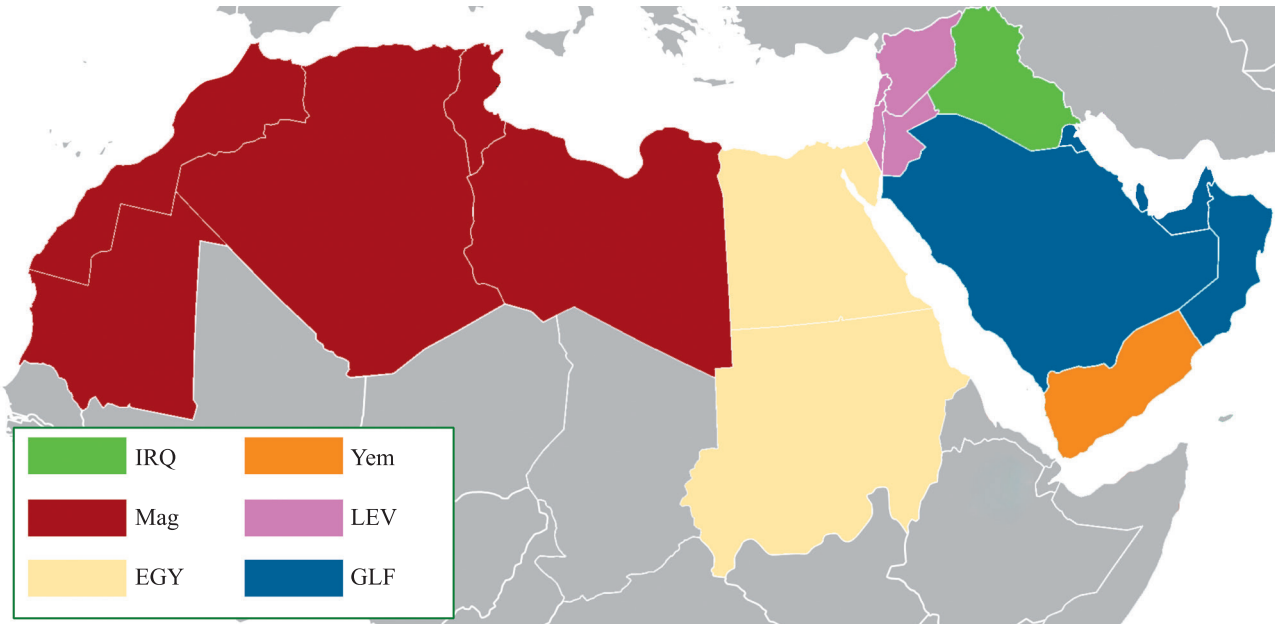

Fig. 1. Regional Arabic dialect map

review the most recent works on SA concentrated on Arabic dialects when it possible.

In [9], the authors presented a German SA corpus of labelled tweets. The methodology had employed CNN to construct a corpus of 9738 German tweets. According to the authors, computer science and linguistics students were chosen to annotate the unlabelled tweets. The conclusion of the study was that the proposed method outperformed the comparative SVM in some cases, while other cases it did not.

In [10], the authors used the ASTD (Arabic Sentiment Tweets Dataset) with 10K tweets involving the Arabic-Egyptian social sentiment, with the crawling process being done over two stages. The first stage involved SocialBakers, and specified the most effective accounts. The second stage used the most Egypt-trending hashtags. According to authors, the classification process was done in four steps. Finally, the validation was done using F1 measure.

In [11], a Saudi-Arabic sentiment corpus consists of 32063 tweets annotated using deep learning method. SVM was used as a comparative method based on accuracy criteria.

The authors of [12] experimented on IAD corpus dataset. Furthermore, the feature extraction process was based on their latent words. Four methods were applied: logistic regression, Naïve Bayes, SVM and decision tree. In conclusion, the experimental performance showed that SVM and logistic regression have better performance that the other methods.

In [13], the authors introduce a Shami (Levantine) corpus which covers four spoken countries: Syria, Lebanon, Jordan and Palestine. The corpus contains a large volume of Levantine dialects that have been collected from social media using Twitter API. Different blogs are manually elected during the data collection step. Preprocessing step had been done using four steps: diacritics removal, non-Arabic words and letters removal, unifying spelling styles, and freeing up the non-Levantine sentences. Which was done manually. Finally, the experimental results were conducted using the Naïve Bayes classifier and n-gram model.

In [14], the authors had introduced a new corpus of Arabic SA. The dataset was built from a large number of Arabic tweets. The preprocessing mechanism included removing repeated letters, dividing each tweet into multiple tokens, and eliminating the stop words. The following step emphasized on the creation of three types of lexicons: emoticons, social acronyms, and interjections. Finally, in the last step, SA was done based on three methods: Naïve Bayes, SVM and Maximum Entropy. Meanwhile, the behavior of each method was studied using representation vector model.

In [15] a Moroccan (Maghrebi) SA method for tweets microblogs was developed to investigate the Moroccan users' emotions based on negative or positive responses. According to the authors, a number of challenges had to be addressed since the Moroccan speakers use many dialects such as "Darija dialect", "Amazigh dialect" and mix of Arabic and other languages. Therefore, a number of tools were employed in the preprocessing step. Such as HDFS, to clean up tweets by removing unwanted symbols, hashtag and Uniform Resource Locators (URLs). Then, Naïve Bayes was used as the main method to classify the tweets into positive or negative. Finally, the main topics were revealed by using an Latent Dirichlet Allocation algorithm.

To sum up, the works reviewed in this sections show that different SAs of Arabic dialects and their corpora have been investigated. However, it has been found that an Arabic Iraqi Dialect corpora not available for public access. We also noted that the most dominant method, which has been used for SA, is SVM. Therefore, it has been used in the validation process on this research.

Research Methodology

The research methodology presented in this paper is illustrated in Fig. 2. The proposed approach consists of four main stages. Each stage consists of various sub-stages. In the following subsections, each stage and sub-stage are explained extensively.

Corpus Creation

The prepared corpus aims to help researchers to separate the hate speech from non-hate speech of CIAD. CIAD collects data from twitter by using the twitter Auth listener API[1]. The fetched data is filtered by hashtags and by twitter accounts from the Iraqi society, where the hashtags and accounts of Iraqi politicians and influencers have been chosen. The former were selected from trending Iraqi hashtags such as "#العراق_ينتفض", "#العراق" and "#قاطعو المنتجات الأجنبية", which means "Iraq_Uprising", "Iraq", and "Boycott imported products", while the latter were selected based on the highest number of followers.

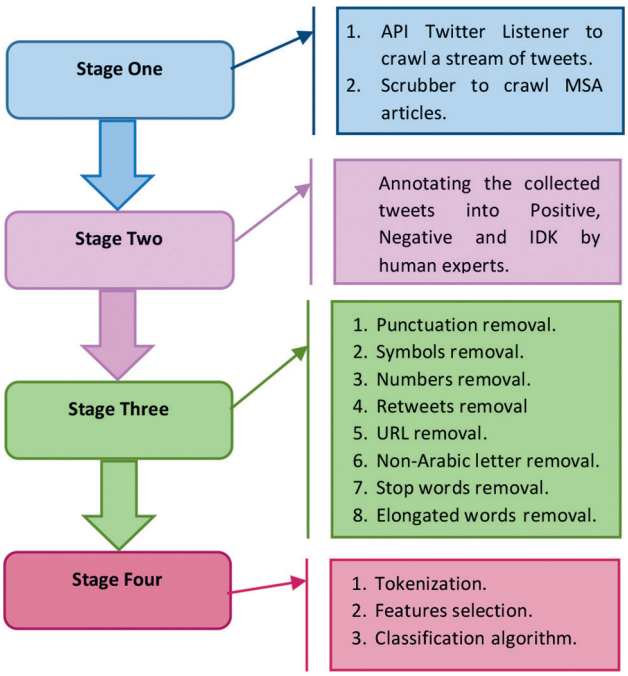

Fig. 2. Research methodology

[1] Available at: https://developer.twitter.com/en/docs (accessed 11.03.2022).

Data Annotation

The 1K tweets were annotated by three Arabic Iraqi native linguistics experts. Binary classification was used in the annotation decision since that leads to high accuracy results [16] we propose a pattern-based approach that goes deeper in the classification of texts collected from Twitter (i.e., tweets).

The experts annotated the microblogs as one clause, either as positive or negative. An expert who couldn't figure out whether a clause was positive (P) or negative (N) would annotate it as "I Don't Know" (IDK). Finally, the annotations of the three experts were collected and compared. The decision of microblog polarity is made based on the rules illustrated in Table 1. For instance, if two experts labelled a tweet as positive and negative, while the other expert annotated the same tweets as IDK, then the final annotation decision would be neutral (Nu). Table 1 shows the results for the three experts' annotation decisions.

As a rule of thumb, the number of positive and negative words and the overall context of the tweets were taken into consideration during the annotation process. The following examples discuss sample of the tweets annotated in this research and the associated authors' decisions.

Table 1. The critical cases according to experts

ID	Expert A	Expert B	Expert C	Decision
1	P	P	P	P
2	P	P	N	P
3	P	P	IDK	P
4	P	N	P	P
5	P	N	N	N
6	P	N	IDK	Nu
7	P	IDK	P	P
8	P	IDK	N	Nu
9	P	IDK	IDK	IDK
10	N	P	P	P
11	N	P	N	N
12	N	P	IDK	Nu
13	N	N	P	N
14	N	N	N	N
15	N	N	IDK	N
16	N	IDK	P	Nu
17	N	IDK	N	N
18	N	IDK	IDK	IDK
19	IDK	P	P	P
20	IDK	P	N	Nu
21	IDK	P	IDK	IDK
22	IDK	N	P	Nu
23	IDK	N	N	N
24	IDK	N	IDK	IDK
25	IDK	IDK	P	IDK
26	IDK	IDK	N	IDK
27	IDK	IDK	IDK	IDK

Example 1: sample text "أوجه لها كل الحب والاحترام لأنها تستحق ذالك" which means "I'd like to express my love and respect for here because she deserves that". It can be seen that the positive words like "love" and "respect" reflect a positive content. Additionally, the overall context of the tweet also expresses a positive felling. Therefore, the three experts labelled this tweet as positive.

Example 2: sample text "يعوزك جرت اذن" which means "your ears need to be popped". Although, the tweet did not contain any negative words, its overall context means threating in Iraqi slang dialect. Hence, all the three experts labelled it as negative.

Example 3: sample text" العراقي للمطالبة بحقوقه. دعوات أم الشهيد التي فارقت ابنها اقوى من رصاصات القتله أم الشهيد ريمون ريان تدعو الشعب" which means "Mother of the martyr Ryan Ramon, urge the Iraqi people to ask for their rights. The call of the martyr's mother who missed her son is stronger than the assassination bullets". Such tweets do not contain clear positive or negative words. Hence, the annotators relied on the context of the tweet to make their decisions. Accordingly, the three experts labelled this tweet as IDK, positive and negative, respectively.

All in all, the experts annotated 1170 tweets. Since the ultimate aim of the research was to classify each tweet as either negative or positive, the neutral (78 tweets) and IDK (540 tweets) have been excluded. Only 552 tweets have been used, 251 positive and 289 negative. The 1170 annotated tweets are freely accessible on GitHub[1]. Fig. 3 depicts the annotation process implemented by the three experts.

Data Pre-processing

The data pre-processing stage, implicitly, consists of eight sub-stages. The ultimate aim of this stage is to convert the collected data into a format that can be consumed by the mining algorithm. The eight sub-stages included in the data pre-processing stage are explained in detail below.

Punctuation removal. Preserving the original semantics of the posted tweet is an essential step in the data pre-processing stage. Hence, removing punctuation is implemented in this sub-stage. Table 2 gives an example of the executed process.

Symbols removal. It is a common practice for social media users to add various symbols in the posts. Some of these symbols reflect various form of emotional status, while others are randomly added. In this sub-stage, symbols are removed in an attempt to maintain only textual data which will be utilized in the upcoming stages. An example of the symbols removal process is explained in Table 3.

Numbers removal. The approach adopted in this research relied on the textual data only. Therefore, numbers are excluded from the original tweets. An example of the numbers removal process is illustrated in Table 4.

Retweets removal. It has been proven that the accuracy of the textual data-mining process is highly affected by the quality of the mined data. Hence, to create rich textual data,

[1] Available at: https://github.com/ebady/Iraqi-Arabic-Dialect-Dataset (accessed 28.02.2022).

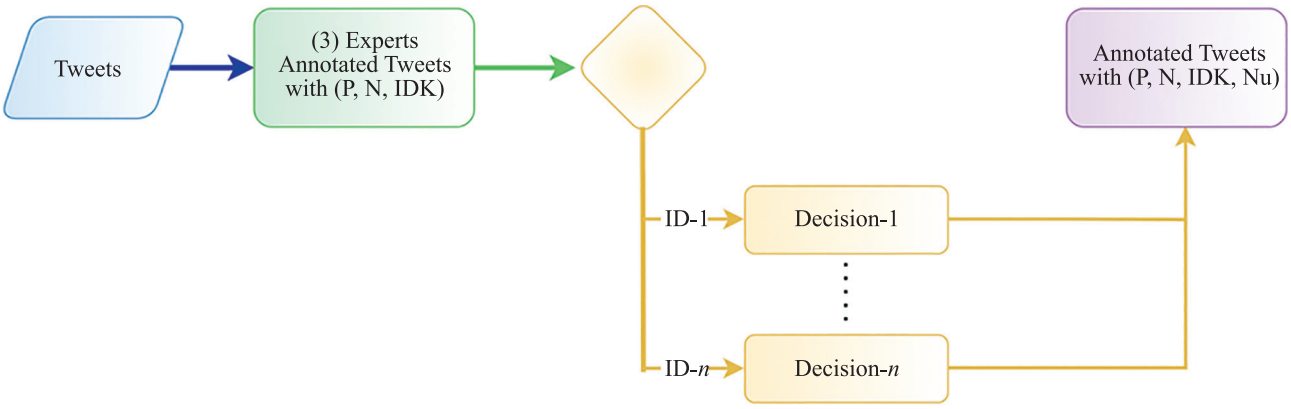

Fig. 3. Annotation process

Table 2. Punctuation removal process

Original tweets	Processed tweets
قيل لـ #رجل عاقل : - ما نراك تعيب أحداً ! - فقال : - لستُ عن #نفسي راضياً.. حتى أتفرغ لـ ذم #الناس. #العراق	قيل ل رجل عاقل ما نراك تعيب أحداً فقال لستُ عن نفسي راضياً حتى أتفرغ ل ذم الناس العراق

Table 3. Symbols removal process

Original tweets	Processed tweets
☢⚠الطائفية على الأبواب لإنهاء المظاهرات في #العراق . من خلال استخدام حرب الاسماء الطائفية القذرة بين المتظاهرين .	الطائفية على الأبواب لإنهاء المظاهرات في العراق . من خلال استخدام حرب الاسماء الطائفية القذرة بين المتظاهرين

Table 4. Number removal process

Original tweets	Processed tweets
كورونا
بعد أن وصل سعر الكمامة الواحدة في بعض مدن #العراق 15 دولار، خياط عراقي يخيط #الكمامة الطبية ويوزعها مجاناً	كورونا بعد أن وصل سعر الكمامة الواحدة في بعض مدن العراق دولار، خياط عراقي يخيط الكمامة الطبية ويوزعها مجاناً

a decision was made to eliminate repeated tweets in attempt to introduce new posts and exclude repeated sentences.

URL removal. As part of the corpus creation process, URLs are implicitly included in the collected tweets. Since this research targeted a text-based SA approach, URLs are ignored. Table 5 illustrates the URL removal process.

Non-Arabic letter removal. Social media users express their opinions using nonstandard Arabic language. Mixing Arabic and Non-Arabic letters, such as English, is widely used. Since the aim of this research is to extract Arabic text, letters from other languages are excluded. An example of the Non-Arabic letter removal process is shown in Table 6.

Stop words removal. The authors of [17] used a comprehensive list of Arabic stop words to clean the collected data. In the same way, this research utilized the same stop words list. The consulted Arabic stop words list consists of 750 words.

Elongated words processing. To emphasis their opinions or highlight a point of view, social media users widely repeat various letters in modern Arabic language.

Table 5. URL removal process

Original tweets	Processed tweets
لا احد ابداً ينسى وطنه مهما كانت ذكرياته مره &..#10;#العراق https://t.co/LpNeCMbo2X	لا احد ابداً ينسى وطنه مهما كانت ذكرياته مره &..#10;#العراق

Table 6. Non-Arabic letter removal

Original tweets	Processed tweets
لا احد ابداً ينسى وطنه مهما كانت ذكرياته مره&..#10;#العراق	لا احد ابداً ينسى وطنه مهما كانت ذكرياته مره العراق

Such behaviors generate elongated words which are inaccurate from a spelling perspective. For example, the word "عاجل", which means "breaking news" in English, is written as "عاااااااااااااجل" by repeating the letter "ا" (in English "Aleph") more than once. To deal with such cases a two-steps procedure is proposed in this research, dictionary creation and spelling correction.

Dictionary creation. The ultimate aim of this step is to create a comprehensive list of Arabic words which contain repeated letter. Two source of textual dataset have been utilized, e-newspaper article and Facebook posts, to achieve this step. Fig. 4 depicts the dictionary creation process.

Initially, articles from local Iraqi e-newspaper are extracted using web scraper software. The extracted articles are processed to identify words with repeated letters. The identified words are added to the created word dictionary. All in all, 5330 words were extracted from five local Iraqi e-newspapers.

To ensure that various sources of modern Iraqi dialect are used in the dictionary creation step, another textual dataset was extracted from Facebook. Facebook textual data is processed to identify words with repeated letters. The identified words are examined by human experts to check their spelling accuracy. Eventually, 195 words were identified form Facebook textual datasets and added to the created dictionary. Table 7 gives a breakdown of the number of the words identified from each dataset.

The identified words are added to the word dictionary in such a way that duplicated words are removed. Furthermore, the growth rate is calculated each time a new word list is added to the dictionary. The growth rate is used to indicate the percentage of newly added words. Fig. 5 shows that for the addition of the last three datasets the growth rate almost levelled out. Hence, a decision was made that no more datasets were required. Overall, 5452 words were added to the created dictionary.

Spelling correction. The process of elongated word spelling correction is illustrated in Fig. 6. The core step is comparing the elongated word extracted from the original tweets with the words in the dictionary constructed on the previous step. The comparison is performed in such a way that the number of consecutive repeated letters is disregarded if a match is found. The spelling of the extracted word is altered to match the spelling of the word in the dictionary. In addition, a special case, words starting with the Arabic letter "و" ("waw" in English) are checked.

In the Arabic language the letter "و" ("waw" in English) is used to join to words. So, if the second word starts with

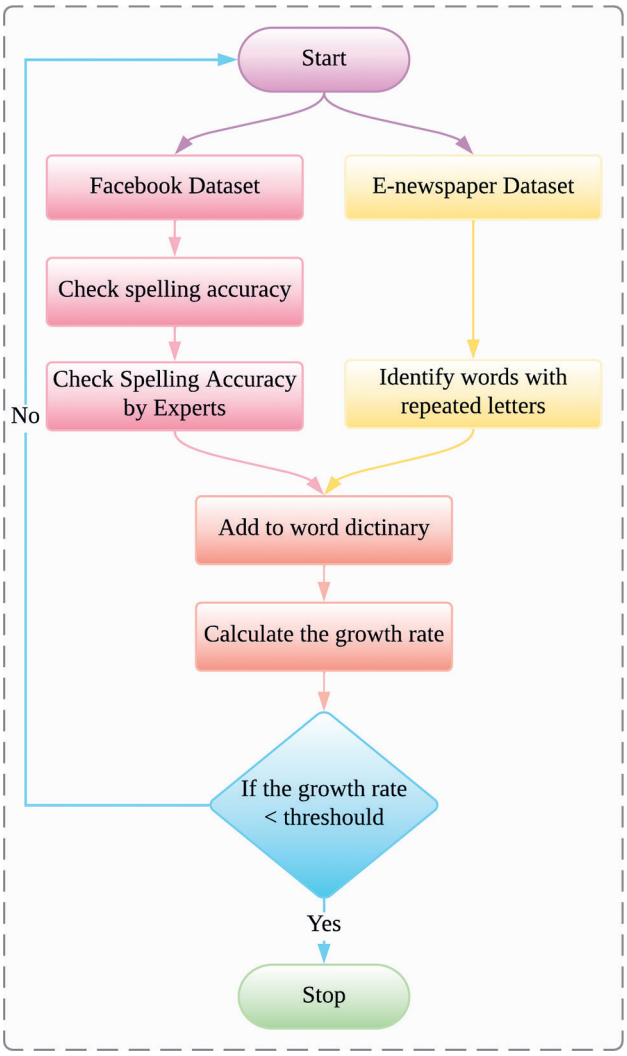

Fig. 4. Dictionary creation process

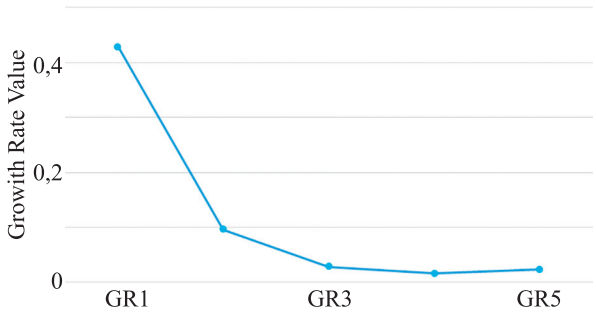

Fig. 5. Data dictionary merge growth rate level

Table 7. Words dictionary and growth rate

Dataset name	Size, KB	No. of elongated words	No. of elongated words at the word dictionary after each merge				
			Merge 1	Merge 2	Merge 3	Merge 4	Merge 5
E-News DS1	7116	3259	4655				
E-News DS2	4978	2808		5103			
E-News DS3	2354	1568			5247		
E-News DS4	1818	712				5330	
E-News DS5	1002	495					5452
FB-DS1	1049	195					

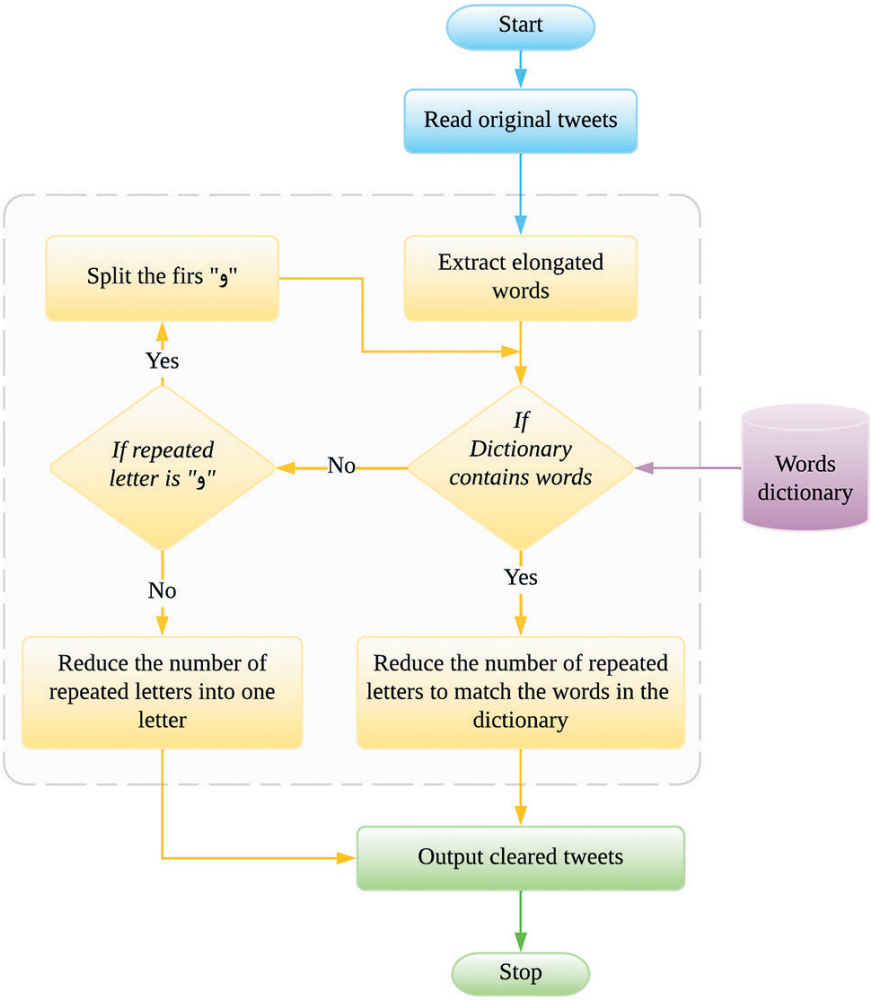

Fig. 6. Elongated words spelling correction process

the letter "waw" then the letter will be repeated more than once. Hence, to tackle such a situation the following word is compared with the words in the dictionary. As a result, the number of the repeated letters is reduced to one letter. Eventually, more accurate tweets are generated.

Data Mining

The goal of this stage is to classify the collected tweets into two categories, positive and negative. A three-step process is implemented to achieve this goal.

Tokenization. In this step the collected tweets are converted into a words vector. A Term Frequency-Inverse Document Frequency (TF-IDF) approach is executed with the help of Weka. All in all, 4541 tokens were identified.

Feature selection. To reduce the collected dataset dimensions, four feature-selection approaches are executed. Accordingly, the number of features (i.e. tokens) are reduced from 4541 to a maximum of 28 features. Table 8 shows the executed approaches and the number of features obtained on each approach.

Classification. Recent research [8, 18] reported that the SVM method has generated good results in social media text mining, consequently, this paper implemented two versions of SVM. The sequential minimal optimization (SMO) [19] was executed first, the obtained results were compared with LibSVM method. A detailed illustration of the obtained results is presented in the next section.

Table 8. Feature selection

No.	Attributes selection approach		No. of selected attributes
	Evaluator	Search method	
1	cfsSubsetEva	Best first	22
2	cfsSubsetEval	Greedy stepwise	23
3	Correlation Attribute Eval	Ranker	28
4	InfoGainAttributeEval	Ranker	23

Table 9. Performance criteria

No.	Algorithm	Accuracy, %	Precision, %	Recall, %	F1-Measure, %
1	SMO	76.2	81.2	63.7	71.4
	LibSVM	**78.1**	81.2	68.9	74.5
2	SMO	76.2	81.2	63.7	71.4
	LibSVM	65.7	95.8	27.4	42.7
3	SMO	68.14	89.1	35.9	51.1
	LibSVM	73.89	100	43.8	60.9
4	SMO	67.78	88.9	35.1	50.3
	LibSVM	72.22	100	40.2	57.4

Results Analysis

The validity of the proposed research methodology was tested using eight combinations of four feature-selection approaches and two versions of SVM classification algorithms. Additionally, performance criteria such as F1-Measure, Recall, Precision and Accuracy were calculated as well. The obtained results are summarized in Table 9.

Fig. 7 shows that, in terms of accuracy, LibSVM has outperformed SMO in three out of four tests. Furthermore, LibSVM has secured the highest accuracy obtained across all datasets. The data on the results table also indicate that LibSVM has surpassed SMO with respect to precession measure, except for the first test where both algorithms secured the same result. Similarly, LibSVM achieved better results regarding the Recall and F1-measure in three tests out of four. Together, these results suggest that LibSVM has better performance than SMO and the proposed methodology provides a satisfactory result.

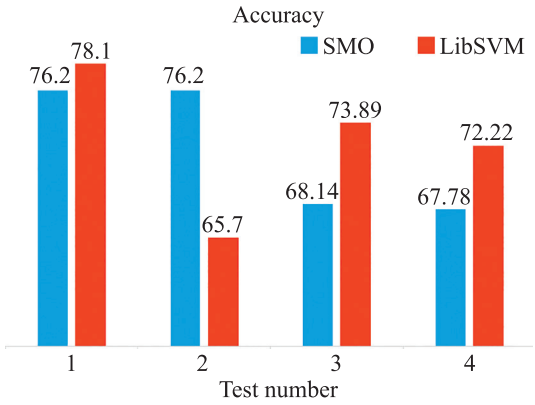

Fig. 7. Classification algorithms accuracy

Conclusion

An open access of CIAD corpus is presented in this paper. The compiled corpus consists of 1170 tweets. The collected tweets have been manually annotated by three experts. CIAD would facilitate further studies in the SA for the Iraqi Dialect to identify the implications of positive/negative tweets. An elongated words dictionary has been created to tackle the similarity between elongated words and repeated letters of standard words. Various text sources, formal and informal, have been employed to construct the presented dictionary. Eventually, 450 out of 1170 tweets were used to test two versions of SVM classifier, LibSVM and SMO. LibSVM outperformed SMO and produced the best accuracy at 78 %. The obtained results are relatively acceptable and subject to further improvements in future studies.

References

1. Stone M.L. et al. BIg data for Media. *Reuters Institute for the Study of Journalism*, 2014, november.
2. Badaro G., Baly R., Hajj H., El-Hajj W., Shaban K.B., Habash N., Al-Sallab A., Hamdi A. A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2019, vol. 18, no. 3, pp. 27. https://doi.org/10.1145/3295662
3. Zaidan O.F., Callison-Burch C. Arabic dialect identification. *Computational Linguistics*, 2014, vol. 40, no. 1, pp. 171–202. https://doi.org/10.1162/COLI_a_00169
4. Habash N.Y. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 2010, vol. 3, no. 1. https://doi.org/10.2200/S00277ED1V01Y201008HLT010
5. Alnawas A., Arici N. The corpus based approach to sentiment analysis in modern standard Arabic and Arabic dialects: A literature review. *Journal of Polytechnic — Politeknik Dergisi*, 2018, vol. 21, no. 2, pp. 461–470. https://doi.org/10.2339/politeknik.403975
6. Alshutayri A., Atwell E. Classifying Arabic dialect text in the Social Media Arabic Dialect Corpus (SMADC). *Proc. of the 3rd Workshop on Arabic Corpus Linguistics*, 2019, pp. 51–59.

Литература

1. Stone M.L. et al. BIg data for Media // Reuters Institute for the Study of Journalism. 2014. November.
2. Badaro G., Baly R., Hajj H., El-Hajj W., Shaban K.B., Habash N., Al-Sallab A., Hamdi A. A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations // ACM Transactions on Asian and Low-Resource Language Information Processing. 2019. V. 18. N 3. P. 27. https://doi.org/10.1145/3295662
3. Zaidan O.F., Callison-Burch C. Arabic dialect identification // Computational Linguistics. 2014. V. 40. N 1. P. 171–202. https://doi.org/10.1162/COLI_a_00169
4. Habash N.Y. Introduction to Arabic natural language processing // Synthesis Lectures on Human Language Technologies. 2010. V. 3. N 1. https://doi.org/10.2200/S00277ED1V01Y201008HLT010
5. Alnawas A., Arici N. The corpus based approach to sentiment analysis in modern standard Arabic and Arabic dialects: A literature review // Journal of Polytechnic-Politeknik Dergisi. 2018. V. 21. N 2. P. 461–470. https://doi.org/10.2339/politeknik.403975
6. Alshutayri A., Atwell E. Classifying Arabic dialect text in the Social Media Arabic Dialect Corpus (SMADC) // Proc. of the 3rd Workshop on Arabic Corpus Linguistics. 2019. P. 51–59.

7. Abo M.E.M., Raj R.G., Qazi A. A review on Arabic sentiment analysis: State-of-The-Art, taxonomy and open research challenges. *IEEE Access*, 2019, vol. 7, pp. 162008–162024. https://doi.org/10.1109/ACCESS.2019.2951530
8. Kumar A., Jaiswal A. Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency and Computation: Practice and Experience*, 2020, vol. 32, no. 1, pp. e5107. https://doi.org/10.1002/cpe.5107
9. Cieliebak M., Deriu J., Egger D., Uzdilli F. A Twitter corpus and benchmark resources for German sentiment analysis. *Proc. of the 5th International Workshop on Natural Language Processing for Social Media, SocialNLP*, 2017, pp. 45–51. https://doi.org/10.18653/v1/W17-1106
10. Nabil M., Aly M., Atiya A.F. ASTD: Arabic sentiment tweets dataset. *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 2515–2519. https://doi.org/10.18653/v1/D15-1299
11. Alahmary R.M., Al-Dossari H.Z., Emam A.Z. Sentiment analysis of Saudi dialect using deep learning techniques. *Proc. of the 18th International Conference on Electronics, Information, and Communication (ICEIC)*, 2019, pp. 8706408. https://doi.org/10.23919/ELINFOCOM.2019.8706408
12. Alnawas A., Arici N. Sentiment analysis of Iraqi Arabic dialect on Facebook based on distributed representations of documents. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2019, vol. 18, no. 3, pp. a20. https://doi.org/10.1145/3278605
13. Kwaik K.A., Saad M., Chatzikyriakidis S., Dobnik S. Shami: A corpus of levantine Arabic dialects. *Proc. of the 11th International Conference on Language Resources and Evaluation. (LREC-2018)*, 2019, pp. 3645–3652.
14. Oussous A., Lahcen A.A., Belfkih S. Impact of text pre-processing and ensemble learning on Arabic sentiment analysis. *ACM International Conference Proceeding Series*, 2019, vol. part F148154, pp. 65. https://doi.org/10.1145/3320326.3320399
15. El Abdouli A., Hassouni L., Anoun H. Sentiment analysis of moroccan tweets using naive bayes algorithm. *International Journal of Computer Science and Information Security*, 2017, vol. 15, no. 12, pp. 191–200.
16. Bouazizi M., Ohtsuki T. Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter. *Proc. of the IEEE International Conference on Communications (ICC)*, 2016, pp. 7511392. https://doi.org/10.1109/ICC.2016.7511392
17. Altamimi M., Alruwaili O., Teahan W.J. BTAC: A twitter corpus for Arabic dialect identification. *Proc. of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018)*, 2018, pp. 5.
18. Al-Yasiri E.K., Al-Azawei A. Improving Arabic sentiment analysis on social media: A comparative study on applying different pre-processing techniques. *Compusoft*, 2019, vol. 8, no. 6, pp. 3150–3157.
19. Platt J.C. Sequential Minimal Optimization: A fast algorithm for training support vector machines. *CiteSeerX*, 1998, vol. 10, no. 1.43, pp. 4376.

7. Abo M.E.M., Raj R.G., Qazi A. A review on Arabic sentiment analysis: State-of-The-Art, taxonomy and open research challenges // IEEE Access. 2019. V. 7. P. 162008–162024. https://doi.org/10.1109/ACCESS.2019.2951530
8. Kumar A., Jaiswal A. Systematic literature review of sentiment analysis on Twitter using soft computing techniques // Concurrency and Computation: Practice and Experience. 2020. V. 32. N 1. P. e5107. https://doi.org/10.1002/cpe.5107
9. Cieliebak M., Deriu J., Egger D., Uzdilli F. A Twitter corpus and benchmark resources for German sentiment analysis // Proc. of the 5th International Workshop on Natural Language Processing for Social Media (SocialNLP). 2017. P. 45–51. https://doi.org/10.18653/v1/W17-1106
10. Nabil M., Aly M., Atiya A.F. ASTD: Arabic sentiment tweets dataset // Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2015. P. 2515–2519. https://doi.org/10.18653/v1/D15-1299
11. Alahmary R.M., Al-Dossari H.Z., Emam A.Z. Sentiment analysis of Saudi dialect using deep learning techniques // Proc. of the 18th International Conference on Electronics, Information, and Communication (ICEIC). 2019. P. 8706408. https://doi.org/10.23919/ELINFOCOM.2019.8706408
12. Alnawas A., Arici N. Sentiment analysis of Iraqi Arabic dialect on Facebook based on distributed representations of documents // ACM Transactions on Asian and Low-Resource Language Information Processing. 2019. V. 18. N 3. P. a20. https://doi.org/10.1145/3278605
13. Kwaik K.A., Saad M., Chatzikyriakidis S., Dobnik S. Shami: A corpus of levantine Arabic dialects // Proc. of the 11th International Conference on Language Resources and Evaluation. (LREC-2018). 2019. P. 3645–3652.
14. Oussous A., Lahcen A.A., Belfkih S. Impact of text pre-processing and ensemble learning on Arabic sentiment analysis // ACM International Conference Proceeding Series. 2019. V. Part F148154. P. 65. https://doi.org/10.1145/3320326.3320399
15. El Abdouli A., Hassouni L., Anoun H. Sentiment analysis of moroccan tweets using naive bayes algorithm // International Journal of Computer Science and Information Security. 2017. V. 15. N 12. P. 191–200.
16. Bouazizi M., Ohtsuki T. Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter // Proc. of the IEEE International Conference on Communications (ICC). 2016. P. 7511392. https://doi.org/10.1109/ICC.2016.7511392
17. Altamimi M., Alruwaili O., Teahan W.J. BTAC: A twitter corpus for Arabic dialect identification // Proc. of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018). 2018. P. 5.
18. Al-Yasiri E.K., Al-Azawei A. Improving Arabic sentiment analysis on social media: A comparative study on applying different pre-processing techniques // Compusoft. 2019. V. 8. N 6. P. 3150–3157.
19. Platt J.C. Sequential Minimal Optimization: A fast algorithm for training support vector machines // CiteSeerX. 1998. V. 10. N 1.43. P. 4376.

Authors

Mohammed M. Hassoun Al-Jawad — PhD, Academic Staff, Lecturer, University of Kerbala, College of Computer Science and Information Technology, Karbala, 51001, Iraq, SC 57216978290, https://orcid.org/0000-0001-6750-0294, mohammedm@uokerbala.edu.iq

Hasaneun Alharbi — Academic Staff, Lecturer, University of Babylon, IT College, Babylon, 51002, Iraq, https://orcid.org/0000-0003-2577-278X, hasanein.alharbi@uobabylon.edu.iq

Ahmed Almukhtar — PhD, Academic Staff, Lecturer, University of Kerbala, College of Computer Science and Information Technology, Karbala, 51001, Iraq, https://orcid.org/0000-0002-4585-3977, ahmed.ahmukhtar@uokerbala.edu.iq

Anwar A. Alnawas — PhD, Head of Department, Southern Technical University, Nasiriyah Technical Institute, Nasiriyah, Iraq, SC 57205432645, https://orcid.org/0000-0001-9181-9377, anwar.alnawas@stu.edu.iq

Received 14.12.2021
Approved after reviewing 28.02.2022
Accepted 31.03.2022

Авторы

Хассун Аль-Джавад Мохаммед М. — PhD, научный работник, преподаватель, Университет Кербелы, Колледж компьютерных наук и информационных технологий, Кербела, 56001, Ирак, SC 57216978290, https://orcid.org/0000-0001-6750-0294, mohammedm@uokerbala.edu.iq

Альхарби Хасанейн — научный сотрудник, преподаватель, Университет Вавилона, Колледж информационных технологий, Вавилон, 51002, Ирак, https://orcid.org/0000-0003-2577-278X, Hasanein.alharbi@uobabylon.edu.iq

Альмухтар Ахмед Ф. — PhD, научный работник, преподаватель, Университет Кербелы, Колледж компьютерных наук и информационных технологий, Кербела, 56001, Ирак, https://orcid.org/0000-0002-4585-3977, ahmed.ahmukhtar@uokerbala.edu.iq

Алнавас Анвар Аднан — PhD, начальник департамента, Южный технический университет, Технический институт в Насири, Насири, Ирак, SC 57205432645, https://orcid.org/0000-0001-9181-9377, anwar.alnawas@stu.edu.iq

Статья поступила в редакцию 14.12.2021
Одобрена после рецензирования 28.02.2022
Принята к печати 31.03.2022