Data in Brief 42 (2022) 108089

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Medical dataset classification for Kurdish short text over social media

Ari M. Saeed[a,*], Shnya R. Hussein[a], Chro M. Ali[a], Tarik A. Rashid[b]

[a] Computer Science Department, University of Halabja, KRG, Halabja, Kurdistan, Iraq
[b] Computer Science and Engineering Department, University of Kurdistan-Hawlêr, KRG, Erbil, Kurdistan, Iraq

ARTICLE INFO

Article history:
Received 17 January 2022
Revised 14 March 2022
Accepted 16 March 2022
Available online 23 March 2022

Dataset link: Medical Sentiment Analysis Dataset for Kurdish Short Text over Social Media (Original data)

Keywords:
Machine learning
Medical text classification
Kurdish short text
Text pre-processing

ABSTRACT

The Facebook application is used as a resource for collecting the comments of this dataset, The dataset consists of 6756 comments to create a Medical Kurdish Dataset (MKD). The samples are comments of users, which are gathered from different posts of pages (Medical, News, Economy, Education, and Sport). Six steps as a preprocessing technique are performed on the raw dataset to clean and remove noise in the comments by replacing characters. The comments (short text) are labeled for positive class (medical comment) and negative class (non-medical comment) as text classification. The percentage ratio of the negative class is 55% while the positive class is 45%.

© 2022 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

* Corresponding author.
E-mail address: ari.said@uoh.edu.iq (A.M. Saeed).

https://doi.org/10.1016/j.dib.2022.108089
2352-3409/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Specifications Table

Subject	Applied Machine Learning
Specific subject area	Medical dataset classification for Kurdish short text over social media
Type of data	Text Figure Table
How the data were acquired	Facepager application is used for collecting the comments after configuring.
Data format	Raw
Description of data collection	Each post is separated accurately to describe the type of class (medical or non-medical), then the link of the post is copied and pasted in the Facepager application for gathering the specified comments.
Data source location	Kurdish post link in Facebook Application
Data accessibility	+Repository name: Mendeley Data Data identification number: 10.17632/f2yfz4r9fr.1 Direct link to the dataset: https://dx.doi.org/10.17632/f2yfz4r9fr.1

Value of the Data

- This is an effort of collecting a dataset in the field of medical text classification for the Kurdish language. Moreover. It can be beneficial for supporting and modeling patient health systems, health policies, and regulations.
- The data is preprocessed and ready for implementation by those researchers and scholars who conduct research work on the Arabic Alphabet, such as Persian, Arabic, and Urdu.
- The dataset can be used with several preprocessing steps such as stemming and lemmatization.

1. Data Description

The Kurdish language is one of the languages of the Middle East that is used for speaking by Kurdish people. Central Kurdish (Sorani) and Kurmanji are two popular dialects of the Kurdish language [1,2]. In this project, the Sorani dialect is used for collecting the database. The texts of Sorani Kurdish are like Persian, Arabic, and Urdu that are written from right to left. The characters among those languages are almost like each other but sometimes have different Unicode as shown in Table 1. The number of Sorani Kurdish alphabets is 36 that is divided into vowels (ا, ە, و, ۆ, وو, ی, ێ) and consonants (ئ, ب, پ, ت, ج, چ, ح, خ, د, ر, ڕ, ز, ژ, س, ش, ع, غ, ف, ڤ, ق, ک, گ, ل, ڵ, م, ن, ه, (و, ی)) [3,4]. The character (و, ی) are used as vowels and constants based on the positions of the word, for example, the word (لوگ) (gull) means (Flower), the (و) is a vowel, while the word (وازی) (wazi) means (game), the (و) is constant, the word (یاری) (yari) means (play), and the first (ی) is constant, by contrast, the second one is a vowel. The Kurdish language is complex and has different scripts (no standard) for Sorani, dialect for example, in some sources (ك) is used instead of (ک) [3,4].

1.1. Data collection

In this era, the health of people is a serious subject that researchers work on it closely [5,6]. For this purpose, it is important to read humans' views over social media. In this work, the Facebook application is used as a social media for creating a proper MKD. Nevertheless, to say that for predicting the right sight of humans by using machines, a good resource (dataset) is necessary. As it is clear, there are so many channels, websites, and live posts that can be used for this purpose. The database in this work id consisted of 6756 samples, which are divided into two different classes (medical and non-medical). The samples were collected from various pages

Table 1
Alphabet similarities among (Kurdish, Arabic, Persian, Urdu) Languages.

NO.	Kurdish alphabetic	Arabic Language	Persian Language	Urdu Language
1	ئ	ا	آ or ا	آ or ا
2	ب	ب	ب	ب
3	پ	ت	پ	پ
4	ت	ث	ت	ت
5	ج	ج	ث	ٹ
6	چ	ح	ج	ث
7	ح	خ	چ	ج
8	خ	د	ح	چ
9	د	ذ	خ	ح
10	ر	ر	د	خ
11	ڕ	ز	ذ	د
12	ز	س	ر	ڈ
13	ژ	ش	ز	ذ
14	س	ص	ژ	ر
15	ش	ض	س	ڑ
16	ع	ط	ش	ز
17	غ	ظ	ص	ژ
18	ف	ع	ض	س
19	ڤ	غ	ط	ش
20	ق	ف	ظ	ص
21	ک	ق	ع	ض
22	گ	ك	غ	ط
23	ل	ل	ف	ظ
24	ڵ	م	ق	ع
25	م	ن	ک	غ
26	ن	ه	گ	ف
27	ه	و	ل	ق
28	و	ي	م	ک
29	ی	ء	ن	گ
30	ا		و	ل
31	ه		ه	م
32	و		ی	ن
33	ۆ			و
34	وو			ہ
35	ی			ی
36	ێ			ے

Table 2
Number and percentage of collected dataset.

Class	Field	No. of Samples	Percentage
Medical	Medical	3076	45%
Not	News	890	55%
Medical	Economy	720	
	Education	1140	
	Sport	930	

and different areas as shown in Table 2. The number of medical comments (positive class) is 3076 while the non-medical comments (negative class) are 3680.

1.2. Methodology

On social media, the data can be viewed in various types, such as image, video, text. In this work, the data set is collected from the text. Facebook application is used for collecting the comments of users. Some different tools and techniques can be utilized for collecting the comments,

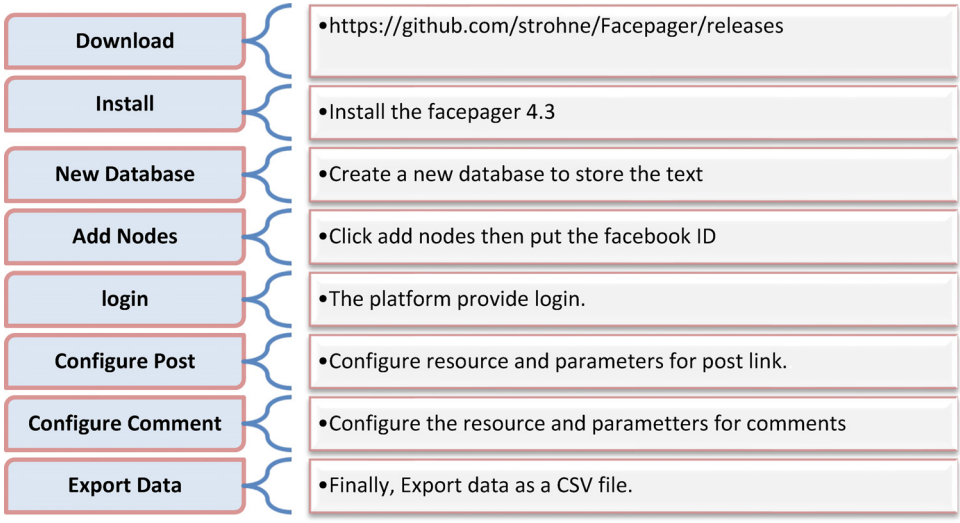

Fig. 1. Steps to dataset collection.

the Facepager tool is one of them that has been used for this reason [7]. The following steps should be followed for obtaining the data as shown below in Fig. 1.

As shown in Fig. 1, the first step is downloading the Facepager software for collecting the comments. The second step is locating and installing the files. The third step is to open the software and create a new database for saving the text file in (.db) format. The fourth step is adding nodes and putting the Facebook ID of the specified link after converting it over the internet. The fifth step is to log into Facebook via the Facepager tool. The sixth important step is configuring resources as (/<page-id>/posts) and parameters filed as (message) and specifying a start date and end date to fetch posts between those specific dates as shown in Fig. 2.

The seventh step is configuring a tool for fetching comments by clicking on a specific post and configuring resources as (/<post-id>/comments) and parameters filed as (message) as shown in Fig. 3.

The last and final step is exporting the comments as a CSV file as shown in Fig. 4.

1.3. Data set preprocessing

Preprocessing is one of the most important challenges for decreasing the noise on social media. Due to Kurdish users on the Facebook application using different Unicode to share their opinion and views. This causes a big issue for recognizing text and makes different characters shape. Using different scripts also increases the number of features (word) [1,4,8,9]. Accordingly, python language is used to create a new tool for implementing the below steps on the text as shown in Table 3:

1. Removing noise (URL, User mentions, and Hashtag) on social media users will provide extra information for their relatives and friends by using URL, mentions (@user name), and hashtags (#special topic) that information are helpful for users but it is noise for the machine. It has to be removed.
2. Replacing elongated characters: users on social media sometimes use elongated words purposely to emphases about special things, such as (چیییییییئ) (chiye), which means (Whaaaaaaat), which should be replaced with a base word (چیئ) (chiye), which means (what).

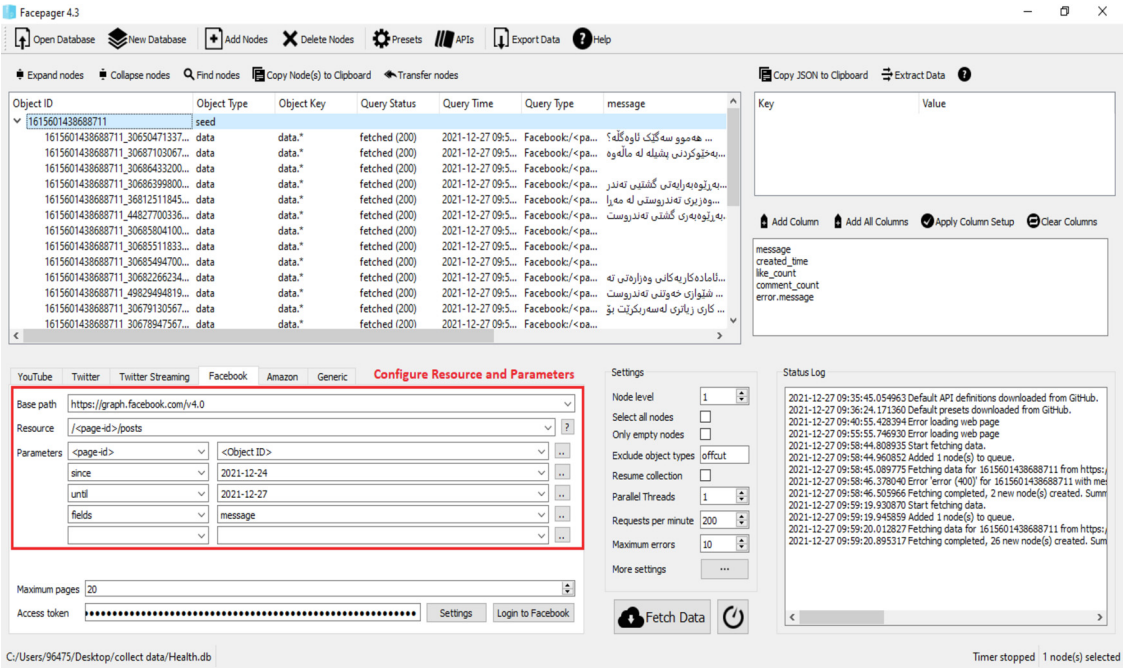

Fig. 2. Configure facepager posts.

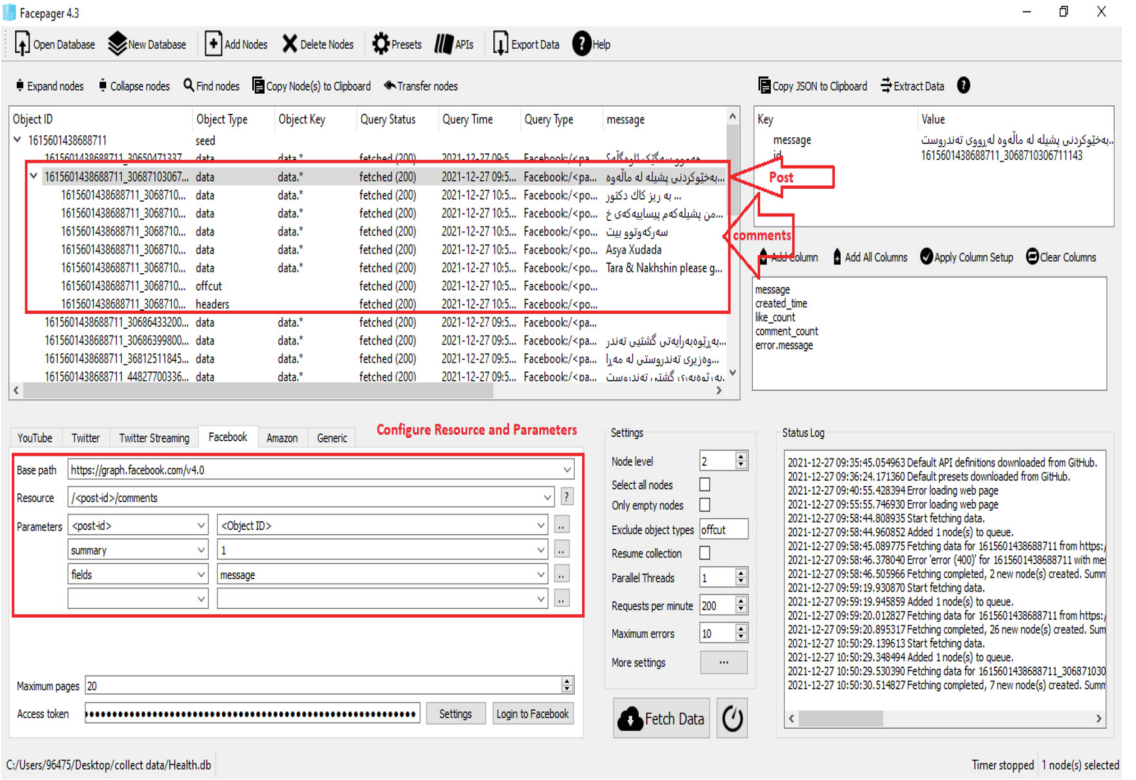

Fig. 3. Configure facepager comments.

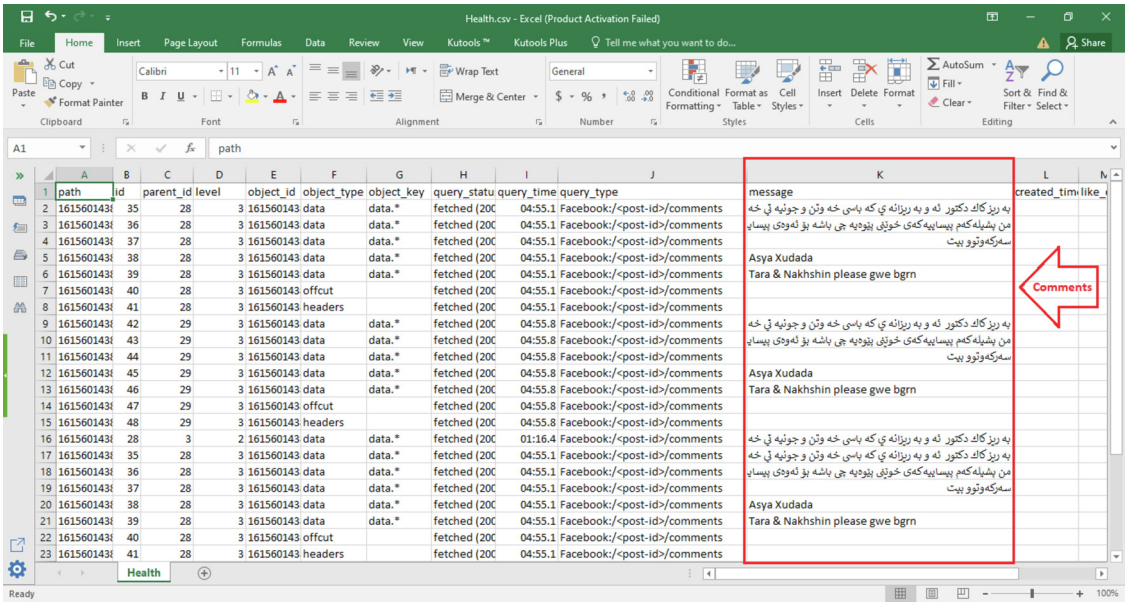

Fig. 4. Configure facepager comments.

Table 3
Preprocessing steps.

<table>
<tr><th>NO</th><th>Preprocessing steps</th><th>Natural comments with the Arabic alphabet</th><th>Natural comments with Latin alphabet</th><th>Natural comments in English</th><th>Preprocessing comments with the Arabic alphabet</th><th>Preprocessing comments with Latin alphabet</th><th>Preprocessing comments in English</th></tr>
<tr><td>1.</td><td>Removing noise (URL, User mentions and Hashtag):</td><td>دایز ایران|ەس|ەرۆتکد
یرۆپسپ .ە|ن|ەگ|ن|ەز
و یر|ەگر|ەتش|ەن
.واچ ینا|ک|ەیشۆخ|ەن
#Medical_Knowledge
یکشیزپ_یتسناز#</td><td>dktore| se|nariya ziyad zen|ge|ne|. psporî ne|şite|rge|rî û ne|xoşiye|kanî çaw. #Medical_Knowledge #zanistî_pzîşkî</td><td>Doctor Sanaria Zyad Zangan. Expert in eye surgery and disease. Medical_Knowledge #یکشیزپ_یتسناز</td><td>دایز ایران|ەس|ەرۆتکد
یرۆپسپ ە|ن|ەگ|ن|ەز
و یر|ەگر|ەتش|ەن
واچ ینا|ک|ەیشۆخ|ەن</td><td>dktore| se|nariya ziyad ze|nge|ne| psporî ne|şite|rge|rî û ne|xoşiye|kanî çaw</td><td>Doctor Sanaria Zyad Zangan expert in eye surgery and disease</td></tr>
<tr><td>2.</td><td>Replacing elongated characters</td><td>ةيةه مكيكشوخ روتكد والس
ةواركيب يرةس اسيداح ةل
يةليبليب وودرةه اتسيا
ينيلوجان يشاب ةب يواج
يييييييييياكت
ببببباوج ةييييي
راكوه بببب
ةييييييييج</td><td>slaw dktur xuşkîkm heye le ḧadîsa serî bîkrawe aysta herdû bîlbîleyi cawî be başî naculînî tkayîyîyîyîyîyîyîye cwabibbibbibbibb hukar cîyîyîyîye</td><td>Hi doctor, my sister was hurt in an accident now her cornea cannot move perfectly pleaaaaaaaaaaase answeeeeeer and whaaaaaaat is the cause</td><td>ەیەه مکیکشوخ روتکد والس
ەوارکیب یرەس اسیداح ەل
یەلیبلیب وودرەه اتسیا
ینیلوجان یشاب ەب یواج
یج راکوه باوج ەیاکت</td><td>slaw dktur xuşkîkm heye le ḧadîsa serî bîkrawe aysta herdû bîlbîleyi cawî be başî naculînî tkaye cwab hukar ciye</td><td>Hi doctor, my sister has hurt in an accident now her cornea cannot move perfectly please answer and what is the cause</td></tr>
<tr><td>3.</td><td>Incorrect spelling and grammars</td><td>وەئ راج دنەچرەه اڵڵاشام
نیزنەب یساب ەیاراباک
تێبەد رتنارگ تاکب</td><td>maşallla herçend car ew kabraye basî benzîn bikat grantir debêt</td><td>Allah has willed it, any time that man talks about gasoline, it will be more expensive</td><td>وەئ راج دنەچرەه ەللااشام
نیزنەب یساب ەیاراباک
تێبەد رتنارگ تاکب</td><td>maşaalle herçend car ew kabraye basî benzîn bikat grantir debêt</td><td>Allah has willed it, how any time that man talks about gasoline, it will be more expensive</td></tr>
<tr><td>4.</td><td>Removing punctuation</td><td>ەتاڵو مەئ تێبۆت یەسقەب
!!!!ەفیرەش یماش</td><td>beqseyi tobêt em wullate şamî şerîfe!!!!</td><td>According to your speech, this country is peaceful!!!!</td><td>ەتاڵو مەئ تێبۆت یەسقەب
ەفیرەش یماش</td><td>beqseyi tobêt em wullate şamî şerîfe</td><td>According to your speech, this country is peaceful</td></tr>
<tr><td>5.</td><td>Removing numbers:</td><td>يوم ەڵاس ٣٠ نم ەروتكد
ەتيد ر ەە منيد ر ەد ردينم
وە</td><td>dkturh min 30 salh mwî rdînm dh r dînm hh r dîth wh</td><td>Doctor, it is about 30 years, I have been pulling out beard hair, yet it grows back</td><td>ەد منیدر یوم ەڵاس نم ەروتکد
وە ەتید ر ەە منید ر</td><td>dkturh min salh mwî rdînm dh r dînm hh r dîth wh</td><td>Doctor, it is about 30 years, I have been pulling out beard hair, yet it grows back</td></tr>
<tr><td>6.</td><td>Replacing characters</td><td>یرۆز یتةيساسح نم والس
یتةبياتةب ةيةه مواچ
راهةب یزرةوةل
تیروخةد وةتيبةدروس
یتةرةنب یرةسةراچ ايائ
ةيةه ياين رینامیی
؟تةمحةزيب</td><td>slaw min ḧsasiyetî zۆrî çawm heye betaybetî lewerizî behar surdebîtewew dexurît aya çareserî bineretî heye yan rînmayî bîzeḧmet?</td><td>Hi, I have an eye rash, especially in the Spring season and it became red and itchy, is there any essential treatment or any advice.</td><td>یرۆز یتەیساسح نم والس
یتەبیاتەب ەیەه مواچ
راهەب یزرەوەل
تیروخەد وەوەتیبەدروس
یتەرەنب یرەسەراچ ایائ
ەیەه یان رینامیی
تەمحەزیب</td><td>slaw min ḧsasiyetî zorî çawm heye betaybetî lewerizî behar surdebîtewew dexurît aya çareserî bineretî heye yan rînmayî bîzeḧmet</td><td>Hi, I have an eye rash, especially in the Spring season and it became red and itchy, is there any essential treatment or any advice.</td></tr>
</table>

3. Incorrect spelling and grammar: sometimes it is easy for users to correct the misspelling and grammar but machines cannot understand and it is challenging. These three words (ماشاللّٰ, ماشاء اللە) (masha allh), which means (Allah has willed it) used as a misspelling instead the correct word (ماشااللە), which means (Allah has willed it).
4. Removing punctuations: users on social media use them to express special emotions, which are easy for a human to recognize. Nevertheless, those punctuations make usefulness for machines to translate and become inefficient for text classification. These punctuations are removed (!@#$%^&*()[]{};:,،./<>؛ ؟٠☒ ٠٪!! ‹؛‹‹‹؛ £,’~ٍ——————‹؛‘ٌَُّ\ "'''+_=-~`|?؟”,*ًٌٍَُِّْٰٓ ٔ ء؛ ﷲ ٓ ٕ ...©)·
5. Removing numbers: numbers increase the number of features in text datasets on social media and they are not helpful for the machine to understand. However. Kurdish users use different types of numbers, such as (English, Arabic, and Kurdish) numbers as shown.
6. Replacing characters: due to the Kurdish language using the same script of Arabic language for some characters and some users on social media use Arabic Keyboard for writing. This has become an issue for matching and selecting features. However, the issue has been solved by replacing the character as shown below:
   a- ‘ي’ with ‘ی’.
   b- ‘ك’ with ‘ک’.
   c- ‘ۆ’ with ‘ۆ’.
   d- ‘ة’ with ‘ە’.
   e- If the word ends with ‘ئ’ replace with ‘ئ’.
   f- If the word ends with ‘ە‌’ (\u 647\u 200C), then replace it with ‘ە’ (\u0647) as shown in the same shape of characters but different Unicode.

1.4. Dataset labeling

After collecting the dataset, another important step is labeling the samples. For this purpose, three annotators read the samples accurately and manually labeled the unlabeled samples for two classes (medical and non-medical). This process needs a huge effort and consumes time. For labeling each sample, the annotator annotates the sample based on some special words in the medical domain and the meaning of each sentence as shown in Table 4:

Table 4
Labeling of comments.

NO.	Samples (Comments) in the Arabic alphabet	Samples (Comments) in Latin alphabet	Samples (Comments) in English	Classes
1	سلاو من دە مو جاو م خالی قاوەیی یە جی باشە بیزەحمە ت	slaw min de mu caw m xalî qaweyi ye cî başe bîze ḣme t	Hello, my face has a brown spot, please what is good for me to do	medical
2	پۆلێک ژمارەی تە لە بە کانی ٦٠ بۆ سە رەوە بێ ئە وانە ەە مووی بێ مانان وەهیچ سوودی نیە!	polêk jmareyi te lh bih kanî ٦٠ bo sh rewe bê e wanh he muwî bê manan wehîç sûdî niye!	A class that has several students, more than 60, that is no sense and does’t have any benefit	Not medical
3	سلاو دكتور مناڵە كە م ماوە يي سكی ی ريش هويب یجنیل تيج ئە قوتو ئە خوات	slaw dktur minalh kh m mawh yh skî ih cît lînçî biyuh şîr yi qtu ih xwat	Hi doctor, my baby has diarrhea and viscidity and eats condensed milk	medical
4	ئەو ڕۆژەی بەخت دەرگایت لێ ئەکاتەوە بەڵام ئەقڵ دائەخاتەوە	ew rojeyi bext dergayt lê ekatewe belam eql dayexatewe	The day that opens the luck for you, yet, it closes mind	Not medical

Ethics Statement

All omments in the dataset belong to users in the Facebook application and it is scrapped. The data has been distributed over Facebook and thus, it has been collected and labeled. Moreover, we confirm that all the data is insensitive and anonymized data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Data Availability

Medical Sentiment Analysis Dataset for Kurdish Short Text over Social Media (Original data) (Mendeley Data).

CRediT Author Statement

Ari M. Saeed: Supervision, Data curation, Conceptualization, Methodology, Visualization, Project administration, Funding acquisition, Writing – original draft, Writing – review & editing; **Shnya R. Hussein:** Software, Formal analysis, Investigation, Resources; **Chro M. Ali:** Software, Formal analysis, Investigation, Resources; **Tarik A. Rashid:** Methodology, Supervision, Validation, Writing – review & editing.

Acknowledgments

The authors would like to thank the University of Halabja and the University of Kurdistan Hewler for providing all the facilities needed for conducting this research work.

References

[1] A. Saeed, T. Rashid, A. Mustafa, R. Agha, A. Shamsaldin, N. Al-Salihi, An evaluation of Reber stemmer with longest match stemmer technique in Kurdish Sorani text classification, Iran J. Comput. Sci. 1 (2018) 99–107, doi:10.1007/s42044-018-0007-4.
[2] T.A. Rashid, A.M. Mustafa, A. Saeed, A robust categorization system for Kurdish Sorani text documents, Inf. Technol. J. 16 (2016) 27–34, doi:10.3923/itj.2017.27.34.
[3] T. Rashid, A. Mustafa, A. Saeed, Automatic Kurdish text classification using KDC 4007 dataset, in: advances in internetworking, data & web technologies. EIDWT 2017, Lecture Notes on Data Engineering and Communications Technologies, Springer, Cham, Wuhan, 2017.
[4] A. Saeed, T. Rashid, A. Mustafa, P. Fattah, B. Ismael, Improving Kurdish web mining through tree data structure and Porter's Stemmer algorithms, UKH J. Sci. Eng. 2 (2018) 48–54, doi:10.25079/ukhjse.v2n1y2018.pp48-54.
[5] R. Meena, V. hulasi Bai, Study on machine learning based social media and sentiment analysis for medical data applications, in: Proceedings of the Third International Conference on ISMAC (IoT in Social, Mobile, Analytics and Cloud) (ISMAC), IEEE, 2019.
[6] G. Saranya, G. Geetha, C. K, M. Meenakshi K, S. Karpagaselvi, Sentiment analysis of healthcare tweets using SVM classifier, in: Proceedings of the International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), IEEE, 2020.
[7] W. He, S. Zha, L. Li, Social media competitive analysis and text mining: a case study in the pizza industry, Int. J. Inf. Manag. 33 (2013) 464–472, doi:10.1016/j.ijinfomgt.2013.01.001.
[8] R. Ahmed, T. Rashid, P. Fatah, A. Alsadoon, S. Mirjalili, An extensive dataset of handwritten central Kurdish isolated characters, Data Brief 39 (2021) 107479, doi:10.1016/j.dib.2021.107479.
[9] U. Naseem, I. Razzak, P. Eklund, A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter, Multimed. Tools Appl. 80 (2020) 35239–35266, doi:10.1007/s11042-020-10082-6.