Medical Dataset Classification for Kurdish Short Text over Social Media

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

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

ari.said@uoh.edu.iq

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%.

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

1- 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.

2- 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.

3- The dataset can be used with several preprocessing steps such as stemming and lemmatization.

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 (ئ, ب, پ, ت, ج, چ, ح, خ, د, ر, ڕ, ز, ژ, س, ش, ع, غ‌ف, ڤ, ق, ک, گ, ل, ڵ, م, ن, ه, (و, ى)) and consonants (ئ, ب, پ, ت, ج, چ, ح, خ, د, ر, ڕ, ز, ژ, س, ش, ع, غ‌ف, ڤ, ق, ک, گ, ل, ڵ, م, ن, ه, (و, ى)) [3, 4]. The character (و, ى) are used as vowels and constants based on the positions of the word, for example, the word (گول) (gull) means (Flower), the (و) is a vowel, while the word (وازى) (wazi) means (game), the (و) is constant, the word (بازى) (yari) means (play), and the first (ى) 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 (ک) is used instead of (ک) [3, 4].

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

| NO. | Kurdish alphabetic | Arabic Language | Persian Language | Urdu Language |
|-----|-------------------|----------------|-----------------|--------------|
| 1   | ئ | ا | ا | ا | ئ |
| 2   | ب | ب | ب | ب | ب |
| 3   | پ | ت | ت | ت | ت |
| 4   | ت | ت | ت | ت | ت |
| 5   | ج | ج | ج | ج | ج |
| 6   | ج | ج | ج | ج | ج |
Data collection

In this era, the health of people is a serious subject that researchers work on it closely [7, 8]. 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 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.
### Table 2 Number and Percentage of Collected Dataset

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

### 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, the Facepager tool is one of them that has been used for this reason [9]. The following steps should be followed for obtaining the data as shown below in Figure 1.

Figure 1 Steps to Dataset Collection

- **Download**
  - https://github.com/strohne/Facepager/releases
- **Install**
  - Install the facepager 4.3
- **New Database**
  - Create a new database to store the text
- **Add Nodes**
  - Click add nodes then put the facebook ID
- **login**
  - The platform provide login.
- **Configure Post**
  - Configure resource and parameters for post link.
- **Configure Comment**
  - Configure the resource and parameters for comments
- **Export Data**
  - Finally, Export data as a CSV file.

As shown in Figure 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 Figure 2.

![Figure 2 Configure Facepager Posts](image1.png)

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 Figure 3.

![Figure 3 Configure Facepager Comments](image2.png)

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

![Figure 4 Export as CSV](image3.png)
Figure 4 Configure Facepager Comments

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, 5, and 10]. 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 (جيبسية) (chiye), which means (Whaaaaaat), which should be replaced with a base word (جي) (chiye), which means (what).
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 (مашاشلا) (masha allh), which means (Allah has willed it) used as a misspelling instead the correct word (مашاشا) (masha), 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 (!@#$%^&*()[]{]};; /<>+=•••!!"""£$%^\____+~`♥♫®).
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. 'ؤ' with 'و'
   d. 'ئ' with 'ی'
   e. If the word ends with 'ي' replace with 'ی'
   f. If the word ends with 'و' (u 647u 200C), then replace it with 'و' (u0647) as shown in the same shape of characters but different Unicode.

### Table 3 Preprocessing Steps

| NO  | Preprocessing Steps          | Natural Comments with the Arabic alphabet | Natural Comments with Latin alphabet | Natural Comments in English | Preprocessing Comments with the Arabic alphabet | Preprocessing Comments with Latin alphabet | Preprocessing Comments in English |
|-----|------------------------------|-------------------------------------------|-------------------------------------|-----------------------------|-----------------------------------------------|-------------------------------------------|---------------------------------|
| 1   | Removing noise (URL, User mentions and Hashtag): | دکتوردی سمانیا و زاد زەنگەنە، پەسپۆری و نەخۆشەکەیان جاو Medical_Knowledge حکیم عەیینەیەیەن | کەرەن وەئەوەیەنە و نەخۆشەکەیان جاو Medical_Knowledge حکیم عەیینەیەن | Doctor Sanaria Ziad Zangan. Expert in eye surgery and disease. Medical_Knowledge حکیم عەیینەیەن | دکتوردی سمانیا و زاد زەنگەنە، پەسپۆری و نەخۆشەکەیان جاو Medical_Knowledge حکیم عەیینەیەن | Doctor Sanaria Ziad Zangan expert in eye surgery and disease |
| 2   | Replacing elongated characters | سلاو دکتور خوشکێکی وەئەوەیەنە، جەوەیەنەدا بە ژواکەیەنەدا بەبەنەیەنەدا بەنەیەنەدا بەتەکێنییەنێیەنی | کەرەن وەئەوەیەنە و نەخۆشەکەیان جاو Medical_Knowledge حکیم عەیینەیەن | Hi doctor, my sister was hurt in an accident now her cornea cannot move perfectly | slab dktur xusëkikm heye le hadisa seri bikrawe aysta herdû bîblîyeyî cawî be başi nacûlînî tkaye cwbab hukar ciye | slab dktur xusëkikm heye le hadisa seri bikrawe aysta herdû bîblîyeyî cawî be başi nacûlînî tkaye cwbab hukar ciye | Hi doctor, my sister has hurt in an accident now her cornea cannot move perfectly please answer and what is the cause |
| Incorrect spelling and grammars | ماماًلا همچندن چار کارایی بسیار نیاز به زنی بکت گرینت دیبته | Allah has will ed it, any time that man talks about gas olene, it will be more expensive | ماماًلا همچندن چار کارایی بسیار نیاز به زنی بکت گرینت دیبته | Allah has will ed it, how any time that man talks about gasoline, it will be more expensive |
|--------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| Removing punctuation | بهسیای توبته یم وولاشه شامی شریفه!!! | According to your speech, this country is peaceful!!! | بهسیای توبته یم وولاشه شامی شریفه!!! | According to your speech, this country is peaceful!!! |
| Removing numbers: | دکتوه من ۳۰ ساله می رنیم دلاً هکی ر دیته و ه | دکتوه من ساله موی دیمل دیه ر دیته و ه | دکتوه من ساله موی دیمل دیه ر دیته و ه | Doctor, it is about 30 years, I have been pulling out beard hair, yet it grows back |
| Replacing characters | سلاو من حساسیتی نزوری چاوم همچندکه بنا به این سوال دول حیرت دیه سوکوئینیو دهورینی تاً خاف نیزیه یهای بیز چمته؟ | 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. | سلاو من حساسیتی نزوری چاوم همچندکه بنا به این سوال دول حیرت دیه سوکوئینیو دهورینی تاً خاف نیزیه یهای بیز چمته؟ | 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. |
|                | با کیفیت بیز چمته | 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. | با کیفیت بیز چمته | 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. |
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 دەمەوە جەوا بەنەی قاوەیە یەنە | slaw min de mu caw m xalî qawey yi ye cî bâse bîze êmê t | Hello, my face has a brown spot, please what is good for me to do | medical       |
| 2   | پۆڵەک ژمەرەی تە لە بە گالێ ١٠ یۆ سە رەوە | polêk jmareyi te lh bih kanî ١٠ bo sh rewe bê ë wanh he muwî bê manan wehiç 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 دەکتوڕ مەنالەکە مەوەیە سکی نە | slaw dktur minalh kh m mawh yh skî ih cît lîncî biyuh śîr yî qtu ih xwat | Hi doctor, my baby has diarrhea and viscidity and eats condensed milk | medical       |
| 4   | نەر رۆژەیە مەخە دەرگایت لە نەکاتەوە بە لەم نەقفل دایەنەخانەوە | ew rojeyi bext dergayt lê ekatewê belam eql dayexatewe | The day that opens the luck for you, yet, it closes mind | Not medical   |
Ethics Statement

All comments 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.

CRediT Author Statement

Ari M. Saeed: Supervision, Data curation, Conceptualization, Methodology, Visualization, Project administration, Funding acquisition, Writing - Original Draft, Writing –review and editing; Shnya R. Hussein: Software, Formal analysis, Investigation, and Resources; Chro M. Ali: Software, Formal analysis, Investigation, and 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 Journal Of Computer Science. 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, Information Technology Journal. 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 Journal Of Science And Engineering. 2 (2018) 48-54. doi:10.25079/ukhjse.v2n1y2018.pp48-54.

[5] R. Ahmed, T. Rashid, P. Fatah, A. Alsadoon, S. Mirjalili, An extensive dataset of handwritten central Kurdish isolated characters, Data In Brief. 39 (2021) 107479. doi:10.1016/j.dib.2021.107479.
[6] B. Hassan, T. Rashid, S. Mirjalili, Performance evaluation results of evolutionary clustering algorithm star for clustering heterogeneous datasets, Data In Brief. 36 (2021) 107044. doi:10.1016/j.dib.2021.107044.

[7] R. Meena, V. hulasi Bai, Study on Machine learning based Social Media and Sentiment analysis for medical data applications, in: 2019 Third International Conference On I-SMAC (Iot In Social, Mobile, Analytics And Cloud) (I-SMAC), IEEE, 2019.

[8] G. Saranya, G. Geetha, C. K, M. Meenakshi K, S. Karpagaselvi, Sentiment analysis of healthcare Tweets using SVM Classifier, in: 2020 International Conference On Power, Energy, Control And Transmission Systems (ICPECTS), IEEE, 2020.

[9] W. He, S. Zha, L. Li, Social media competitive analysis and text mining: A case study in the pizza industry, International Journal Of Information Management. 33 (2013) 464-472. doi:10.1016/j.ijinfomgt.2013.01.001.

[10] 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, Multimedia Tools And Applications. 80 (2020) 35239-35266. doi:10.1007/s11042-020-10082-6.