# KMD: A New Kurdish Multilabel Emotional Dataset For the Kurdish Sorani Dialect

**Soran Badawi**
Charmo Center for Research, Training and Consultancy, Charmo University, KRG, Iraq
soran.sedeeq@charmouniversity.org

## Abstract

This paper presents a new Kurdish Multilabel Emotional Dataset (KMD) for the Kurdish Sorani dialect, which contains emotional labels for four different categories: Fear, Sadness, Joy, and Surprise. The dataset was collected using Twitter API, pre-processed, and manually labelled by three independent annotators. We also conducted experiments using classical machine learning classifiers, including Naive Bayes and Support Vector Machine (SVM), and deep learning models, BLSTM and BERT, to evaluate the efficiency of the dataset. Our results show that the multilingual BERT model outperforms the traditional machine learning classifiers in emotional labelling accuracy. The KMD dataset can be used for various natural language processing tasks, including sentiment analysis, emotion detection, and opinion mining, for the Kurdish Sorani dialect.

## 1 Introduction

Emotions play a significant role in human communication. The presence and significance of emotions can be found in every area of our existence. Our emotions impact the choices we make, how we interact with others, and our daily conduct (Erol et al., 2019). They even persist beyond our memories. As the quantity of text-based content that conveys emotions grows rapidly (e.g., microblog posts, blogs, and forums), there is a pressing demand and potential to create automated tools that can recognize and evaluate emotions expressed in written language. In many situations, machines must engage with or observe humans, and emotion recognition has numerous practical uses in such settings. For example, in an online learning platform, an automated tutor could give more effective feedback to a student by taking into account her level of motivation or frustration. Similarly, a car that has the capability to assist a driver could take action or sound an alert if it detects that the driver is fatigued or anxious (Kosti et al., 2019).

Therefore, recognizing emotions is an essential task for machines to understand human behaviour. Natural Language Processing (NLP) is a rapidly growing field that aims to enable machines to understand human language (Khurana et al., 2023). One of the essential components of NLP is the availability of labelled datasets that can be used for the training and evaluation of machine learning models. However, there is a lack of such labelled datasets for some languages, including the Kurdish Sorani dialect.

In this paper, We endeavour to construct an emotion detection system for the Kurdish language, utilizing the Sorani dialect. Our efforts involved creating the initial emotion-annotated dataset for the Kurdish language entirely annotated by human annotators. Additionally, we utilized both machine learning classifiers and deep learning models to train our dataset and documented the findings of our experiment. We also conducted an experiment on the KMD dataset using classical machine learning classifiers and deep learning BLSTM and BERT to evaluate the efficiency of the dataset. For the machine learning classifiers, Naive Bayes and Support Vector Machine (SVM) classifiers were implemented. These models were selected because of their wide usage in machine-learning workflows. The availability of the KMD dataset will enable researchers to develop and evaluate emotion recognition models for the Kurdish Sorani dialect. The dataset will also be useful for various NLP applications, such as sentiment analysis and opinion mining. In summary, this paper presents a significant contribution to the field of NLP and emotional analysis by introducing a new dataset for the Kurdish Sorani dialect.

The following sections comprise the remainder of this paper: Section 2 provides the related works, while Section 3 introduces the data collection methods. In Section 4, we detail our robust baselines and experiments. Sections 5 and 6 discuss the find-

ings. Lastly, we conclude the paper in Section 7.

## 2 Related Works

The Kurdish language belongs to the Indo-Iranian branch of the larger Indo-European language family and consists of 33 letters. It is relatively similar to Persian and is spoken by approximately 30-40 million people in Iran, Turkey, Iraq, and Syria (Badawi, 2023b). In Iraq, the Kurdish language is recognized as one of the official languages (Ahmadi et al., 2019). There are two main dialects of Kurdish: Central Kurdish (Sorani) and Northern Kurdish (Badawi, 2023c). However, several minor dialects, such as Gorani (Hawrami), are used by small communities in Iraq and Iran, and Zazaki spoken in Turkey (Buran, 2011).

There have been several efforts to create emotional datasets for different languages, such as English (Plaza del Arco et al., 2020), Chinese (Feng et al., 2022), Arabic (Al-Khatib and El-Beltagy, 2017), and Hindi (Singh et al., 2022). However, the efforts on the Kurdish language increasingly focused on building sentiment analysis datasets. Sentiment analysis is a subfield of NLP that aims to identify and extract opinions and emotions expressed in text. Several studies have focused on sentiment analysis for Kurdish Sorani. Awla and Veisi (Awlla and Veisi, 2022) built a sentiment analysis dataset from gathering Facebook comments. A total of 18,450 comments were extracted from 13 popular pages. After collecting the data, the authors performed preprocessing on the comments by removing noisy comments and those that were not written in the Central Kurdish language. Three annotators were assigned to annotate each comment: Positive, Negative, or Neutral. The work of Hameed, Ahmedi and Rezai (Hameed et al., 2023)is another example of dataset construction in the field of sentiment analysis. Using Twitter API, the authors were able to collect and annotate 1769 tweets. The labels include positive, negative, mixed, neutral, and none. Furthermore, several studies worked on building Kurdish datasets in the field of text classification. Currently, we are aware of only two annotated corpora. The first one is the medical corpus, which contains 6756 samples obtained from Facebook comments and is divided into medical and non-medical (Saeed et al., 2022). The second one is KDC-4007, comprising 4,007 text files categorized into eight groups: Sports, Religions, Arts, Economics, Education, Socials, Styles, and Health (Rashid et al., 2018). Notably, no datasets for detecting emotions in the Kurdish language are currently available. Finally, KNDH (Kurdish News Dataset Headlines) is a dataset which includes a collection of 50,000 news headlines, equally distributed among Health, Science, Social, Economic and Sports categories (Badawi et al., 2023).

There is a concerning scarcity of annotated datasets for emotion detection in the Kurdish language. Previous studies that have focused on dataset creation for the Kurdish language have primarily focused on sentiment analysis, which involves categorizing texts into positive, negative, or neutral categories based on the expressed sentiment. However, emotions are more complex than simply positive or negative sentiments. Our work aims to fill this gap by building a new emotional dataset for the Kurdish Sorani dialect that includes a wider range of emotions, such as fear, sadness, joy, and surprise. By including a broader range of emotions, we can gain a deeper understanding of the nuances of language and how emotions are expressed in different contexts. This will enable us to build more accurate and effective natural language processing models that can be used for a variety of applications, including sentiment analysis, emotion detection, and text classification.

## 3 Dataset Benchmark

### 3.1 Data collection

To gather data, either a software program or specialized libraries must be employed through coding. In this study, we utilized the latter method and leveraged the Twitter developer API to extract tweets while removing user identities to comply with Twitter's policies and ensure security. The dataset resulting from this process is freely available on the Mendeley repository, accessible via the*URL ,[1]. Figure 1 outlines the steps taken to construct this dataset.

Undoubtedly, raw data can be contaminated with noise. Online Kurdish data, for instance, often contains words from other languages, special characters, elongated letters, symbols, and irrelevant numbers. During the preprocessing phase, we removed all non-Kurdish characters from HTML links. In the second phase, special characters were examined. If a specific character is used to convey senti-

[1] https://data.mendeley.com/datasets/dntxt73dm6/1

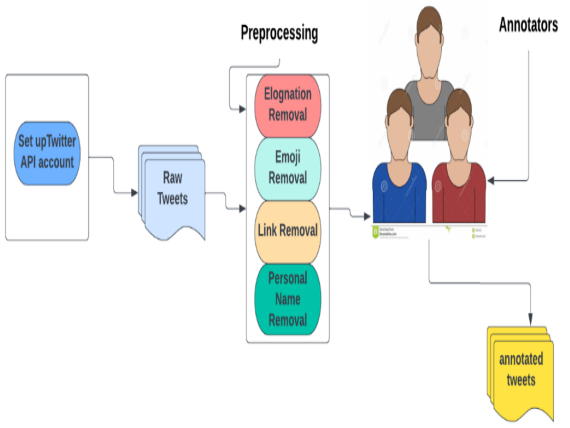


Figure 1: Data collection process

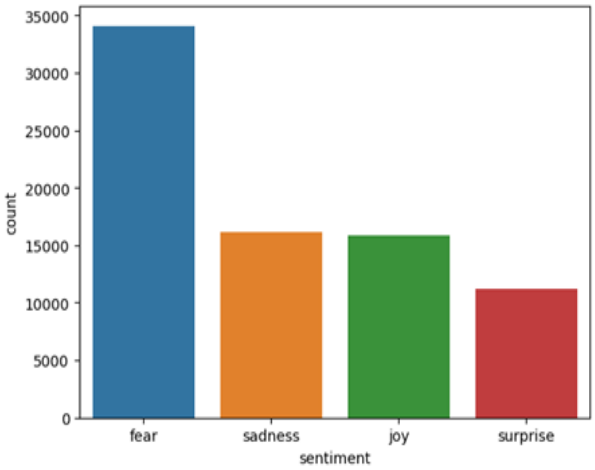


Figure 2: Class distribution in the dataset

| Annotated | R1 + R2 | R1 + R3 | R2 + R3 |
|---|---|---|---|
| kappa | 0.88 | 0.79 | 0.86 |

Table 1: KAPPA score among raters, R1,R2,R3 refer to Rater 1, Rater 2 and Rater 3 respectively

ment, it is left unchanged. Characters that have no meaning were removed from the text. Sometimes, numbers can express feelings or sentiments for the user, and therefore, they are included in the text.

### 3.2 Annotation Process

A group of three individuals who are knowledgeable in the Kurdish language were selected to annotate the corpus. These annotators have an educational background in the Kurdish language. To assist them in the annotation process, the annotators were provided with instructions, which included the following:

1. The annotators were required to determine the label of each texts.

2. The corpus was updated with the most noteworthy annotations for each piece of text, based on the number of votes it received.

The entire dataset was given to the annotators. The degree of agreement between the annotators was measured using Kappa coefficients (Cao et al., 2016). The Kappa coefficients demonstrated a strong level of agreement during the sentiment annotation process, ranging from 0.79 to 0.88, as presented in Table 1. These numbers are within an acceptable range according to previous studies.

### 3.3 Dataset Statistics

The whole dataset consists of 77270 texts distributed among fear, sadness, joy and surprise labels as depicted in Figure 3.

Figure 2 shows that fear was the most commonly existing emotion in the dataset, as it had the highest frequency count of 34085. The second most common emotion was sadness, with a frequency count of 16099. Joy was reported less frequently than both fear and sadness, with a frequency count of 15857. On the other hand, surprise was the least frequently reported emotion with a frequency count of 11204.

## 4 Experiment

In this paper, we aim to conduct an experiment on the KMD dataset using both classical machine classifiers and deep learning BLSTM and BERT to evaluate the efficiency of the dataset. For the machine learning classifiers, Naive Bayes and Support Vector Machine (SVM) classifiers were implemented. These models were selected because of their wide usage in machine-learning workflows.

The Naive Bayes algorithm is a probabilistic algorithm that is commonly used for text classification tasks (Abbas et al., 2019). Specifically, the Multinomial Naive Bayes variant of the algorithm is often used in natural language processing applications. To implement the Naive Bayes classifier in the paper, the authors used the Scikit-Learn library in Python. The library provides an implementation of the Multinomial Naive Bayes algorithm that is easy to use and has good performance. We first preprocessed the data by filtering out noisy comments and non-Central Kurdish letters. Then, the data

was split into training and testing sets with a ratio of 80:20. The training set was used to train the classifier, while the testing set was used to evaluate its performance. The next step was to convert the text data into numerical feature vectors that can be used as input to the classifier. We used a technique called CountVectorizer to represent the text data as feature vectors. CountVectorizer is a commonly used technique in text classification that converts text into a matrix of token counts. Once the data was preprocessed and converted into feature vectors using CountVectorizer, We trained the Naive Bayes classifier on the training set. The trained classifier was then evaluated on the test set to measure its accuracy in predicting the emotional sentiment of the tweets.

SVM (Support Vector Machine) is a popular machine-learning algorithm used for classification tasks. It finds the best hyperplane that separates the data into different classes by maximizing the margin between the hyperplane and the nearest data points (Cervantes et al., 2020). In this paper, We used the SVM algorithm as one of the machine learning classifiers to evaluate the efficiency of the KMD dataset. The Scikit-learn library in Python was used to implement the SVM algorithm.Overall, the SVM classifier was implemented using the Scikit-learn library in Python.

Throughout the experimentation phase, we selected BLSTM and BERT as our preferred options. BLSTM, a neural network type, excels in handling sequential data, encompassing both time series and text data. It comprises of two LSTM layers that process the input sequence in both forward and backward directions, merging their outputs at every time step. This comprehensive approach enables a deeper comprehension of the input sequence's past and future context, enhancing performance for tasks such as sentiment analysis, named entity recognition, and machine translation (Badawi, 2023a)

The BERT model is a deep learning technique that has been widely used for natural language processing tasks (Koroteev, 2021). In this paper, the BERT model is employed for the classification of emotions in the KMD dataset. To implement the BERT model, the training data is tokenized and fine-tuned using the Hugging Face Transformer library. The BERT tokenizer is used to split the text into tokens and add special tokens such as [CLS] and [SEP] as shown in Fig 3. A maximum length

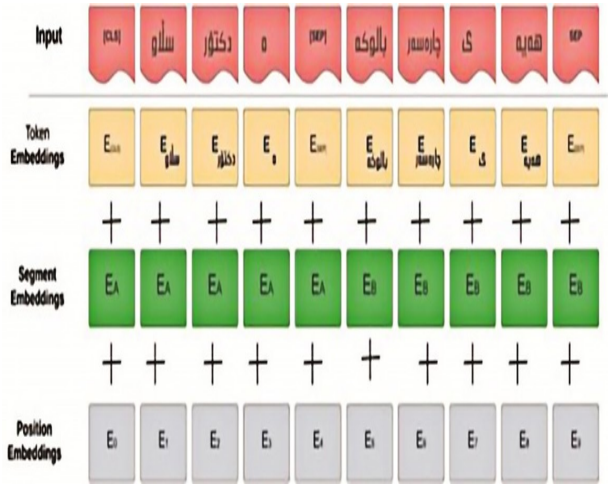


Figure 3: Representation of BERT Iokanization

of 128 is utilized for the BERT tokenizer. The BERT model is known for its ability to capture the context and meaning of words in a sentence. It achieves this through its attention mechanism, which enables it to focus on important parts of the input text. The fine-tuned BERT model is capable of accurately predicting the emotional sentiment of a given text. It is worth noting that, similar to the machine learning classifiers, the text data used for the BERT model is also represented as feature vectors using CountVectorizer, a text. preprocessing technique that converts the text data into numerical features.

## 5 Results And Discussions

To establish a basic benchmark for determining the sentiments present in our dataset, we utilized different machine-learning algorithms. The method used for splitting the data can significantly affect the accuracy of the model. Since our dataset is relatively big, we chose to use the holdout method for splitting the data. In the first phase, we used the 80% train 20% test technique. Furthermore, we split the train set using the holdout method to create a validation set. Having a validation set is vital, especially in the case of deep learning. The outcomes obtained from each classifier are presented in Table 2.

The results of the experiments show that all four classifiers (Naïve Bayes, SVM, BLSTM and BERT) have achieved reasonably good performance on the KMD dataset. In general, BERT performed the best among the four classifiers, with an overall F-score of 0.65, followed closely by BLSM with an overall F-score 0.64, SVM with an F-score

| Labels | Naive Bayes | SVM | BLSTM | BERT |
|---|---|---|---|---|
| Fear | 0.74 | 0.75 | 0.76 | **0.77** |
| Sadness | 0.54 | 0.55 | 0.56 | **0.58** |
| Joy | 0.57 | 0.56 | 0.58 | **0.58** |
| Surprise | 0.32 | 0.38 | 0.42 | **0.49** |

Table 2: F1-score obtained by each classifiers

| Labels | Naive Bayes | SVM | BLSTM | BERT |
|---|---|---|---|---|
| Fear | 0.77 | 0.78 | 0.79 | **0.80** |
| Sadness | 0.55 | 0.58 | 0.59 | **0.63** |
| Joy | 0.59 | 0.57 | 0.56 | **0.60** |
| Surprise | 0.34 | 0.39 | 0.43 | **0.48** |

Table 3: Precision-score obtained by each classifiers

of 0.63, and Naïve Bayes with an F-score of 0.62. The results also show that the classifiers performed differently across the different emotion labels. For the fear and sadness labels, Naïve Bayes and SVM performed similarly with F-scores of around 0.7 and 0.5, respectively, while BLSM and BERT performed slightly better with F-scores of 0.76, 0.56, 0.77 and 0.58, respectively. For the joy label, all four classifiers performed similarly with F-scores ranging from 0.56 to 0.58.

However, for the surprise label, BERT outperformed the other three classifiers with an F-score score of 0.49, while Naïve Bayes and SVM performed much worse with F-scores f 0.32 and 0.38, respectively. However, BLSM performed slightly better with a score of 0.42. This suggests that the surprise label may be more difficult to classify accurately than the other emotion labels in the KMD dataset. The performances of all models are illustrated in Fig 4.

Table 3 displays the precision scores of four classifiers (Naive Bayes, SVM, BLSTM, and BERT) for four different emotions: Fear, Sadness, Joy, and Surprise. Precision is a crucial metric to measure the accuracy of a classifier's positive predictions. The precision scores in the table offer valuable insights into the performance of these classifiers in emotion classification. BERT turns out to be the top-performing classifier in all emotions. It consistently outperforms other classifiers such as BLSTM, SVM, and Naive Bayes. In the "Fear" emotion category, BERT achieves the highest precision score of 0.80, followed closely by BLSTM with a precision of 0.79, while SVM and Naive Bayes lag slightly behind at 0.78 and 0.77, respectively. This trend is observed consistently across the various emotions. BLSTM shows competitive performance, ranking second across all emotions. It closely follows BERT's precision scores, indicating its effectiveness in emotion classification. SVM ranks third in most cases, followed by Naive Bayes, which consistently has the lowest precision scores. This suggests that advanced models like BERT and BLSTM tend to outperform traditional classifiers like SVM and Naive Bayes for this emotion classification task. Notably, the "Surprise" emotion is the most challenging for all classifiers. This is indicated by the especially lower precision scores for "Surprise" compared to the other emotions. The fact that all classifiers struggle to accurately classify "Surprise" suggests that this emotion may have unique characteristics that are difficult to capture, possibly due to its nuanced expression in the dataset or semantic complexities.

Table 4 provides a comprehensive overview of recall scores for four different classifiers used for emotion classification. Recall, also referred to as sensitivity or the true positive rate, is a critical metric that measures a classifier's ability to correctly identify all positive instances. Upon analyzing the recall scores in the table, we can draw important insights into the performance of these classifiers in the context of emotion classification. A distinct pattern emerges for each of the four emotions. For the "Fear" emotion category, BERT has the highest recall score of 0.83, followed closely by BLSTM at 0.80, SVM at 0.79, and Naive Bayes at 0.78. This pattern continues across the other emotions, where BERT consistently shows the highest recall scores, affirming its position as the best-performing classifier for emotion classification based on recall. The BLSTM model ranks second for most emotions, including "Fear," "Sadness," and "Joy," providing recall scores that are very close to BERT's. SVM and Naive Bayes tend to perform lower in terms of recall performance across all emotions. This consistent trend highlights the superiority of advanced models such as BERT and the effectiveness

| Labels | Naive Bayes | SVM | BLSTM | BERT |
|---|---|---|---|---|
| Fear | 0.78 | 0.79 | 0.80 | **0.83** |
| Sadness | 0.56 | 0.59 | 0.61 | **0.64** |
| Joy | 0.58 | 0.56 | 0.57 | **0.61** |
| Surprise | 0.35 | 0.40 | 0.44 | **0.49** |

Table 4: Recall-score obtained by each classifiers

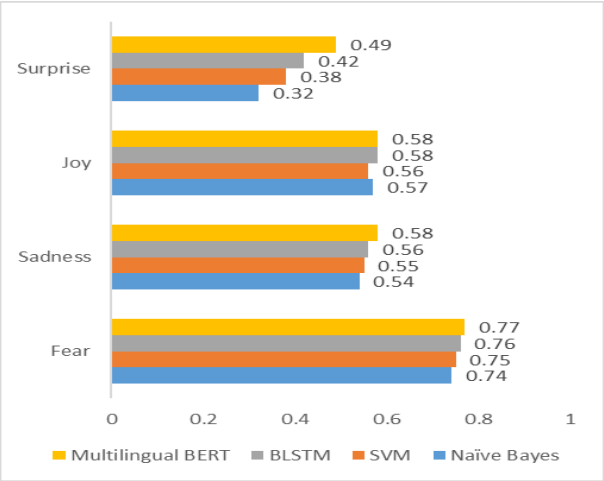


Figure 4: PErformance of each methods

of BLSTM to deliver competitive results in emotion classification tasks. Moreover, it also highlights that the "Surprise" emotion is a considerable challenge for all classifiers, as indicated by their notably lower recall scores for this emotion. This suggests that "Surprise" may possess unique characteristics that make it more difficult to accurately classify. These challenges may stem from the subtle and nuanced expressions associated with this emotion in the dataset or the intricacies of semantic interpretation.

Overall, these results indicate that the KMD dataset can be used to effectively train emotion classifiers for the Kurdish Sorani dialect using both classical machine learning methods and deep learning techniques like BERT. Furthermore, the results highlight the importance of evaluating classifiers across multiple emotion labels to better understand their performance and limitations.

## 6 Conclusion

The KMD dataset offers a new and valuable resource for researchers interested in the Kurdish Sorani dialect and emotional analysis. The dataset's comprehensive structure, annotation guidelines, and categories provide researchers with a robust foundation for further analysis and experimentation. The proposed model, which utilizes classical machine learning algorithms and deep learning BERT, demonstrates superior performance compared to traditional machine learning algorithms, such as Naive Bayes and Support Vector Machine classifiers and deep learning BLSTM, on all classes. The KMD dataset, along with the proposed model, paves the way for further research on the Kurdish Sorani dialect and provides a benchmark for future studies in natural language processing and machine learning (Badawi, 2023d).

## Limitations

The study presents significant findings in the realm of natural language processing and emotional analysis, yet several limitations deserve consideration. Firstly, the study predominantly focuses on the Kurdish Sorani dialect, which may restrict the generalizability of its results to other Kurdish dialects, such as Kurmanji and Gorani. Additionally, due to the extensive dataset used in the study, the training process demanded substantial GPU resources. Furthermore, the Kurdish language shares similarities with the Arabic and Persian alphabets, and users may have employed characters from these languages in their texts. Given the lack of specialized libraries or tools for automatic letter correction or replacement in Kurdish, such texts had to be excluded from the dataset. This removal potentially led to a loss of valuable data and linguistic diversity, particularly considering the low-resource nature of the Kurdish language.

## Ethics Statement

This study on the development of a Kurdish Multilabel Emotional Dataset (KMD) for the Kurdish Sorani dialect was conducted with a strong commitment to ethical research practices. The research team recognized the importance of ethical considerations throughout the project, including data collection, annotation, and analysis. This ethical statement outlines the key principles and practices adhered to during the study:

1. **Data Collection and Usage:** Data for the KMD dataset was collected using the Twitter API. We ensured that the data collection process adhered to Twitter's policies and guidelines. All data were obtained without compromising the privacy or consent of Twitter users. No personally identifiable information was collected or disclosed.

2. **Informed Consent:** As the dataset involved publicly available social media content, we did not seek explicit consent from individual Twitter users. However, we ensured that the dataset was used solely for research purposes and that no harm or undue exposure was caused to individuals or communities.

3. **Data Anonymization:** To protect the privacy and anonymity of individuals, all identifying information, including usernames and profile pictures, were removed from the collected data. The dataset was processed to ensure that it contained no personally identifiable information.

## Acknowledgements

We would like to extend our heartfelt gratitude to Mr. Fatih Kurt for his invaluable assistance in writing the Python code for the data collection process. His contributions were instrumental in the successful execution of this research project. We would also like to express our sincere thanks to the annotators who dedicated their time and effort to the annotation process. Their meticulous work and commitment were essential in creating a high-quality emotional dataset.

## References

Muhammad Abbas, Kamran Ali, Saleem Memon, Abdul Jamali, Saleemullah Memon, and Anees Ahmed. 2019. Multinomial naive bayes classification model for sentiment analysis. *IJCSNS International Journal of Computer Science and Network Security*, 19:62–67.

Sina Ahmadi, Hossein Hassani, and John McCrae. 2019. Towards electronic lexicography for the kurdish language. In *In Proceedings of the sixth biennial conference on electronic lexicograph (eLex)*, pages 881–906, Sintra, Portugal.

Amr Al-Khatib and Samhaa El-Beltagy. 2017. Emotional tone detection in arabic tweets. In *18th International Conference, CICLing 2017*, pages 105–114, Budapest, Hungary. Computational Linguistics and Intelligent Text Processing.

Kozhin Awlla and Hadi Veisi. 2022. Central kurdish sentiment analysis using deep learning. *Journal of University of Anbar for Pure Science*, 16:119–130.

Soran Badawi. 2023a. Data augmentation for sorani kurdish news headline classification using back-translation and deep learning model. *Kurdistan Journal of Applied Research*, 08:27–34.

Soran Badawi. 2023b. Kurdsum: A new benchmark dataset for the kurdish text summarization. *Natural Language Processing Journal*, 05:100043.

Soran Badawi. 2023c. Transformer-based neural network machine translation model for the kurdish sorani dialect. *UHD Journal of Science and Technology*, 7:15–21.

Soran Badawi. 2023d. Using multilingual bidirectional encoder representations from transformers on medical corpus for kurdish text classification. *Aro-The scientific Journal of Koya University*, 11:10–15.

Soran Badawi, Ari Saeed, Sara Ahmed, Peshraw Abdalla, and Diyari Hassan. 2023. Kurdish news dataset headlines (kndh) through multiclass classification. *Data in Brief*, 48:109120.

Ahmet Buran. 2011. Kurds and kurdish language. *Journal of Turkish Studies*, 6:43–57.

Hongyuan Cao, Pranab Sen, Anne Peery, and Evan Dellon. 2016. Assessing agreement with multiple raters on correlated kappa statistics. *Biometrical journal. Biometrische Zeitschrift*, 58:935–943.

Jair Cervantes, Farid García-Lamont, Lisbeth Rodríguez, and Asdrubal Lopez-Chau. 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215.

Berat Erol, Abhijit Majumdar, Patrick Benavidez, Paul Rad, Kim-Kwang Raymond Choo, and Mo. Jamshidi. 2019. Toward artificial emotional intelligence for cooperative social human-machine interaction. *IEEE Transactions on Computational Social Systems*, 7:1–13.

Xiang Feng, Keyi Yuan, Xiu Guan, and Longhui Qiu. 2022. An emotion analysis dataset of course comment texts in massive online learning course platforms. *Interactive Learning Environments*, pages 1–15.

Razhan Hameed, Sina Ahmadi, and Fatemeh Daneshfar. 2023. Transfer learning for low-resource sentiment analysis. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 37:Article 111. 14 pages.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, 82:3713–3744.

Mikhail Koroteev. 2021. Bert: A review of applications in natural language processing and understanding. Computing Research Repository, arXiv:2103.11943. version 2.

Ronak Kosti, Jose Alvarez, Adria Recasens, and Àgata Lapedriza. 2019. Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:2755–2766.

Flor Miriam Plaza del Arco, Carlo Strapparava, L. Alfonso Urena Lopez, and Maite Martin. 2020. EmoEvent: A multilingual emotion corpus based on different events. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France. European Language Resources Association.

Tarik Rashid, Arazo Mustafa, and Ari Saeed. 2018. Automatic kurdish text classification using kdc 4007 dataset. In *The 5th International Conference on Emerging Internetworking, Data and Web Technologies*, pages 187–198.

Ari Saeed, Riam Hussein, Chro Ali, and Tarik Rashid. 2022. Medical dataset classification for kurdish short text over social media. *Data in Brief*, 42:108089.

Gopendra Vikram Singh, Priyanshu Priya, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. EmoInHindi: A multi-label emotion and intensity annotated dataset in Hindi for emotion recognition in dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5829–5837, Marseille, France. European Language Resources Association.