Available online at www.garmian.edu.krd

Passer Journal

Passer 6 (Issue 2) (2024) 265-274

http://passer.garmian.edu.krd/

Enhancing Low-Resource Sentiment Analysis: A Transfer Learning Approach

Fatemeh Daneshfar[1*]

[1]*Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran.*

Received 30 January 2024; revised 20 March 2024;
accepted 21 April 2024; available online 12 May 2024

DOI: 10.24271/PSR.2024.440793.1484

ABSTRACT

The identification and extraction of subjective information from text, known as sentiment analysis, has seen advancements in employing cross-lingual approaches. However, the effective implementation and evaluation of sentiment analysis systems necessitate language-specific data to account for diverse sociocultural and linguistic variations. This paper outlines the process of collecting and annotating a dataset for sentiment analysis in Central Kurdish. We investigate classical machine learning and neural network-based techniques for this purpose. Furthermore, we adopt a transfer learning approach to enhance performance by leveraging pre-trained models for data augmentation. Our results demonstrate that despite the challenging nature of the task, data augmentation contributes to achieving high F1 scores and accuracy.

https://creativecommons.org/licenses/by-nc/4.0/

Keywords: Sentiment Analysis, Less-Resourced Languages, Kurdish, Natural Language Processing.

1. Introduction

Sentiment analysis has been a growing field of research that aims to classify the sentiment expressed in text automatically. The ability to accurately identify sentiment in text has numerous practical applications, such as market research, political analysis, and social media monitoring[61]. This problem has been widely studied in natural language processing (NLP) for some languages on various sources of data such as Twitter[1], YouTube[2, 3] and microblogs[4] across several domains as in finance[5], politics[6] and disaster management[7]. However, sentiment analysis, along with other related topics such as emotion recognition and hate speech detection, has yet to be addressed for many less-resourced languages, where collecting data and creating evaluation datasets pose significant challenges[8].

In this paper, we focus on sentiment analysis for Central Kurdish, a less-resourced Indo-European language spoken by over 30 million speakers. One of the major limitations in the previous studies on Kurdish sentiment analysis is the lack of evaluation datasets and models[63], making it impossible to carry out a comparative study. As such, our primary focus is to benchmark this task by manually annotating a dataset for evaluation purposes. Moreover, we explore the effect of emojis on sentiment analysis by training and testing the proposed models in two different settings, with and without emojis in the text.

Our experimental results demonstrate the effectiveness of our proposed approaches and provide insights into the challenges and opportunities of sentiment analysis for Central Kurdish. Furthermore, this benchmark can pave the way for more comparative studies in the future as models and datasets are openly available at https://github. Com/Hrazhan/sentiment. Given the existing gap in the literature for an open-source dataset and model for sentiment analysis of Central Kurdish, this study is beneficial for future progress in the field.

The paper is organized into several key sections to provide a comprehensive understanding of the research topic. In Section 2, we review the existing literature. Following this, in Section 3, we outline our approach in detail, covering aspects such as data collection, data annotation, transfer learning, and classification algorithms. In Section 4, we present our evaluation framework, comparing the baseline system, upsample system, and merged system, to understand the impact of individual components. Finally, in Section 5, we summarize our findings and propose directions for future research.

2. Related Work

With the increasing accessibility and popularity of opinion-rich resources such as online review sites, personal blogs, and social networks, opportunities, and challenges have arisen in this area. People can now use information technology to discover the opinions of others. Therefore, due to the increasing interest in systems that can directly explore thoughts and opinions, a lot of attention was drawn to the topics of opinion mining and sentiment analysis. These two fields deal with the computational encounter with the ideas, feelings, and mentality in the text. Sentiment analysis systems help gather information and provide insightful

* Corresponding author
E-mail address: f.daneshfar@uok.ac.ir (Instructor).
Peer-reviewed under the responsibility of the University of Garmian.

remarks from structured and unstructured online text such as email, blog posts, support tickets, web chats, social media networks, forums, and online comments.

In this section, the related work in sentiment analysis is examined. Also, previous research in this field is reviewed generally, and for the Kurdish language in particular. In addition, some of the previous methods that are useful in this field in dealing with low-resource languages are pointed out.

2.1 General Sentiment Analysis

Opinion mining algorithms replace feeling recognition and data processing with rule-based, automated, or hybrid methods. Rule-based systems perform sentiment analysis based on predefined rules and meaningful lexicons, while automated systems learn to extract opinions using machine learning[59] and labeled datasets. A wide range of techniques for the application of sentiment analysis has been proposed, including the hidden Markov model, Gaussian mixture model, support vector machine, neural network, contextual embeddings[60] and, zero-shot learning. In fact, there is no agreement regarding the most suitable classification method for the application of sentiment analysis and opinion mining yet.

Birjali et al.[9] review the existing methods for sentiment analysis using NLP. Among these methods, the use of techniques based on machine learning as well as statistical solutions has been beneficial. Furthermore, Rice and Zorn[10] present the resources and corpora that have been produced so far for speech sentiment analysis and shed light on how to annotate them. In this paper, the different methods that are available for evaluating the results related to the annotations of different users are compared with the same one, and the application of each one in languages with different specifications is presented, too.

With the progress in neural networks, many studies have focused on applying neural networks for the task of sentiment analysis. A new method for text emotion recognition is proposed by Basiri et al.[11] using a combination of four deep learning methods and a supervised machine learning model to recognize text emotions in annotated tweets. This model has obtained good results on the corpora of high-resource languages. Similarly, Jing et al.[12] present a framework with a new and robust model for detecting sentiment in text and stock price prediction using deep learning.

More recently, deep neural networks based on transformer architecture and contextual embeddings, such as BERT[13], and other variants, such as multilingual mBART[14], have achieved state-of-the-art performance in many natural language classification tasks, including standard emotion and offensive speech recognition. Miok et al.[15] propose a Bayesian method using Monte Carlo in the attention layers of transformer models to be used to estimate the probability of hateful sentences with high reliability. Plaza-del Arco et al.[16] propose several methods for detecting text emotions in Spanish and compare their results. In this paper, the performance of models based on deep learning and machine learning, pre-trained language models, the transformer architecture with attention mechanism, as well as multilingual and monolingual models of the pre-trained Spanish models are compared. The performance of each one has been evaluated in terms of error analysis to understand the difficulty of the task.

Since most of the automatic sentiment analysis approaches classify the problem as binary (positive or negative), without addressing the local focus or the goal-oriented nature of the speech, Chiril et al.[17] use a multi-objective method to detect both the speech emotion and hate speech. In the presented method, a dataset labeled in one language is used to transfer knowledge to different data sets in other languages. In this model, by extracting common linguistic features in a labeled data set, it has been used to transfer this knowledge to detect offensive words in other linguistic data.

In addition, topics like racism, xenophobia, sexism, misogyny, and targets of offensive words have also been identified. Also, the effect of emotional knowledge encoded in emotional computing resources (SenticNet, EmoSenticNet) and hate words with semantic structure (HurtLex) has been studied in determining specific manifestations of offensive speech.

In the case of the availability of pre-trained language models with annotated sentiment analysis datasets, it is also possible to implement a zero-shot or few-shot approach for this task. For instance, Pamungkas et al.[18] use a joint learning architecture based on the zero-shot approach for cross-lingual sentiment analysis. Initially, this method was employed on a resource-rich language, leading to the development of models for less-resourced languages.

2.2 Kurdish Sentiment Analysis

Despite the numerous efforts made in recent years to develop and process the Kurdish language, it remains in the early stages of development, thus lagging considerably behind current language processing technologies[19]. Although much research has been published on tasks such as machine translation[20], tokenization[21], and developing toolkits[22], only a handful of studies address sentiment analysis for Kurdish and its varieties.[1]

As one of the earliest studies, Abdulla and Hama[23] developed a sentiment analyzer using a naive Bayes classifier with bag-of-words containing frequent words in 15,000 text documents, among which 8000 are labeled as positive reviews and the rest as negative. The documents were collected from various social networks such as Facebook, Twitter, and Google+, and an F1 measure of 0.72 was reported for this classifier. Similarly, Amin et al.[24] investigate the challenges of applying sentiment analysis approaches in the Kurdish language. These challenges are related to all stages of sentiment analysis processing, from data collection to feature extraction and classification. Also, two different proposed methods, a machine learning-based method and a lexicon-based method, are presented to face these obstacles.

Furthermore, Awlla and Veisi[25] address sentiment analysis for Kurdish and describe the creation of a dataset containing 14,881 comments from various Facebook pages. To create an analyzer, Word2vec embeddings, along with a recurrent neural network

[1] An updated list of the existing tools and works on Kurdish language processing is provided at https://github.com/sinaahmadi/awesome-kurdish.

classifier, are used with a reported accuracy of 71.35%. In the same vein, Azad et al.[26] address fake news detection for the Kurdish language. In this regard, a corpus is collected from a few news websites and then annotated and evaluated for the task using a few classical machine learning algorithms.

Table 1 summarizes the previous studies along with their datasets and techniques proposed for Kurdish sentiment analysis. Although the task has been tackled a few times, they are of little or no avail, chiefly due to the lack of open-source tools and resources. This challenge not only impedes progress in the field but also makes experimental comparisons impossible. Therefore, our main focus in this work is on the

Table 1: Previous works in sentiment analysis for Central Kurdish in comparison with the current paper. Inter-annotator agreement (IAA) is provided if reported in the papers. # refers to the number. Unlike the current work, G and S refer to gold-standard and si

Study	Data source	# instances	Methodologies	Key findings	Limitations	IAA	Open-source
Abdulla and Hama [23]	Facebook, Twitter, YouTube	15,000	Classical ML	Considering the unique writing styles in Kurdish social network texts	Handling diverse writing styles in Kurdish texts, Difficulties in accurately interpreting sentiments in informal language	-	✗
Awlla and Veisi [25]	Facebook	14,881	Neural networks	Dataset of 14,881 comments, Word2Vec and LSTM models	Limited labeled data for training, Challenges in accurate sentiment labeling, Dependency on Facebook comments	-	✗
This work	Twitter	1,185 (G) 4,500 (S)	classical ML, neural networks, transfer learning			0.84	✓

2.3 Sentiment Analysis in Low-Resourced Scenarios

Most of the methods presented so far have been based on annotated corpora and high-resource languages. However, some models have utilized approaches such as zero-shot, few-shot, or transfer learning to address the challenges posed by low-resource languages. Zero-shot learning[57] involves training a model on a source task with labeled data and then applying it directly to a target task without any additional training data, leveraging shared features or representations between the tasks. Few-shot learning[58], on the other hand, aims to train models with only a small amount of labeled data, often by adapting pre-trained models or using meta-learning techniques to quickly generalize to new tasks. Transfer learning[44] involves pre-training a model on a large dataset from a source domain and then fine-tuning it on a smaller dataset from a target domain, allowing the model to transfer knowledge and features learned from the source to the target domain. These techniques enable the adaptation of models trained on rich languages to low-resource ones, facilitating sentiment analysis in diverse linguistic contexts.

In less-resource environments, that lack suitable annotated data, sentiment analysis can be performed using methods such as transfer learning[27–29], semi-supervised learning[30–32, 62], and unsupervised learning[33, 34]. Transfer learning[44] involves leveraging knowledge gained from one task or domain to improve performance in another, thus enabling sentiment analysis even with sparse data. Semi-supervised learning utilizes a combination of labeled and unlabeled data to train models, making it particularly useful when labeled data is scarce. Unsupervised learning, on the other hand, focuses on extracting patterns and structures from unlabeled data alone, offering insights into sentiment without the need for annotated datasets. By employing these techniques, sentiment analysis can adapt and thrive in resource-constrained environments, offering valuable insights despite data limitations. In these models, techniques such as semi-supervised manifold regularization[30, 35], methods based on recursive automatic encoders[36, 37], and hidden latent variable models can be utilized. Also, a group of researchers has used machine translation systems to translate other languages into English, and to use English sources in sentiment analysis of low-resource languages[38–40]. Such techniques facilitate sentiment analysis in multilingual and cross-lingual setups[41].

However, when assessing sentiment analysis systems in low-resourced contexts, it is necessary to consider the specifics of the evaluation methods employed. This involves highlighting the unique challenges and considerations inherent in evaluating sentiment analysis for languages such as Kurdish. Given the diverse cultural and linguistic nuances at play, it becomes imperative to rely on annotated data curated by native speakers. This not only ensures the accuracy and relevance of the evaluation process but also takes into account various cultural and linguistic observations specific to each language. Thus, in the case of a low-resource language like Kurdish, the development of an annotated corpus and subsequent creation of an opinion analysis model based on it emerge as necessary and indispensable steps.

In this paper, in contrast to these previous works, we first introduce an annotated corpus and then use several simple methods to evaluate it. In the same spirit of works applying transfer learning, we also rely on a pre-trained model to leverage

information from a richly-resourced language, English, for Kurdish sentiment analysis in this work.

3. Methodology

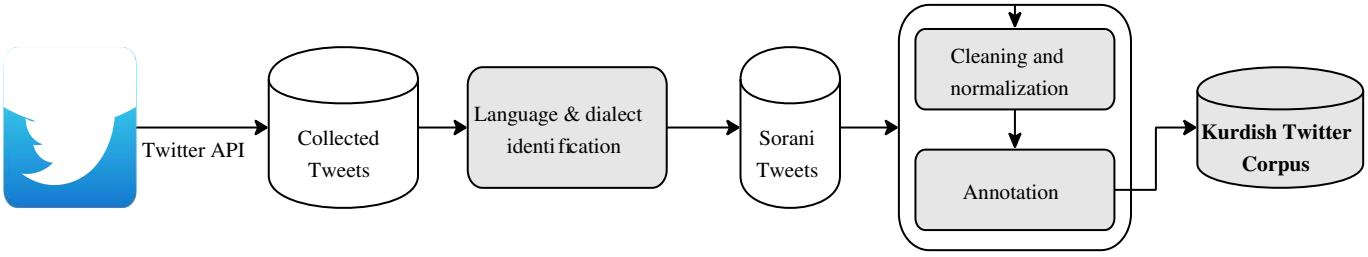

Figure 1: A schema of data collection and annotation for Central Kurdish sentiment

3.1 Data Collection

We selected Twitter as our source of data. The tweets were mostly collected between June 6th and June 14th of 2024. Using Twitter API[2], we collected a few thousand tweets and filtered out the potentially Kurdish ones using regular expressions based on distinctive characters such as ڕ and ڵ along with fastText language identifier[42]. Twitter supports searching language-specific content with the keyword "lang:ckb". Furthermore, only tweets that are correct with regard to grammar or spelling are included. In cases where a slight modification in the tweet structure (but not wording) can make it conform to this condition, the annotator is asked to apply modifications as long as rewriting the whole tweet is not required. In addition, poems, famous quotations, and code-switched tweets are removed. The tweets are manually verified to be written in the Perso-Arabic script of Central Kurdish. Finally, we use KLPT[22] for text preprocessing and orthography normalization.

3.2 Data Annotation

The annotation process was carried out by two annotators native of Central Kurdish. The annotation was conducted following an annotation guide that describes the selection of appropriate tweets and provides instructions on how to annotate them. The annotation is aimed at document-level sentiment classification to find the general sentiment of the author in an opinionated text, including emojis. Therefore, annotators determine a tweet as "subjective" or "objective" first, then, in the case of subjectivity, a label among the following ones is selected: positive, negative, mixed, neutral, and none. Usually, a subjective sentence is supposed to represent sentiment, not an objective one. Overall, 1769 instances are annotated.

In order to evaluate the quality of the annotations, we calculate the inter-annotator agreement (IAA) with Krippendorff's alpha[43] for nominal sentiment labels between the two annotators. As indicated in Table 1, the annotations achieve 0.84 of IAA in sentiment analysis. In other words, annotators agree on 84% of the labels they were expected to disagree on by chance. Krippendorff's alpha provides a useful measure of how often labels from different annotators agree in such a way that isolates annotators' skills. Finally, the annotations are aggregated by a third annotator where common annotations are finalized and those that are in conflict, are rectified. For instance, if one annotator marks a tweet as subjective and the other one as objective, it is considered mixed.

It is worth mentioning that the annotation campaign initially intended to include hate speech detection where tweets were tagged as offensive and targeting individuals or groups. This, however, was a more challenging task with low agreement among annotators and a considerable imbalance and bias. The imbalance is primarily due to under-represented classes. Therefore, we did not include hate speech detection in this project. Detection of hate speech is particularly challenging as hatefulness is often a relative notion. For instance, the remarks of an atheist may be intellectual and factual but be considered hate speech toward religious groups. Similarly, how religious practitioners describe social issues and natural phenomena might be aggressive and offensive utterances.

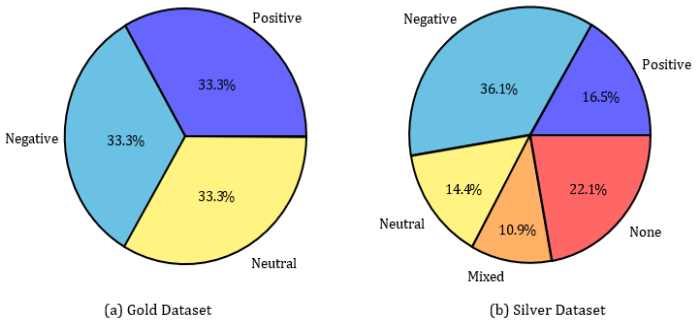

Figure 2: Distribution of the instances among the classes in our gold-standard and silver standard datasets used for sentiment analysis of Central Kurdish.

3.3 Transfer Learning

The annotated dataset is imbalanced with remarkably more negative than positive and neutral instances. This imbalance can be problematic for machine learning models, as they may struggle to accurately classify the minority class. To remedy this issue, we also create another dataset with a more balanced distribution of

[2] Used in June 2021

sentiment using transfer learning. By the latter dataset, we are able to generate a larger dataset with a more even distribution of sentiment. Following this, we train several classical machine learning models on both datasets as well as a bidirectional long short-term memory model.

Robust classification systems require a considerable set of data. Given the small size of the annotated data, hereafter referred to as gold-standard dataset, due to limitations in time and cost, we also extend the data for Central Kurdish sentiment analysis using transfer learning[44]. Transfer learning refers to a set of methods where a model developed for a task is reused as another task; it has been previously used in low-resource NLP[45], particularly in sentiment analysis[46]. In our approach, we rely on a neural machine translation model pre-trained on many languages including Central Kurdish and English along with a sentiment analysis model pre-trained on English.

To that end, we use the No Language Left Behind (NLLB)[47] to translate the collected tweets to English described in §3.1, excluding those that are annotated and included in the gold-standard dataset. The NLLB model is used to translate each tweet into English, because the RoBERTa model we used to annotate the silver-standard data only performed well in English. When the English translated tweets got labeled, the labels were put back on the original Central Kurdish tweets.

Following this, we determine the translated tweets with their sentiment labels using a RoBERTa-base model trained on 124M tweets in English from January 2018 to December 2021[48].[3] Additionally, we made sure that out of the 1500 tweets, at least 500 tweets in each sentiment class (positive, negative, and neutral) contained an emoji. This was done to allow us to examine the impact of emojis on the performance of the proposed models.

We believe that the resulting dataset, henceforth called the silver-standard dataset, can remedy the lack of data for Central Kurdish sentiment analysis. The number of instances per class is provided in Table 2.

Table 2: Number of instances in the annotated dataset (gold-standard) and the silver-standard one created using transfer learning

Dataset	Positive	Negative	Neutral	Total
Gold-standard	292	639	254	1185
Silver-standard	1500	1500	1500	4500

3.4 Classification Algorithms

We train a total of five mainstream classification algorithms that are commonly used in sentiment analysis tasks due to their effectiveness. Four of these algorithms, namely logistic regression, decision trees, random forest trees, and support vector machines, are classical machine learning methods. On the other hand, a deep learning method is employed using a bi-directional long short-term memory model. Datasets are split into train and test with an 80%-20% ratio. The hyperparameter tuning process involved systematically experimenting with different parameter settings to optimize the performance of our machine learning models. This iterative approach, often complemented by cross-validation techniques, helped us identify parameter values that enhanced the predictive capabilities of the proposed models, ensuring they are well-suited to the task of sentiment analysis.

3.4.1 Logistic Regression (LR)

This is a statistical method that we used to classify data into one of two categories[49]. It is a linear model that is simple to implement and can be used for the multiclass classification task. In this study, we used LR to classify tweets into positive, negative, or neutral classes. To optimize the performance of the model, we set the hyperparameters, including the optimizer to L-BFGS and the multinomial loss function, i.e. softmax.

3.4.2 Decision Trees (DT)

This is a popular algorithm used for classification tasks[50]. It works by recursively partitioning the data into subsets based on the values of the input features.

3.4.3 Random Forest (RF)

This is an ensemble method that combines multiple decision trees to improve the performance of the model[51]. By averaging the predictions of multiple DTs, RF reduces the overfitting and variance of a single DT. We utilize cross-entropy as the loss function and set the number of estimators to 30. We determined the optimal number of estimators by conducting a grid search, where we tested a range of values and selected the one that maximized model accuracy on the validation set.

3.4.4 Support Vector Machines (SVM)

This works by finding the hyperplane that maximally separates the data into different classes[52]. We used Linear SVC, a type of SVM, which operates by identifying the hyperplane that can best divide the data points into distinct categories. This approach is based on maximizing the margin between the data points and the hyperplane, ensuring that the model can accurately classify future observations.

3.4.5 Bidirectional Long Short-Term Memory (BiLSTM)

This is a type of recurrent neural network (RNN) that is able to process sequential data by considering past and future contexts[53] which makes it a powerful tool for our task. Additionally, it has been previously widely used for the same task. We employ embeddings to represent the data, wherein the embedding size is set to 100. The network architecture consists of two bidirectional layers, with sizes 64 and 32, respectively. We also include a dropout of 0.3 between the layers to prevent overfitting. Finally, we apply the softmax activation function to the last dense layer. Similarly, we adjust the embedding size through a series of trials, evaluating the impact of various sizes on model performance.

To train each of the classical machine learning models, we used the text features of the tweets as inputs. These features were

[3] The latest version of 2022 is available at https://huggingface.co/cardiffnlp/twitter-roberta-base-sentimentlatest.

represented using the term frequency-inverse document frequency (TF-IDF)[54] representation method.

It is worth mentioning that due to the high imbalance among the five labels, we only include 'positive', 'negative', and 'neutral' in the classification task. This limitation can be further studied in future work. The distribution of different classes is illustrated in Figure 2. A few examples are provided in Table A1 in Appendix.

4. Evaluation

4.1 Evaluation Metrics

We use several common evaluation metrics to evaluate the performance of the proposed models, namely accuracy, F_1 score, precision, and recall as described by Powers [55] as follows: • Accuracy is the ratio of the number of correct predictions to the total number of predictions.

- Precision is the ratio of true positive predictions to the total number of positive predictions.
- Recall is the ratio of true positive predictions to the total number of actual positive instances.
- F_1 score is the harmonic mean of precision and recall.

While accuracy, precision, recall, and F1 score are widely used to evaluate model performance, they have limitations, especially in imbalanced datasets. Accuracy can be misleading, as it might reflect the dominance of the majority class rather than the model's ability to identify all classes accurately. Precision and recall offer a more nuanced view by considering false positives and false negatives, respectively, but they may still not fully capture the performance across highly imbalanced classes.

4.2 Results

4.2.1 Baseline system

As a baseline system, we evaluate the performance of the models by testing on the gold-standard dataset without including emojis and training on the baseline, upsampled and merged training sets. The experiment results are presented in Table 3. The baseline results indicate that SVM achieves the highest accuracy of 56% and F_1 score of 53% among all models. This said the BiLSTM model outperforms all of the classical models in accuracy, achieving an accuracy of 57%. On the other hand, the analysis of the F_1 score, precision, and recall metrics reveals a remarkable imbalance among all models indicating a bias toward the negative class, which has the highest number of samples (639). This highlights the need to address data imbalance and improve the model's ability to distinguish between the different sentiment classes. The BiLSTM model outperforms classical models due to its ability to capture sequential information and context in text data

Although data augmentation in the unsampled and merged systems does not substantially improve the F_1 score and accuracy in general, the transfer learning approach (twitter-robert model and NLLB system in Table 3) increases the baseline with a 0.54 F_1 score and 0.53 accuracy.

4.2.2 Upsample system

As the second system, we proceed to upsample the gold-standard dataset by incorporating additional samples from the silver-standard dataset. As such, we increase the number of samples to 700 instances per class to reduce the imbalance between classes. This leads to an improvement in performance for all the models, as demonstrated in Table 3. LR achieves an accuracy of 59%, which is a notable improvement. Surprisingly, the BiLSTM model experiences a decline in performance, although the F_1 score, recall, and precision metrics are more balanced and exhibit similar values. This implies a superior ability to generalize to all classes, in contrast to the previous evaluation that was conducted on the imbalanced data.

4.2.3 Merged system

In a last experimental setting, we merge both the gold and silver standard datasets, while still maintaining a balanced distribution of 1700 samples per class. We observe an overall improvement in performance, with more balanced metrics indicating better generalization across all classes. SVM achieves the highest accuracy and F_1 score of 61%. The performance of all the systems in the three setups of baseline, upsample and merged, respectively denoted by the dataset sizes of 1185, 2100 and 5100 instances, is shown in Figure 3. The F1 score gradually improves with the size of datasets in all systems, except for DT. There could be various factors contributing to this issue, such as the quality of the combined dataset which may include noise, and the need to optimize the hyperparameters to ensure effective generalization for the larger dataset.

To conclude, the upsampling technique using the silver-standard dataset results in a considerable improvement in performance by at least 8%, while also achieving a better balance between F_1 score, recall, and precision, indicating better generalization. One notable observation is that the performance of a BiLSTM model may be highly dependent on the specific task and dataset, and may not always outperform classical machine learning models.

Table 3: Performance of our systems with different test sets and setups without including emojis. Data augmentation improves the results (the highest scores are specified in bold).

Model	Test set	System	Precision	Recall	F_1	Accuracy
LR	Baseline	Baseline	0.49	0.54	0.44	0.54
		Upsample	0.45	0.42	0.43	0.42
		Merged	0.48	0.48	0.48	0.48
	Upsample	Upsample	**0.61**	0.59	0.59	0.59
	Merged	Merged	**0.61**	0.6	0.6	0.6
SVM	Baseline	Baseline	0.53	0.56	0.53	0.56
		Upsample	0.47	0.45	0.45	0.45
		Merged	0.47	0.47	0.47	0.47
	Upsample	Upsample	0.54	0.53	0.53	0.53
	Merged	Merged	**0.61**	**0.61**	**0.61**	**0.61**
RF	Baseline	Baseline	0.48	0.54	0.46	0.54
		Upsample	0.52	0.47	0.48	0.47
		Merged	0.47	0.44	0.45	0.44
	Upsample	Upsample	0.55	0.52	0.52	0.52
	Merged	Merged	0.55	0.55	0.55	0.55
DT	Baseline	Baseline	0.45	0.44	0.44	0.44
		Upsample	0.46	0.42	0.43	0.42
		Merged	0.42	0.37	0.39	0.37
	Upsample	Upsample	0.5	0.5	0.5	0.5
	Merged	Merged	0.48	0.48	0.48	0.48
BiLSTM	Baseline	Baseline	0.56	0.26	0.36	0.57
		Upsample	0.44	0.41	0.44	0.46
		Merged	0.42	0.42	0.44	0.44
	Upsample	Upsample	0.59	0.52	0.55	0.53
	Merged	Merged	0.56	0.57	0.56	0.57
twitter-roberta	Baseline	NLLB	0.53	0.53	0.54	0.53

4.3 Ablation Analysis

As an ablation analysis, we evaluate the impact of emojis on sentiment analysis for Central Kurdish. To that end, we train and test under the same setups presented in Table 3 with emojis in the tweets.

The experiment results presented in Table 4 indicate that the models achieve superior results when analyzing only the text of the tweets, without including emojis. We observe that the inclusion of emojis decreases the performance of all models by 2-3%. This could be attributed to potential inconsistencies between the sentiment conveyed by the emojis and the actual sentiment expressed in the tweet. Similarly, the presence of emojis continues to decrease the performance of all classical ML models except the BiLSTM, which is a neural network-based model. Neural networks tend to be more data-hungry and may benefit from the increased amount of data resulting from the dataset merge.

Nevertheless, the impact of emojis on performance is only evident when the number of samples is increased. Increasing the size of the dataset leads to more resilient models

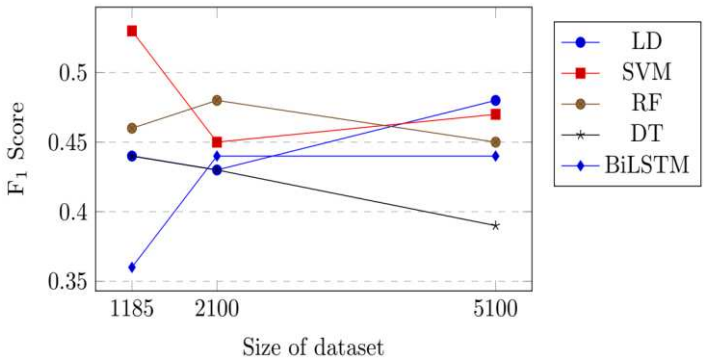

Figure 3: Performance of various systems (without emojis) in different setups that are tested on the gold-standard data (20% of the 1185 instances) and trained on the gold-standard train set (1185 instances), up-sampled data (2100 instances) and those two merged (5100 instances), all with an 80-20 train-test ratio.

to emojis. While emojis tend to decrease the performance of classical ML models, they actually improve the performance of only the BiLSTM model.

Table 4: Performance of our systems in different setups with data containing emojis.

Model	System	Precision	Recall	F_1	Accuracy
LR	Baseline	0.49	0.54	0.43	0.54
	Upsample	0.59	0.58	0.58	0.58
	Merged	0.62	0.61	0.61	0.61
SVM	Baseline	0.53	0.55	0.53	0.55
	Upsample	0.53	0.53	0.53	0.53
	Merged	0.58	0.57	0.57	0.57
RF	Baseline	0.46			
	Upsample	0.51	0.54 0.5	0.42 0.5	0.54 0.5
	Merged	0.57	0.56	0.57	0.56
DT	Baseline	0.46	0.45	0.46	0.45

	Upsample	0.46	0.45	0.45	0.45
	Merged	0.48	0.48	0.48	0.48
BiLSTM	Baseline	0.51	0.5	0.5	0.51
	Upsample	0.49	0.48	0.49	0.49
	Merged	0.59	0.57	0.58	0.58

5. Conclusion and Future Work

In this study, we addressed the task of sentiment analysis for Central Kurdish with the primary goal of providing a benchmark. We described the collection and annotation of data based on tweets and also, another data augmentation approach using transfer learning to automatically generate sentiment analysis instances. This dataset is openly available and can pave the way for future developments for Central Kurdish. To train and evaluate the proposed models, we employed a variety of classical machine learning techniques, namely logistic regression, decision trees, random forest trees, and support vector machines along with the neural network-based model of bi-directional long short-term memory.

Comparing the performance of different techniques under various setups, we demonstrated that data augmentation is beneficial to increase the accuracy and F_1 score remarkably. We also carried out an ablation analysis and found that sentiment analysis without the actual emoji characters can slightly improve the results.

One of the limitations of the current study is the number of sentiment labels being positive, negative and neutral. This along with the lack of further annotated resources, particularly a polarity lexicon or translation of emojis, can be addressed in the future by incorporating lexicon-based information, as suggested by Lemmens et al.[56]. Exploring cross-lingual approaches for sentiment analysis of other Kurdish varieties and dialects is also suggested as a future work.

6. Acknowledgments

The authors would like to thank Sina Ahmadi, Razhan Hameed, Behshad Davoudi, Iman Ghavami, and Arvin Rasooli at the University of Kurdistan for annotating the sentiment analysis data in Central Kurdish.

This paper is derived from the research project of “Kurdistan Studies Institute” of University of Kurdistan. We express our gratitude and appreciation for providing financial resources.

7. Appendix

We showcase a few examples of sentiment analysis predictions by the BiLSTM model. Specifically, we present examples of both correct and incorrect predictions made by the model, in order to provide insight into its strengths and weaknesses. By examining these examples, readers can gain a deeper understanding of the challenges and opportunities involved in sentiment analysis for Central Kurdish, and the potential limitations of current machine learning models in this domain.

Table A1: A few examples in our annotated dataset with references and BiLSTM model predictions. Code-switched words like سپۆیلد‘spoiled’ and incorrect spellings are highlighted in red and yellow respectively. The translations are free. Note that these sentences are selected as examples and do not reflect authors opinions.

Reference	Prediction	Tweet
Negative	Negative	ئاخ منالّی سپۆیلد و هیچ نەدیو چەن تێنەگەشتوو چەن ناشیرین چەن بێ سوود. Oh, how ugly, useless and fool (is) a spoiled and bad-mannered kid.
Neutral	Negative	یادی بە خێر 😁 بە منالّی خەونەکانمان چەند گەورە بوون 😅😅💔💔💔 Those were the days 😁 How big our childhood dreams were 😅😅💔💔💔
Neutral	Positive	عەشقی ڕاستەقینە وەک و نوێژ وایە، دوای ئەوەی نیەتت هێنا نابێ سەیری دەور و بەرت بکەی. Real love is like prayer. You should not get distracted when doing it.
Negative	Negative	بە درۆی پیاوە گەورەکان ئەڵێن سیاسەت The big lies of big men (people) are called politics
Positive	Positive	ئەو کەسانەی بەقسەم ئەکەن لام زۆر شیرینننننن I find those who listen to me so sweeeeeeeet
Positive	Neutral	دڵت بۆ لێدان دروست کراوە و ڕوخسارت بۆ پێکەنین دروست کراوە و ژیانیشت تەنیا مولّکی خۆتە. Your heart is created to beat and your face to smile and your life is yours.
Positive	Negative	ئاشق بوون ئەرکی پیاوە، ژن خۆی عیشقە. 💛 Falling in love is man’s work. Woman is love. 💛
Neutral	Neutral	تۆ سەرنجی ئەم هەموو هاتن و چوونی ئینسانانە بدەن؛ وەکوو خۆرمان لێ هاتووە، ئەم ئاوا ئەبێت و ئەوی تر هەڵدێت. Look how humans arrive and leave. We are like the sun: coming and going.

References

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.J.: Sentiment analysis of Twitter data. In: Proceedings of the Workshop on Language in Social Media (LSM 2011), pp. 30–38 (2011)
2. Asghar, M.Z., Ahmad, S., Marwat, A., Kundi, F.M.: Sentiment analysis on YouTube: A brief survey. arXiv preprint arXiv:1511.09142 (2015)
3. W¨ollmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., Morency, L.-P.: YouTube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intelligent Systems 28(3), 46–53 (2013)
4. Aisopos, F., Papadakis, G., Tserpes, K., Varvarigou, T.: Content vs. context for sentiment analysis: a comparative analysis over microblogs. In: Proceedings of the 23rd ACM Conference on Hypertext and Social Media, pp. 187–196 (2012)
5. Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L.T., Trajanov, D.: Evaluation of sentiment analysis in finance: from lexicons to transformers. IEEE access 8, 131662–131682 (2020)
6. Ramteke, J., Shah, S., Godhia, D., Shaikh, A.: Election result prediction using Twitter sentiment analysis. In: 2016 International Conference on Inventive Computation Technologies (ICICT), vol. 1, pp. 1–5 (2016). IEEE
7. Beigi, G., Hu, X., Maciejewski, R., Liu, H.: An overview of sentiment analysis in social media and its applications in disaster relief. Sentiment analysis and ontology engineering: An environment of computational intelligence, 313–340 (2016)
8. Winata, G.I., Aji, A.F., Cahyawijaya, S., Mahendra, R., Koto, F., Romadhony, A., Kurniawan, K., Moeljadi, D., Prasojo, R.E., Fung, P., et al.: Nusax: Multilingual parallel sentiment dataset for 10 Indonesian local languages. arXiv preprint arXiv:2205.15960 (2022)
9. Birjali, M., Kasri, M., Beni-Hssane, A.: A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems 226, 107134 (2021)
10. Rice, D.R., Zorn, C.: Corpus-based dictionaries for sentiment analysis of specialized vocabularies. Political Science Research and Methods 9(1), 20–35 (2021)
11. Basiri, M.E., Nemati, S., Abdar, M., Asadi, S., Acharrya, U.R.: A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets. Knowledge-Based Systems 228, 107242 (2021)
12. Jing, N., Wu, Z., Wang, H.: A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. Expert Systems with Applications 178, 115019 (2021) https://doi.org/10.1016/j.eswa.2021.115019
13. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
14. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics 8, 726–742 (2020)
15. Miok, K., Skrlj, B., Zaharie, D., Robnik-ˇ Sikonja, M.: To ban or not to ban:ˇ Bayesian attention networks for reliable hate speech detection. Cognitive Computation 14(1), 353–371 (2022)
16. Plaza-del-Arco, F.M., Molina-Gonz´alez, M.D., Urena-L´opez, L.A., Mart´ınValdivia, M.T.: Comparing pre-trained language models for spanish hate speech detection. Expert Systems with Applications 166, 114120 (2021)
17. Chiril, P., Pamungkas, E.W., Benamara, F., Moriceau, V., Patti, V.: Emotionally informed hate speech detection: a multi-target perspective. Cognitive Computation 14(1), 322–352 (2022)
18. Pamungkas, E.W., Basile, V., Patti, V.: A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. Information Processing & Management 58(4), 102544 (2021)
19. Ahmadi, S.: On the Current State of Kurdish Language Processing. Proceedings of the 5th International Conference on Kurdish Linguistics (ICKL-5) Conference (2021)
20. Ahmadi, S., Masoud, M.: Towards machine translation for the Kurdish language. arXiv preprint arXiv:2010.06041 (2020)
21. Ahmadi, S.: A tokenization system for the Kurdish language. In: Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 114–127 (2020)
22. Ahmadi, S.: KLPT–Kurdish language processing toolkit. In: Proceedings of
23. Second Workshop for NLP Open Source Software (NLP-OSS), pp. 72–84 (2020)
24. Abdulla, S., Hama, M.H.: Sentiment analyses for kurdish social network texts using naive bayes classifier. Journal of University of Human Development 1(4), 393–397 (2015)
25. Amin, M.H.S.M., Al-Rassam, O., Faeq, Z.S.: Kurdish language sentiment analysis: Problems and challenges. Mathematical Statistician and Engineering Applications 71(4), 3282–3293 (2022)
26. Awlla, K.M., Veisi, H.: Central Kurdish Sentiment Analysis Using Deep Learning. Journal of University of Anbar for Pure science 16(2) (2022)
27. Azad, R., Mohammed, B., Mahmud, R., Zrar, L., Sdiqa, S.: Fake news detection in low-resourced languages “Kurdish language” using machine learning algorithms. Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12(6), 4219–4225 (2021)
28. Bensoltane, R., Zaki, T.: Towards arabic aspect-based sentiment analysis: a transfer learning-based approach. Social Network Analysis and Mining 12(1), 1–16 (2022)
29. Catelli, R., Bevilacqua, L., Mariniello, N., Carlo, V.S., Magaldi, M., Fujita, H., De Pietro, G., Esposito, M.: Cross lingual transfer learning for sentiment analysis of italian tripadvisor reviews. Expert Systems with Applications, 118246 (2022)
30. Kumar, A., Albuquerque, V.H.C.: Sentiment analysis using xlm-r transformer and zero-shot transfer learning on resource-poor indian language. Transactions on Asian and Low-Resource Language Information Processing 20(5), 1–13 (2021)
31. Gupta, R., Sahu, S., Espy-Wilson, C., Narayanan, S.: Semi-supervised and transfer learning approaches for low resource sentiment classification. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5109–5113. IEEE
32. Li, N., Chow, C.-Y., Zhang, J.-D.: Seml: A semi-supervised multi-task learning framework for aspect-based sentiment analysis. IEEE Access 8, 189287–189297 (2020)
33. Liu, P., Marco, C., Gulla, J.A.: Semi-supervised sentiment analysis for underresourced languages with a sentiment lexicon. In: INRA@ RecSys, pp. 12–17
34. Fern´andez-Gavilanes, M., Juncal-Mart´ınez, J., Garc´ıa-M´endez, S., CostaMontenegro, E., Gonz´alez-Castan˜o, F.J.: Creating emoji lexica from unsupervised sentiment analysis of their descriptions. Expert Systems with Applications 103, 74–91 (2018)
35. Rana, T.A., Shahzadi, K., Rana, T., Arshad, A., Tubishat, M.: An unsupervised approach for sentiment analysis on social media short text classification in roman urdu. Transactions on Asian and Low-Resource Language Information Processing 21(2), 1–16 (2021)
36. Gupta, R.: Data augmentation for low resource sentiment analysis using generative adversarial networks. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7380–7384 (2019).
37. Nankani, H., Dutta, H., Shrivastava, H., Rama Krishna, P., Mahata, D., Shah, R.R.: Multilingual sentiment analysis, pp. 193–236. Springer, ??? (2020)
38. Zhuang, H., Guo, F., Zhang, C., Liu, L., Han, J.: Joint aspect-sentiment analysis with minimal user guidance. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1241–1250 (2020)
39. Ekbal, A., Bhattacharyya, P.: Exploring multi-lingual, multi-task, and adversarial learning for low-resource sentiment analysis. Transactions on Asian and LowResource Language Information Processing 21(5), 1–19 (2022)
40. Sazzed, S.: Cross-lingual sentiment classification in low-resource Bengali language. In: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020), pp. 50–60
41. Sazzed, S.: Bengsentilex and bengswearlex: creating lexicons for sentiment analysis and profanity detection in low-resource bengali language. PeerJ Computer Science 7, 681 (2021)
42. Nankani, H., Dutta, H., Shrivastava, H., Rama Krishna, P., Mahata, D., Shah, R.R.: Multilingual sentiment analysis. Deep Learning-Based Approaches for Sentiment Analysis, 193–236 (2020)
43. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pretraining distributed word representations. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018) (2018)
44. Krippendorff, K.: Computing Krippendorff’s Alpha-Reliability. (2011)
45. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proceedings of the IEEE 109(1), 43– 76 (2020)
46. Zoph, B., Yuret, D., May, J., Knight, K.: Transfer learning for low-resource neural machine translation. arXiv preprint arXiv:1604.02201 (2016)
47. Bensoltane, R., Zaki, T.: Towards Arabic aspect-based sentiment analysis: A transfer learning-based approach. Social Network Analysis and Mining 12, 1–16 (2022)

48. Costa-juss`a, M.R., Cross, J., C¸elebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al.: No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672 (2022)

49. Loureiro, D., Barbieri, F., Neves, L., Anke, L.E., Camacho-Collados, J.: TimeLMs: Diachronic language models from Twitter. arXiv preprint arXiv:2202.03829 (2022)

50. Peng, C.-Y.J., Lee, K.L., Ingersoll, G.M.: An introduction to logistic regression analysis and reporting. The journal of educational research 96(1), 3–14 (2002)

51. Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D.: An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society 18(6), 275–285 (2004)

52. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)

53. Noble, W.S.: What is a support vector machine? Nature biotechnology 24(12), 1565–1567 (2006)

54. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)

55. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of documentation 28(1), 11–21 (1972)

56. Powers, D.M.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061 (2020)

57. Lemmens, J., Markov, I., Daelemans, W.: The LiLaH Emotion Lexicon of Greek,

58. Kurdish, Turkish, Spanish, Farsi and Chinese (2023)

59. Romera-Paredes, B. and Torr, P., 2015, June. An embarrassingly simple approach to zero-shot learning. In International conference on machine learning (pp. 2152-2161). PMLR.

60. Wang, Y., Yao, Q., Kwok, J.T. and Ni, L.M., 2020. Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur), 53(3), pp.1-34.

61. Berahmand, K., Daneshfar, F., Salehi, E.S. et al. Autoencoders and their applications in machine learning: a survey. Artif Intell Rev 57, 28 (2024). https://doi.org/10.1007/s10462-023-10662-6

62. Daneshfar, F., Soleymanbaigi, S., Nafisi, A. and Yamini, P., 2024. Elastic deep autoencoder for text embedding clustering by an improved graph regularization. Expert Systems with Applications, 238, p.121780.

63. Hussein, D.M. and Beitollahi, H., 2022. A Hybrid Deep Learning Model to Accurately Detect Anomalies in Online Social Media. Tikrit Journal of Pure Science, 27(5), pp.105-116.

64. Daneshfar, F., S. Soleymanbaigi, P. Yamini and M. S. Amini (2024). "A survey on semi-supervised graph clustering." Engineering Applications of Artificial Intelligence 133: 108215.

65. Daneshfar F, Barkhoda W, Azami BZ. Implementation of a Text-to-Speech System for Kurdish Language. In2009 Fourth International Conference on Digital Telecommunications 2009 Jul 20 (pp. 117-120). IEEE.