Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09716-6

ORIGINAL PAPER

KurdiSent: a corpus for kurdish sentiment analysis

Soran Badawi[1] · Arefeh Kazemi[2] · Vali Rezaie[2]

Accepted: 13 December 2023

Abstract

Language is essential for communication and the expression of feelings and sentiments. As technology advances, language has become increasingly ubiquitous in our lives. One of the most critical research areas in natural language processing (NLP) is sentiment analysis, which aims to identify and extract opinions and attitudes from text. Sentiment analysis is particularly useful for understanding public opinion on products, services, and topics of interest. While sentiment analysis systems are well-developed for English, this differs for other languages, such as Kurdish. This is because less-resourced languages have fewer NLP resources, including annotated datasets. To bridge this gap, this paper introduces KurdiSent, the first manually annotated dataset for Kurdish sentiment analysis. KurdiSent consists of over 12,000 instances labeled as positive, negative, or neutral. The corpus covers the Sorani dialect of Kurdish, the most widely spoken dialect. To ensure the quality of KurdiSent, the dataset was trained on machine learning and deep learning classifiers. The experimental results indicated that XLM-R outperformed all machine learning and deep learning classifiers, with an accuracy of 85%, compared to 81% for the best machine learning classifier. KurdiSent is a valuable resource for the NLP community, as it will enable researchers to develop and improve sentiment analysis systems for Kurdish. The corpus will facilitate a better understanding of public opinion in Kurdish-speaking communities.

Keywords Corpus annotation · Kurdish sentiment analysis · Machine learning · Deep learning

✉ Soran Badawi
soran.sedeeq@charmouniversity.org

[1] Language Center, Charmo Center for Research, Training and Consultancy, Charmo University, KRG, Chamchamal, Kurdistan, Iraq

[2] Department of Linguistics, University of Isfahan, Isfahan, Iran

Published online: 02 January 2024

1 Introduction

Sentiment analysis (SA) is a natural language processing (NLP) task that involves extracting, identifying, characterizing, and classifying text information to determine the attitudes and opinions of people (Beigi et al., 2016). It is commonly used to analyze online reviews of products, services, and other entities. Sentiment analysis is classified sentiment as positive, negative, or neutral. In general, SA is performed at four different levels of granularity, including document level, sentence level, aspect level, and concept level (Wankhade et al., 2022).

At the document level, SA takes the entire document as the primary unit of analysis to determine the general information about the document, such as whether the overall sentiment is positive or negative. This level is the most abstract level of SA and is unfit for obtaining precise evaluations (Bhatia et al., 2015). At the sentence level, SA aims to classify the sentiment of individual sentences within a document. It consists of two main tasks: subjectivity classification and polarity classification. Subjectivity classification identifies sentences that express opinions or viewpoints, while polarity classification determines whether the opinion is positive or negative (Rao et al., 2018). It is important to note that analyzing the sentiment of a document or sentence level is essential in many cases, but it does not provide some of the necessary information. For example, just because someone is optimistic about a particular entity does not mean they have a favorable opinion on all aspects of that entity. Similarly, negative sentiments do not necessarily mean the author has a negative impression of all aspects of the entity. That is why aspect-level opinion mining was introduced, which looks at the opinion itself instead of the language structures of documents, sentences, and phrases (Schouten & Frasincar, 2016). This level of opinion mining provides a more in-depth analysis of the target entity. The final level of sentiment analysis focuses on a semantic analysis of the text using web ontologies or semantic networks. Cambria et al. (2013) introduce the concept level of opinion mining as a new approach in Sentiment Analysis. This analysis of emotions at the concept level is based on conceptual information about emotion and sentiment associated with natural language (Birjali et al., 2021).

In sentiment analysis, there are two main approaches: Machine Learning (ML) and the lexicon-based approach. ML includes unsupervised and supervised learning, while lexicon-based methods use dictionary-based and corpus-based approaches. The ML approach involves converting annotated data into feature vectors and training machine learning classifiers to predict the class of unseen data using specific features. On the other hand, the lexicon approach relies on extracting and calculating the polarities of sentiment lexicons using Sentiment Lexicon (Wankhade et al., 2022). It is important to note that this research is based on the ML approach.

This paper focuses on sentiment analysis for the Kurdish language, a less-resourced Indo-European language spoken by over 40 million speakers (Badawi, 2023b). The Kurdish language is part of the Indo-Iranian family of Indo-European languages and has 33 letters. It is similar to Persian and is spoken in Iran, Turkey, Iraq, and Syria. Interestingly, it is even one of the official languages in Iraq. There are two main dialects, Central Kurdish (Sorani) and Northern Kurdish (Kurmanji), as well as more minor dialects like Gorani (Hawrami), used by small communities in Iraq and

Iran, and Zazaki, predominantly practiced in Turkey (Badawi et al., 2023). Using the Sorani Dialect, we attempted to build a sentiment analysis system for the Kurdish language with the first-ever sentiment analysis annotated corpus in which human annotators annotated all the data. Moreover, we trained our dataset on state-of-the-art machine learning classifiers and deep learning classifiers and reported the results.

The remainder of the presented work is organized in the following style: Sect. 2 comprehensively presents related research, providing critical background and illumination into the existing literature. Section 3 delves into the dataset description and the methodology of our study. Section 4 showcases our experimental results and explores their implications to comprehend the findings comprehensively. In conclusion, we summarize the primary findings and their implications in the concluding section.

2 Related work

This section provides an overview of various corpus construction and annotation methodologies utilized in English and multiple languages, highlighting their contributions to the study of natural language processing. Specifically, the Sanford Sentiment Treebank (SST) stands out as a premier corpus that employs a comprehensive, fully labeled parse tree approach to examine the compositional effects of sentiment in language. Composed of 11,855 individually labeled sentences extracted from movie reviews, the SST was annotated by three expert human evaluators (Jiménez-Zafra et al., 2020). Another noteworthy dataset is the Product Review corpus, introduced in 2010, which consists of 2,111 sentence extracts selected from 268 product comments on Google Product Search (Kessler & Kuhn, 2014). The annotations primarily relied on syntactic negation cues, with two experienced annotators participating in the labeling process. The Simon Fraser University Review (SFU) corpus is a collection of 400 review texts from the Epinions.com platform, covering eight domains: books, cars, computers, cookware, hotels, films, music, and phones. Notably, two skilled linguists were involved in the annotation process, with one annotator responsible for labeling the entire corpus. In contrast, a second annotator independently labeled 10% of the randomly chosen documents to assess inter-annotator concurrence (Kolhatkar et al., 2020a). Assessment of inter-annotator agreement yielded a kappa agreement value of 0.872 (Jiménez-Zafra et al., 2020). The annotation methodology was based on the guidelines established for the BioScope corpus, with some adjustments. The SFU ReviewEN corpus is publicly accessible in XML format, licensed under the GNU General Public License. It is a valuable resource for researchers investigating negation and speculation in natural language processing. The PropBank Focus (PB-FOC) corpus concentrates on negation in the English language. It comprises a collection of texts created by two researchers who analyzed the Wall Street Journal portion of the Penn Treebank, identifying 3,993 verbal negations in 3,779 sentences. The corpus features a novel element for annotating negation, performed by two graduate students in computational linguistics. The inter-annotator agreement between the two annotators was 72%, with the remaining instances annotated following disagreement resolution (Banjade & Rus, 2016). The SFU Opinion and Comments Corpus (SOCC) also constitutes a substantial dataset of opinion articles and comments from

The Globe and Mail, a prominent Canadian newspaper. Encompassing five years, it contains over 10,000 articles and 663,000 comments from 303,000 comment threads. Furthermore, a subset of 1,043 comments was annotated with three layers: constructiveness, appraisal, and negation. Two expert annotators utilized meticulous guidelines and Webanno software to label negation cues, scope, xscope, and focus. Inter-annotator agreement was calculated using 50 comments from the beginning and end of the annotation process, demonstrating high agreement rates for keywords, scope, and focus (Kolhatkar et al., 2020b). In conclusion, the SOCC corpus offers a valuable resource for studying online discourse and communication dynamics. Its carefully designed structure, thorough annotation, and open accessibility render it an exceptional tool for expanding knowledge within this field.

In the realm of Spanish language corpora, several datasets have been annotated. The UAM (Universidad Aut6noma de Madrid) Spanish Treebank is identified as one of the earliest Spanish corpora annotated. The dataset includes 1500 sentences from newspaper articles such as El Pas Digital and Compra Maestra. Two linguistic experts annotated the corpus (Kolhatkar et al., 2020b). The second Spanish dataset is IxaMed-GS. The data was collected from 75 outpatient consultation records at Galdakao-Usansolo Hospital in Biscay, Spain. Two experts completed the annotation process (Oronoz et al., 2015). The SFU ReviewSP-NEG is a Spanish language corpus designed explicitly for the annotation of negation, filling a gap in existing resources. It includes 400 reviews from 8 domains, including automobiles, hotels, washing machines, literature, mobile devices, music, personal computers, and cinematic productions. Each domain boasts 50 reviews apiece, with positivity or negativity determined by the cumulative star rating bestowed upon the text by the reviewer. Texts receiving one or two stars are designated negative, whereas those garnering four or five stars are deemed positive. By ignoring texts with three stars, the researchers behind the corpus have effectively sidestepped potential ambiguities arising from ambivalently rated reviews. It is important to note that the annotators formed their annotation guidelines by considering the issues that should be considered vital for the design of a corpus. The UHU-HUVR corpus is a collection of 604 clinical reports from the Virgen del Rocío Hospital in Seville, Spain, annotated for negation, syntax, morphology, and lexicon. The corpus includes both parallel radiology reports and personal history of anamnesis reports transcribed in free text. Two domain experts annotated the corpus using the Thyme corpus guidelines, with some adaptations, resulting in 1,079 sentences in CoNLL format being identified as containing negations out of 3,065 sentences in the anamnesis reports and 1,219 sentences (22.80%) out of 5,347 sentences were annotated with negations in the radiology reports (Jiménez-Zafra et al., 2020). The inter-annotator agreement for the corpus was between 0.70 and 0.95, with most disagreements attributed to human error, such as missing or misplacing words. The UHU-HUVR corpus represents a significant step forward in the study of negation detection and sentiment analysis, and its future availability will undoubtedly provide valuable resources for researchers in these fields.

In Arabic, ASTD (Arabic Sentiment Tweets Dataset) is a collection of 10,000 tweets labeled in four classes: favorable, unfavorable, mixed, and objective. Three annotators labeled the corpus. During the process, tweets at least two annotators agreed upon to possess a specific label were regarded as conflict-free and accepted for further processing. Other tweets that caused conflict among all three raters were ignored (Abo et al., 2019). ArSEntD-LEV Corpus is a Levantine dialect sentiment corpus proposed in 2019. The corpus contains 4000 tweets classified under five categories: very positive, positive, neutral, negative, and very negative. The content of the tweets primarily expresses opinions about personal and daily matters, and a small percentage of the tweets relate to political issues, especially the ongoing conflicts in the Middle East and religious matters, which mainly include quote verses from the holy book Quran. The researchers recruited 5–9 different annotators to annotate each tweet, which is reasonable to perform an aggregation over five classes. The final label for the tweet was decided based on the majority voting(Baly et al., 2019). Recent advancements in natural language processing have led to a surge of interest in detecting user sentiment in texts. This task has gained significant attention due to the proliferation of social media platforms and the increasing number of users engaging on these platforms. As a result, various Arabic sentiment datasets have been compiled, including those presented by Abdul-Mageed et al. (2014), Aly and Atiya (2013), Rushdi-Saleh et al. (2011), Refaee and Rieser (2014), Abdul-Mageed and Diab (2014), Ibrahim et al. (2015), ElSahar and ElBeltagy (2015), Nabil et al. (2015), and ElSahar and El-Beltagy (2015). These datasets vary in size, domain, and genre, providing diverse resources for sentiment analysis tasks. Abdul-Mageed et al. (2014) proposed the SAMAR system, which performs subjectivity and sentiment analysis on Arabic social media texts. They utilized multi-domain datasets from Wikipedia TalkPages, Twitter, and Arabic forums(Abdul-Mageed & Diab, 2012). Aly and Atiya (2013) introduced LABR, a book reviews dataset gathered from GoodReads (Aly & Atiya, 2013). Rushdi-Saleh et al. (2011) presented an Arabic corpus of 500 movie reviews obtained from various web pages (Rushdi-Saleh et al., 2011). Refaee and Rieser (2014) developed a manually annotated Arabic social corpus of 8,868 Tweets and discussed the corpus collection and annotation methods. In addition to these datasets (Refaee & Rieser, 2014), Abdul-Mageed & Diab, (2014) proposed SANA, a large-scale, multi-domain, and multigenre Arabic sentiment lexicon. SANA automatically extends two manually collected lexicons, HUDA (4,905 entries) and SIFFAT (3,325) (Abdul-Mageed & Diab, 2014). Ibrahim et al. (2015) created a manual corpus of 1,000 tweets and 1,000 microblogs for sentiment analysis tasks. ElSahar and ElBeltagy (2015) introduced four datasets in their work towards building a multi-domain Arabic resource for sentiment analysis. Nabil et al. (2015) and ElSahar and El-Beltagy (2015) proposed semi-supervised methods for constructing sentiment lexicons that can be effectively utilized in sentiment analysis (ElSahar & El-Beltagy, 2015).

In the realm of other languages, the EMC Dutch corpus is a collection of clinical texts, specifically comprising entries from general practitioners, specialists' letters, radiology reports, and discharge letters. It was assembled by Afzal et al. (2014) and featured 6740 texts. In order to annotate medical terminology within the corpus, the researchers referenced the Unified Medical Language(Afzal et al., 2014). The corpus was annotated with negation, and the identified terms were annotated for their negation, temporality, and experience properties. Concerning recognizing the negation, a term is labeled as 'Negated' if there is clear evidence in the text to denote that the condition does not occur or exist; otherwise, it is annotated as 'Not negated.'

The corpus constructors selected two independent annotators to annotate the corpus, and their differences were resolved by an expert with a medical background in the four types of clinical texts. The annotators were provided with annotation guidelines, which included information on explaining the process, and each contextual property was provided. However, the guidelines are not available. The kappa inter-annotator agreement for negated terms was 0.90, 0.90, 0.93, and 0.94 for entries from general practitioners, specialists' letters, radiology reports, and discharge letters, respectively. The percentage of negated terms is similar for the different report types: This corpus is deemed the first publicly available Dutch clinical corpus; however, it cannot be accessed online. It is required to email the authors to request the corpus. Additionally, a subset of the Stockholm Electronic Patient Record corpus was annotated by Dalianis et al. (2009) concerning certain uncertain expressions and speculative and negation keywords (Dalianis et al., 2009). Annotating the corpus involved five individuals: three senior-level students, one undergraduate computer scientist, and one undergraduate language consultant. Guideline development benefited from the BioScope corpus guidelines. Inter-annotator agreement was evaluated using a pairwise F-measure, yielding a score of 0.80. Notably, the annotation process focused solely on syntactic negation, while other aspects should have been considered. The corpus is formatted in XML. Qian et al. (2016) developed the Chinese Negation and Speculation (CNeSp) corpus, consisting of three document types annotated with negative and speculative cues and their linguistic scopes. This corpus contains 19 scientific articles, 821 product reviews, and 311 financial articles, totaling 16,841 sentences, of which 4,517 (26.82%) exhibit negation (Qian et al., 2016). The annotation guidelines followed the BioScope corpus's guidelines, with minor adjustments tailored to the Chinese language. Two experts carried out annotation, and disputes were settled by a linguistics specialist who revised the guidelines accordingly. The resulting inter-annotator agreements were 0.96, 0.96, and 0.93 for detecting negation cues and 0.90, 0.91, and 0.88 for identifying scope in scientific literature, financial articles, and product reviews. The corpus only accounted for lexical and syntactic negation(Jiménez-Zafra et al., 2020). RuSentiment is a novel dataset encompassing sentiment analysis of social media posts in Russian, complemented by a newly devised set of exhaustive annotation guidelines readily applicable to other languages. As the most extensive dataset for Russian, RuSentiment proudly showcases 31,185 posts, each assiduously annotated thrice, culminating in a commendable Fleiss' kappa score of 0.58. Employing an active learning technique, a deliberate selection process was implemented, yielding 6,950 posts contributing to the dataset's heterogeneity (Rogers et al., 2018). In Lindén et al. (2023), a 27,000-sentence dataset with sentiment polarity which is named FinnSentiment was annotated independently by three native annotators (Lindén et al., 2023).

In the Kurdish language, the number of annotated datasets in the field of sentiment analysis is alarmingly small. To our best Knowledge, there are four annotated corpora. The first one is the medical corpus. It contains 6756 samples collected from Facebook comments, divided between medical and non-medical classes (Saeed et al., 2022). The corpus was manually annotated, and information about the annotation process was not provided. The second one is KDC-4007. It includes 4,007 text files divided into eight categories: Sports, Religions, Arts, Economics, Education, Socials, Styles, and Health (Rashid et al., 2018). The information about the annotation process regarding this corpus is also not available. Moreover, Awlla and Veisi (2022) presented a dataset comprising 14,881 comments gathered from multiple Facebook pages. To develop a sentiment analyzer, they employ Word2vec embeddings in conjunction with a recurrent neural network classifier, achieving a reported accuracy of 71.35% (Awlla & Veisi, 2022). Badawi (2023a, b, c) explored the detection of headline news in Kurdish, assembling a corpus from several news websites and subsequently annotating and evaluating it utilizing various traditional machine learning algorithms and deep learning augmentation tools (Badawi, 2023a).

Through comprehensive research, two crucial discoveries have come to light. Firstly, languages like English, Arabic, and Spanish have made remarkable strides in the realm of sentiment analysis by producing abundant datasets. Secondly, the Kurdish language needs more datasets that have undergone scientific analysis. The existing Kurdish corpora are either inaccessible or need adequate information about the data collection and annotation processes, which raises concerns about their dependability. To bridge this gap, this paper introduces the first openly accessible sentiment analysis dataset annotated by human annotators. This corpus could significantly benefit the scientific community by offering a reliable dataset to advance sentiment analysis in the Kurdish language.

3 Data collection

The acquisition of data requires deploying a software application or integrating specialized libraries via code composition. In this study, we adopted the latter approach, leveraging the Twitter Developer Tool to retrieve tweets from the Twitter Developer API. As part of our commitment to upholding ethical standards and adhering to Twitter's guidelines, we anonymized users' identities in the collected tweets. Moreover, we applied random sampling techniques to select tweets disseminated during the COVID-19 pandemic.

Undeniably, raw data encompasses a substantial amount of noise, particularly in Kurdish data extracted from online sources. This noise manifests itself in various forms, including words in foreign languages, special characters, elongation (character repetition), symbols, and ineffectual numerals. To address this issue, our preprocessing methodology incorporates a series of stages to filter out irrelevant information. During the initial stage, we eliminate all non-Kurdish characters in HTML links. This is followed by a second stage that focuses on special characters. Our analysis reveals that certain characters utilized to convey emotions were not removed in this stage. However, emotions expressed using incomprehensible symbols such as ":)" were eliminated from the dataset. In order to maintain the integrity of the sentiment expressed, characters that hold emotional significance are retained in their original form. Conversely, characters devoid of meaning are eliminated from the text. It is worth mentioning that numerical values can also convey sentiment or feelings in Kurdish texts. Therefore, we opt to include them in the dataset. Another distinct feature of Kurdish is its use of elongation, which involves extending letters by employing a specific symbol (-). A representative example of elongation in Kurdish words

Table 1 An example of elongation in the Kurdish language

	NO elongation	3-times elongation
Kurdish text	خواردنەکە خۆشە	خواردنەکە خۆۆۆ شە
Transliteration	Xwardneke xoşa	
Translation	The food is tasty	

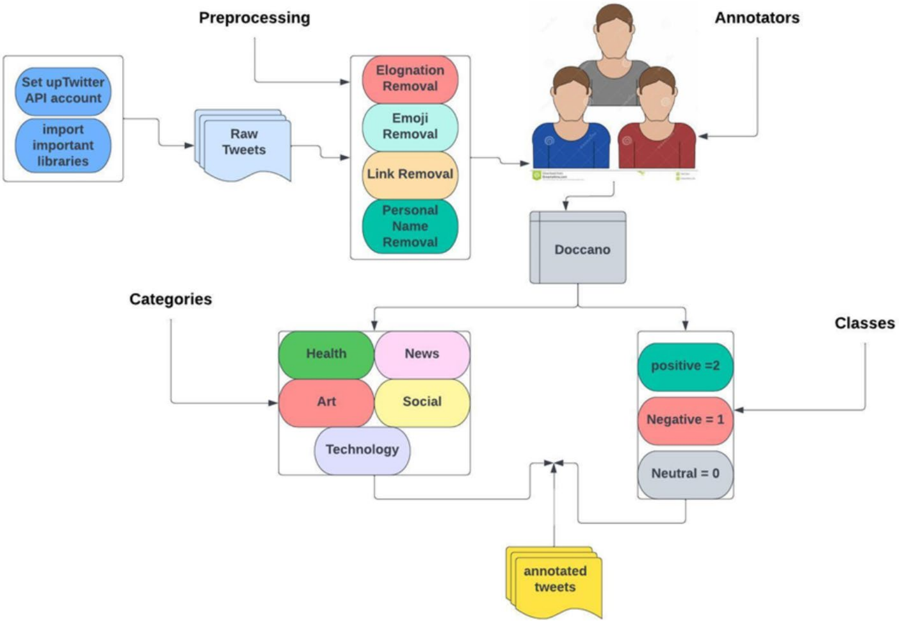


Fig. 1 The steps taken for building KuriSent

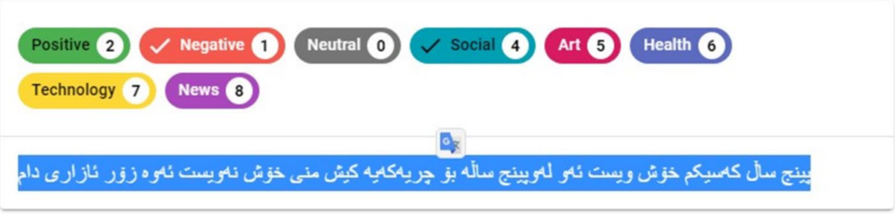
is presented in Table 1. When a character is repeated more than once, it is condensed into a single letter.

In summary, our preprocessing technique is designed to tackle the unique characteristics of Kurdish data, ensuring that the resulting dataset is refined and conducive to sentiment analysis. By meticulously removing unnecessary elements and preserving those that contribute to the expression of sentiment, we can develop a more accurate and efficient sentiment analysis model. The resulting dataset is publicly accessible through the Mendeley repository, bearing the DOI: https://doi.org/10.17632/dntxt73dm6.1. The methodological framework underpinning the construction of the dataset is depicted in Fig. 1.

3.1 Annotation process

Three individuals with academic backgrounds in the Kurdish language were selected to annotate the corpus. The annotators were provided with guidelines to facilitate the annotation process, which included the following tasks:

- Classifying the sentiment of each text as either positive, negative, or neutral.



پینج ساڵ کەسیکم خوش ویست ئەو لەوپینج ساڵە بو چریەکەیە کیش منی خوش نەویست ئەوە زور ئازاری دام

Fig. 2 Annotation using Doccano

Table 2 Annotators agreement

Annotated Agreement	Rater 1 & Rater 2	Rater 1 & Rater 3	Rater 2 & Rater 3
Kappa Coefficient	0.89	0.78	0.87

- Assigning a category to each text.
- Recording the significant votes for each piece of text and integrating them into the corpus.

To support the annotation process, we utilized Doccano, an open-source text annotation tool. First, the corpus was uploaded to the tool's platform. Next, the labels and categories were defined for the annotators. Doccano presents each text on a separate page, allowing the annotators to work efficiently, as depicted in Fig. 2. Upon completion of the annotation process, the corpus was exported in JSON format. It is worth stating that various indispensable tools are available online for annotating corpora; however, we chose to exploit the features offered by Doccano to streamline the annotation process. Its user-friendly interface and versatile functionality made it an ideal choice for our research endeavor.

The annotators were provided with the complete dataset, enabling them to execute their duties effectively. To assess the degree of concurrence among the annotators, we computed Kappa coefficients (Cao et al., 2016) that revealed a substantial harmony in the sentiment annotation process. The outcomes, presented in Table 2, demonstrate a range of 0.78 to 0.89, which falls within the acceptable limits established by prior studies. This outcome signifies a satisfactory level of consistency among the annotators, bolstering the reliability of our findings.

While annotating the dataset, we confronted several challenges requiring careful consideration and innovative solutions. The initial challenge revolved around identifying and securing the services of skilled annotators with expertise in the Kurdish language and the nuances of sentiment analysis. This quest proved time-consuming, as qualified candidates were frequently engaged in other commitments. Another significant hurdle was the inherent subjectivity and ambiguity associated with sentiment analysis. The interpretation of sentiments varied among annotators, and annotating sarcastic tweets related to COVID-19 proved especially problematic. In the initial data collection phase, we gathered a total of 15,694 tweets for annotation. Following the annotation process, we retained 12,309 tweets, excluding those that sparked substantial disagreement among annotators, thereby ensuring the accuracy and reliability of our dataset. Furthermore, we faced the additional challenge of inadequate resources for the Kurdish language. No pre-existing sentiment lexicons or clearly defined annotation guidelines existed, obliging us to develop these fundamental resources from scratch. This undertaking demanded a sizeable investment of time and effort. Moreover, the rich diversity of dialectical variations in Kurdish introduced yet another layer of complexity into the annotation process. Users from diverse regions of Kurdistan employed distinct lexicons, dialects, and idiomatic expressions, making annotation increasingly challenging. To surmount these obstacles, we collaborated closely with native speakers and linguistic experts from various geographic locations across Kurdistan. Additionally, we developed meticulous annotation guidelines explicitly tailored to the Kurdish language. We successfully created a dependable and robust sentiment analysis dataset for the Kurdish language by embracing a comprehensive approach that addressed the singular difficulties of sentiment annotation. Our bespoke methodology enabled us to navigate the complexities of working with a lesser-resourced language, yielding a valuable resource for future research endeavors in this domain.

3.2 Dataset statistics

The underlying structure of the dataset consists of over 12,000 texts distributed evenly across three primary classifications: positive, negative, and neutral. These texts are subdivided into five categories: social, art, health, technology, and news. Notably, each category harbors a diverse array of texts spanning all three labels, facilitating a comprehensive data representation, as provided in Table 3. Social texts examine the complexities of human relationships, community dynamics, and critical societal matters. Art texts explore the realm of creative expression, artistic critique, and cultural analyses. Health texts offer insight into medical topics, wellness, and preventative strategies for various diseases. Technology encompasses various subjects, including hardware, software, artificial intelligence, and the latest digital innovations. Finally, our news category includes an array of reports and articles about local, national, and international events, politics, economics, and business.

These categories are distributed unevenly throughout the dataset, reflecting the relative importance of each one. The Social category dominates the dataset, accounting for 45% of all texts. In contrast, the Art category makes up only 13% of the dataset as presented in Fig. 3. It encompasses creative expressions, artistic critiques, and cultural analyses. While not as prominent as the Social category, the Art category still plays a role in the dataset, albeit a smaller one.

Table 3 Categories in KurdiSent

Label	Categories					Total-label
	Social	Art	Health	Technology	News	
Positive	2486	665	116	40	796	4103
Negative	1570	546	571	67	1349	4103
Neutral	1408	548	37	52	2058	4103
Total-category	5464	1759	724	235	4203	12,309

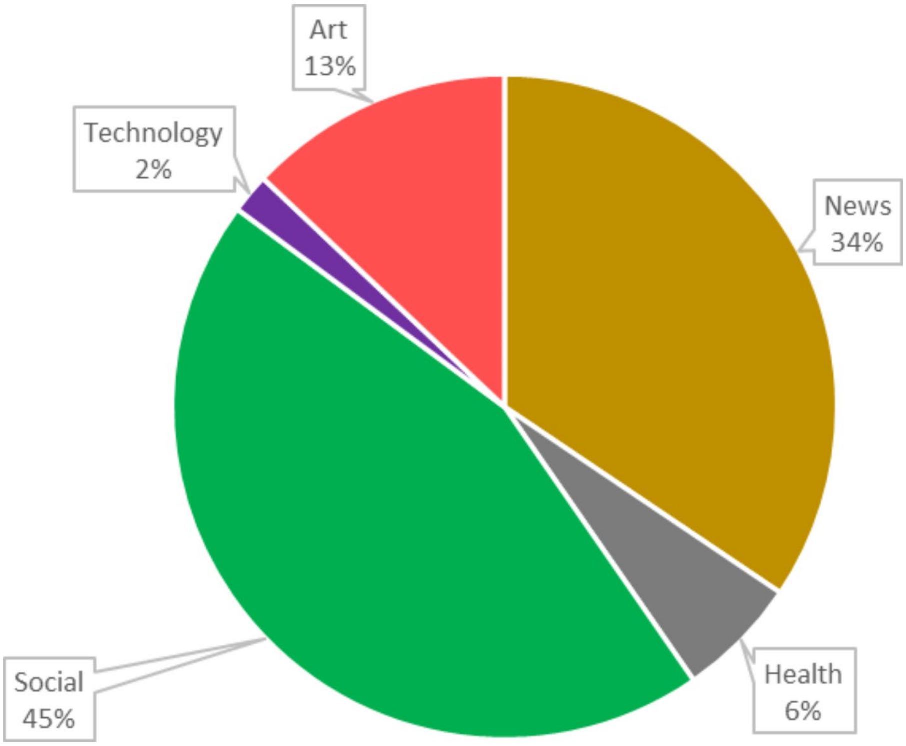


Fig. 3 The distributions of categories in KurdiSent

The Health category comprises 6% of the dataset and covers medical topics, wellness, and disease prevention strategies during Covid-19. The Technology category, with a mere 2% share, includes texts related to hardware, software, artificial intelligence, and digital innovations. Lastly, the News category makes up 34% of the dataset, consisting of reports and articles about local, national, and international events, politics, economics, and business. As the second-largest category, it reinforces the dataset's emphasis on analyzing sentiment in informational and factual contexts. In summary, the KurdiSent dataset strives to offer a comprehensive mix of categories, placing particular attention on social and news-related texts. This diversity enables the development of sentiment analysis models adaptable to various applications, including social media monitoring, customer feedback analysis, and news article summarization.

4 Experiments

The application of sentiment analysis to text employs a machine-learning methodology, which necessitates the existence of an annotated corpus. Before proceeding with the analysis, it is essential to preprocess the data within the corpus, thereby removing any superfluous information and preparing it for the subsequent stages. Once the data has been suitably prepared, the feature construction process commences. The selection of features plays a critical role in determining the efficacy of the machine learning classifiers that will be employed later in the process. Following the completion of feature selection, the classifiers are trained and evaluated using a chosen set of features, as depicted in Fig. 4.

To ensure the proper functioning of machine learning algorithms, it is necessary to represent textual data in a suitable format for the classifier. One widely recognized approach for achieving this goal is the feature or vector model, which transforms texts or documents into features that the classifier can effectively process. This model preserves the fundamental information about the text, enabling the classifier to make informed decisions. Our research employed the Bag-of-the-Word (BOW) feature model as a baseline model for traditional machine learning classifiers. BOW represents each document as a vector where each element corresponds to the presence or absence of a particular word in the vocabulary. This approach allows for the efficient representation of textual data and enables the classifier to focus on the most relevant features when making predictions. We performed this experiment in different classification types and levels and on different domains and ML classifiers.

In order to develop a proficient sentiment analysis model for the Kurdish language, we opted to utilize the most recent advancements in the field of sentiment analysis. Specifically, we chose to employ various state-of-the-art models, including Naïve Bayes (Chakraborty et al., 2017), SVM(Ahmad et al., 2017), Logistic Regression(Ramadhan et al., 2017), Decision Tree (Bayhaqy et al., 2018), Random Forest (Fauzi, 2018), KNN(Huq et al., 2017), CNN-LSTM (She & Zhang, 2018), Bi-LSTM with attention(Almars, 2022), CNN-RNN (Basiri et al., 2021), and BERT (Badawi, 2023a, b, c). Moreover, We used the XLM-R large model, which has 355 M parameters, 24 layers, 1,027 hidden states, 4,096 feed-forward hidden states, and 16 heads. It can take an input of a sequence of no more than 512 tokens and outputs the representation of the sequence (Kumar & Albuquerque, 2021). We also use mT5, a massively multilingual version of the T5 model. It covers more than 100 languages, so its vocabulary is larger (Xue et al., 2021). These models were selected based on their reputation as cutting-edge techniques for detecting cyberbullying in social media. Furthermore, we followed the original papers' setup parameters when implementing these baseline models. Throughout the experimental phase, we leveraged popular libraries such as Keras, TensorFlow, NumPy, NLTK, Scikit-learn, etc. We partitioned the input dataset into training and testing sets to conduct the experiments. Utilizing Google Colab, an online platform equipped with a powerful GPU, we executed the experimentation. Our code was written in Python 3.9, and we operated on a capable personal computer running an advanced OS and processor.

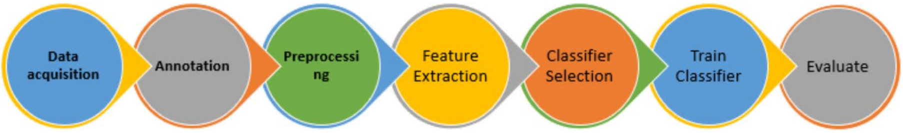


Fig. 4 Machine learning process

Table 4 Accuracy score

Classifiers	Accuracy
KNN	0.40
Decision Tree	0.74
BI-LSTM (ATTN)	0.77
Naive Bayes	0.78
Random Forest	0.79
CNN-LSTM	0.80
SVM	0.81
Logistic Regression	0.81
CNN-RNN	0.81
BERT	0.83
MT5	0.84
XLM-R	**0.85**

Table 5 Precision score

Classifiers	Positive	Neutral	Negative
BI-LSTM (ATTN)	0.80	0.78	0.86
Naive Bayes	0.83	0.67	**0.89**
Decision Tree	0.84	0.67	0.72
CNN-LSTM	0.88	0.75	0.84
KNN	0.90	**0.87**	0.4
BERT	0.90	0.74	0.87
CNN-RNN	0.91	0.73	0.82
Logistic Regression	0.92	0.70	0.87
MT5	0.92	0.76	0.86
XLM-R	0.93	0.80	0.84
Random Forest	**0.94**	0.67	0.77
SVM	**0.94**	0.68	0.88

5 Results and discussions

We applied machine learning algorithms to determine our corpus's sentiments to provide a simple baseline. We distributed our data into train and test. However, the distribution is performed so that each set contains equal labels. The reason behind this was to avoid the case of biasedness by the classifiers. The results each classifier achieves are shown in Tables 4, 5 and 6.

Table 4 displays the accuracy scores of several machine learning models trained on the KurSent dataset, a collection of text data in the Kurdish language. The table shows that the XLM-R model attained the highest accuracy score among all the models, with a value of 0.85. This suggests that XLM-R performed best in classifying sentiments in Kurdish texts. Following XLM-R, the best performers were MT5, BERT, CNN-LSTM and SVM, with an accuracy score of 0.84, 0.83 and 0.81 respectively. The other models had lower accuracy scores, ranging between 0.74 and 0.79. The worst-performing model was KNN, with an accuracy score of only 0.40. Overall, the results indicate that the more advanced models, such as XLM-R, MT5, BERT, CNN-LSTM, and SVM, tend to perform better in sentiment classification

Table 6 Recall score

Classifiers	Positive	Neutral	Negative
KNN	0.28	0.20	**0.89**
Random Forest	0.74	0.82	0.76
Naive Bayes	0.76	0.83	0.76
SVM	0.77	**0.90**	0.77
Logistic Regression	0.79	0.88	0.78
Decision Tree	0.79	0.67	0.76
CNN-RNN	0.80	0.83	0.84
Bi-LSTM (ATTN)	0.81	0.76	0.82
CNN-LSTM	0.82	0.80	0.86
BERT	0.83	0.84	0.82
MT5	0.84	0.87	0.83
XLM-R	**0.85**	0.89	0.86

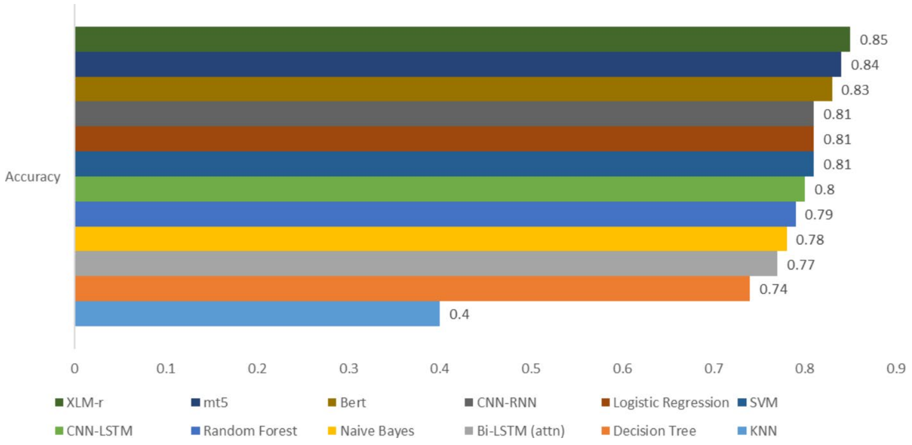


Fig. 5 Accuracy performance

tasks involving Kurdish language texts than simpler models like KNN and decision trees as displayed in Fig. 5.

The Table 5 displays the precision scores of the machine learning and deep learning models trained on the KurSent dataset, which contains text data in the Kurdish language. The models are evaluated based on their ability to predict positive, neutral, and negative sentences. Firstly, it's clear that most models struggle with the neutral class, which could be due to the imbalanced nature of the dataset. The number of positive and negative samples is significantly higher than the number of neutral samples. Despite this challenge, XLM-R emerges as the top performer across all classes, with precision scores of 0.85, 0.89, and 0.85 for positive, neutral, and negative classes, respectively. This suggests that XLM-R is particularly effective in capturing the subtleties of sentiment in Kurdish language texts. When comparing the performance of the other models, we notice that some baseline models, such as Naive Bayes and Logistic Regression, perform relatively well. In fact, Naive Bayes achieves a higher precision score than Decision Tree and Random Forest for the positive class. However, the more advanced models generally outperform the baseline models, indicating that they are better equipped to handle the complexities of sentiment classification. Each model has its unique strengths and weaknesses. For instance, Random Forest excel at identifying positive sentences but struggle with neutral ones. On the other hand, SVM performs well on neutral sentences but falters in classifying positive and negative sentences. Models that incorporate attention mechanisms, such as XLM-R, MT5, BERT and BI-LSTM (attn), tend to perform better overall. This highlights the importance of focusing on specific parts of the input data when making predictions. Additionally, CNN-based models, including CNN-LSTM and CNN-RNN, demonstrate competitive performance across all classes. Their ability to capture local and global context through convolutional filters works effectively for sentiment classification in Kurdish language texts. Lastly, KNN, a simple machine learning algorithm, performs poorly across all classes. This may be due to the lack of robust features extracted from the text data, leading to unreliable predictions (See Fig. 6).

The recall scores for the models trained on the KurdSent dataset can be analyzed as follows: Naive Bayes has a recall score of 0.76 for the positive class, meaning that the model correctly predicted 76% of the actual positive instances. The recall scores for the neutral and negative classes are 0.83 and 0.76, respectively. Logistic Regression performs slightly better than Naive Bayes, with a recall score of 0.79 for the positive class. The recall scores for the neutral and negative classes are 0.88 and 0.78, respectively. Decision Tree has a recall score of 0.79 for the positive class, similar to that of Logistic Regression. The recall scores for the neutral and negative classes are 0.67 and 0.76, respectively. Random Forest has a recall score of 0.74 for the positive class, which is lower than that of Decision Tree. The recall scores for the neutral and negative classes are 0.82 and 0.76, respectively. KNN has a deficient recall score of 0.28 for the positive class, indicating that the model correctly predicted only 28% of the actual positive instances. The recall scores for the neutral and negative classes are 0.20 and 0.98, respectively. SVM has a recall score of 0.77 for the positive class, slightly higher than that of Random Forest. The recall scores for the neutral and negative classes are 0.90 and 0.77, respectively. BERT performs better, with a recall score

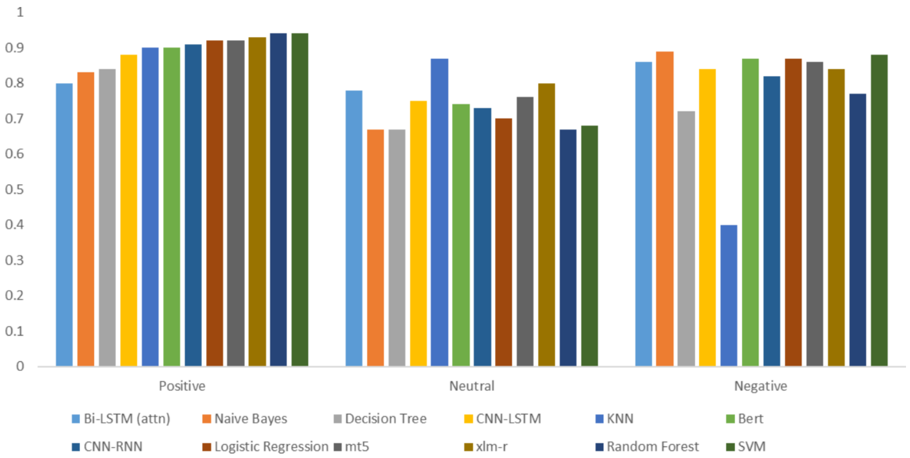


Fig. 6 Precision performance

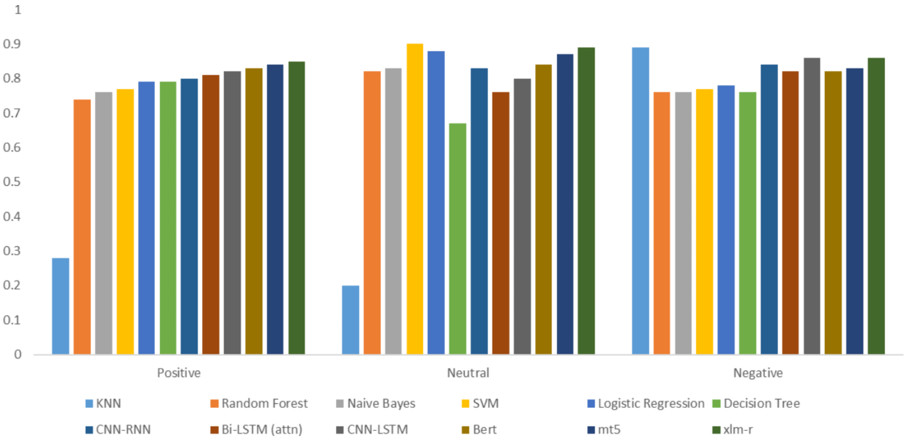


Fig. 7 Recall performance

Table 7 F1_ score

Classifiers	Positive	Neutral	Negative
KNN	0.42	0.33	0.57
Bi-LSTM (ATTN)	0.78	0.75	0.83
Naive Bayes	0.80	0.74	0.82
Decision Tree	0.81	0.67	0.74
Random Forest	0.83	0.74	0.77
CNN-LSTM	0.83	0.77	0.83
SVM	0.84	0.77	0.82
CNN-RNN	0.84	0.76	0.81
Logistic Regression	0.85	**0.78**	0.82
MT5	0.84	0.77	0.82
BERT	0.86	0.78	0.84
XLM-R	**0.88**	0.77	0.85

of 0.83 for the positive class as illustrated in Fig. 7. The recall scores for the neutral and negative classes are 0.84 and 0.82, respectively. CNN-LSTM and CNN-RNN have recall scores of 0.82 and 0.80, respectively, for the positive class. The recall scores for the neutral and negative classes are 0.80 and 0.86, respectively, for CNN-LSTM and 0.83 and 0.84 for CNN-RNN. Finally, BI-LSTM (attn) has a recall score of 0.81 for the positive class, slightly lower than that of BERT. The recall scores for the neutral and negative classes are 0.76 and 0.82, respectively. The recall scores for MT5. are 0.84,0.87, 0.83 for positive, neutral and negative classes respectively. Notably, XLM-R outperforms all of the models by scoring 0.85,0.89,0.86 indicating that the attention-based models work well with Kurdish texts.

Table 7 shows the F1 scores for each classifier on the three classes of sentiment: positive, neutral, and negative. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of both. The results indicate that most classifiers perform well in classifying sentiments in the Kurdish language. In particular, XLM-R, MT5, BERT, SVM, and deep learning models (CNN-LSTM and CNN-RNN)

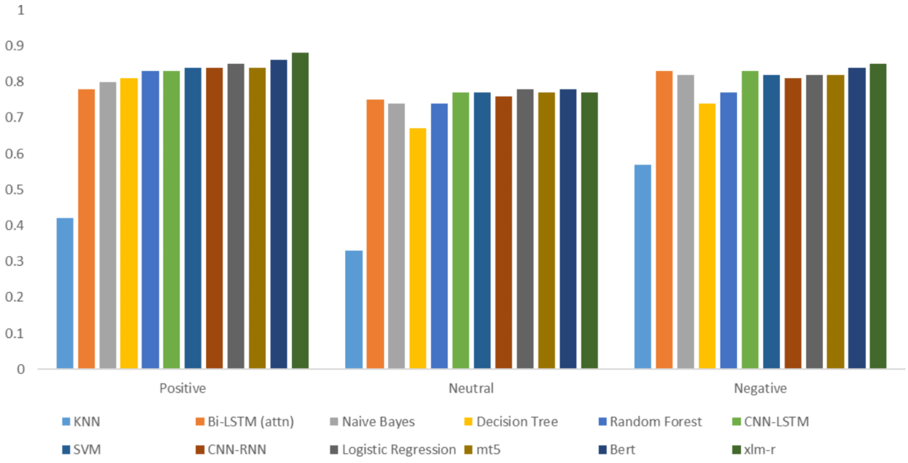


Fig. 8 F1 performance

achieve high F1 scores for all three classes as displayed in Fig. 8. These models are able to effectively capture the nuances of sentiment expression in the Kurdish language. On the other hand, KNN performs poorly across all classes, suggesting that this model may not be suitable for sentiment classification tasks. Overall, the results provide insights into the strengths and weaknesses of different classifiers for sentiment analysis in the Kurdish language. Researchers and practitioners can develop more accurate and efficient sentiment analysis systems for this language by selecting the best-performing models.

6 Conclusion

The Kurdish language does not have any large annotated dataset to study sentiment analysis in the language. We have constructed a new dataset for Kurdish language speakers to perform numerous processes, particularly sentiment analysis. We have also evaluated our dataset on multiple machine learning classifiers and primary deep learning techniques. Attention-based models such as XLM-R, MT5, and BERT provided a significant accuracy score compared to the classical classifiers. Our result showed that machine learning classifiers and deep learning models work well with the dataset. Hopefully, the establishment of this dataset can open numerous opportunities for Kurdish scholars in the scientific community to incorporate the dataset to extensively study the language and develop state-of-the-art models of sentiment analysis for the Kurdish language.

Author contributions Author A was responsible for collecting the data and composing the first draft. Author B was responsible for doing the experimental section of the article. Author C's task involved writing the discussion and reviewing the article.

Funding There was no funding support for this article.

Data availability The corpus can be downloaded on the Mendeley repository for free using this Doi (https://doi.org/10.17632/3yrkswy6ph.2).

Declarations

Ethics declarations Regarding the Terms of service (ToS), the dataset presented in this paper has been compiled using Twitter's services per Twitter's Terms of Service. All data presented in this paper were retrieved officially through Twitter APIs per Twitter's Developer Agreement and Policy. We confirm that we are not collecting data to make money (e.g., business), to deliver DDoS attacks, to steal data or for any other sinister intentions. Twitter's copyright policy states that widespread entities may own tweets. In this dataset, the privacy rights of individuals are protected. We have eliminated each user's identity from the comments and tweets. We confirm that no personal information was collected during the data collection process. We deleted identities from the dataset if they appeared in it. This task aims to build a dataset that can be used to perform sentiment analysis tasks in the Kurdish language, not to attack users.

Competing interests The authors declare no competing interests.

References

Abdul-Mageed, M., & Diab, M. T. (2012). AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. *International Conference on Language Resources and Evaluation*. https://api.semanticscholar.org/CorpusID:11657346.

Abdul-Mageed, M., & Diab, M. T. (2014). SANA: A Large Scale Multi-Genre, Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis. *International Conference on Language Resources and Evaluation*. https://api.semanticscholar.org/CorpusID:10467454.

Abdul-Mageed, M., Diab, M., & Kübler, S. (2014). SAMAR: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language, 28*(1), 20–37. https://doi.org/10.1016/j.csl.2013.03.001.

Abo, M. E. M., Shah, N. A. K., Balakrishnan, V., Kamal, M., Abdelaziz, A., & Haruna, K. (2019). SSA-SDA: Subjectivity and sentiment analysis of Sudanese Dialect Arabic. *2019 International Conference on Computer and Information Sciences (ICCIS)*, 1–5. https://doi.org/10.1109/ICCISci.2019.8716466.

Afzal, Z., Pons, E., Kang, N., Sturkenboom, M. C., Schuemie, M. J., & Kors, J. A. (2014). ContextD: An algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. *Bmc Bioinformatics*, *15*(1), 373. https://doi.org/10.1186/s12859-014-0373-3.

Ahmad, M., Aftab, S., & Ali, I. (2017). Sentiment analysis of tweets using SVM. *International Journal of Computer Applications*, *177*(5), 25–29. https://doi.org/10.5120/ijca2017915758.

Almars, M., A (2022). Attention-based Bi-LSTM model for Arabic Depression classification. *Computers Materials & Continua*, *71*(2), 3091–3106. https://doi.org/10.32604/cmc.2022.022609.

Aly, M. A., & Atiya, A. F. (2013). LABR: A Large Scale Arabic Book Reviews Dataset. *ArXiv*, *abs/1411.6718*. https://api.semanticscholar.org/CorpusID:15980568.

Awlla, K., & Veisi, H. (2022). Central kurdish sentiment analysis using deep learning. *Journal of University of Anbar for Pure Science*, *16*(2), 119–130. https://doi.org/10.37652/juaps.2022.176501.

Badawi, S. (2023a). Data Augmentation for Sorani kurdish News Headline classification using back-translation and deep learning model. *Kurdistan Journal of Applied Research*, *8*(1), 27–34.

Badawi, S. (2023b). Transformer-based neural network machine translation model for the kurdish Sorani Dialect. *UHD Journal of Science and Technology*, *7*(1), 15–21. https://doi.org/10.21928/uhdjst.v7n1y2023.pp15-21.

Badawi, S. S. (2023c). Using Multilingual Bidirectional Encoder representations from transformers on Medical Corpus for kurdish text classification. *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, *11*(1), 10–15. https://doi.org/10.14500/aro.11088.

Badawi, S., Saeed, A. M., Ahmed, S. A., Abdalla, P. A., & Hassan, D. A. (2023). Kurdish news dataset headlines (KNDH) through multiclass classification. *Data in Brief*, *48*, 109120. https://doi.org/10.1016/j.dib.2023.109120.

Baly, R., Khaddaj, A., Hajj, H. M., El-Hajj, W., & Shaban, K. B. (2019). ArSentD-LEV: A Multi-Topic Corpus for Target-based Sentiment Analysis in Arabic Levantine Tweets. *ArXiv*, *abs/1906.01830*. https://api.semanticscholar.org/CorpusID:96438072.

Banjade, R., & Rus, V. (2016). DT-Neg: Tutorial Dialogues Annotated for Negation Scope and Focus in Context. *International Conference on Language Resources and Evaluation*. https://api.semanticscholar.org/CorpusID:37135454.

Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2021). ABCDM: An attention-based bidirectional CNN-RNN Deep Model for sentiment analysis. *Future Generation Computer Systems*, *115*, 279–294. https://doi.org/10.1016/j.future.2020.08.005.

Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (2018). Sentiment analysis about E-Commerce from Tweets using decision Tree, K-Nearest Neighbor, and Naïve Bayes. *2018 International Conference on Orange Technologies (ICOT)*, 1–6. https://doi.org/10.1109/ICOT.2018.8705796.

Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (2016). An Overview of Sentiment Analysis in Social Media and Its Applications in Disaster Relief. *Sentiment Analysis and Ontology Engineering*. https://api.semanticscholar.org/CorpusID:14326757.

Bhatia, P., Ji, Y., & Eisenstein, J. (2015). Better Document-level Sentiment Analysis from RST Discourse Parsing. *ArXiv*, *abs/1509.01599*. https://api.semanticscholar.org/CorpusID:12252194.

Birjali, M., Kasri, M., & Hssane, A. B. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134. https://api.semanticscholar.org/CorpusID:235690410.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems, 28*(2), 15–21. https://doi.org/10.1109/MIS.2013.30.

Cao, H., Sen, P. K., Peery, A. F., & Dellon, E. S. (2016). Assessing agreement with multiple raters on correlated kappa statistics. *Biometrical Journal*, *58*(4), 935–943. https://doi.org/10.1002/bimj.201500029.

Chakraborty, S., Kumar, S., Paul, S., & Kairi, A. (2017). A Study of Product Trend Analysis of Review Datasets using Naive Bayes', K-NN and SVM classifiers. *INTERNATIONAL JOURNAL OF ADVANCED ENGINEERING AND MANAGEMENT*, *2*(9), 204. https://doi.org/10.24999/IJOAEM/02090047.

Dalianis, H., Hassel, M., & Velupillai, S. (2009). The Stockholm EPR Corpus-characteristics and some initial findings. In *Proceedings of ISHIMR* (pp. 243–249).

ElSahar, H., & El-Beltagy, S. R. (2015). *Building Large Arabic Multi-domain Resources for Sentiment Analysis* (pp. 23–34). https://doi.org/10.1007/978-3-319-18117-2_2.

Fauzi, M. A. (2018). Random Forest Approach for Sentiment Analysis in Indonesian Language. *Indonesian Journal of Electrical Engineering and Computer Science*, *12*(1), 46. https://doi.org/10.11591/ijeecs.v12.i1.pp46-50.

Huq, M. R., Ali, A., & Rahman, A. (2017). Sentiment Analysis on Twitter Data using KNN and SVM. *International Journal of Advanced Computer Science and Applications*, *8*. https://api.semanticscholar.org/CorpusID:28413185.

Ibrahim, H.S., Abdou, S.M., & Gheith, M. (2015). Sentiment analysis for modern standard arabic and colloquial. Preprint retrieved from https://doi.org/10.48550/arXiv.1505.03105.

Jiménez-Zafra, S. M., Morante, R., Teresa Martín-Valdivia, M., & Ureña-López, L. A. (2020). Corpora Annotated with negation: An overview. *Computational Linguistics*, *46*(1), 1–52. https://doi.org/10.1162/coli_a_00371.

Kessler, W., & Kuhn, J. (2014). A Corpus of Comparisons in Product Reviews. *International Conference on Language Resources and Evaluation*. https://api.semanticscholar.org/CorpusID:17061218.

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., & Taboada, M. (2020a). The SFU opinion and comments Corpus: A Corpus for the analysis of Online News comments. *Corpus Pragmatics*, *4*(2), 155–190. https://doi.org/10.1007/s41701-019-00065-w.

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., & Taboada, M. (2020b). The SFU opinion and comments Corpus: A Corpus for the analysis of Online News comments. *Corpus Pragmatics*, *4*(2), 155–190. https://doi.org/10.1007/s41701-019-00065-w.

Kumar, A., & Albuquerque, V. H. C. (2021). Sentiment analysis using XLM-R transformer and zero-shot transfer learning on resource-poor Indian Language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *20*(5), 1–13. https://doi.org/10.1145/3461764.

Lindén, K., Jauhiainen, T., & Hardwick, S. (2023). FinnSentiment: A Finnish social media corpus for sentiment polarity annotation. *Language Resources and Evaluation*, *57*(2), 581–609. https://doi.org/10.1007/s10579-023-09644-5.

Nabil, M., Aly, M., & Atiya, A. (2015, September). Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2515–2519).

Oronoz, M., Gojenola, K., Pérez, A., de Ilarraza, A. D., & Casillas, A. (2015). On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics*, *56*, 318–332. https://doi.org/10.1016/j.jbi.2015.06.016.

Qian, Z., Li, P., Zhu, Q., Zhou, G., Luo, Z., & Luo, W. (2016). Speculation and Negation Scope Detection via Convolutional Neural Networks. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 815–825. https://doi.org/10.18653/v1/D16-1078.

Ramadhan, W., Novianty, S., & Setianingsih, S. (2017). Sentiment analysis using multinomial logistic regression. *2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC)*, 46–49. https://api.semanticscholar.org/CorpusID:20471455.

Rao, G., Huang, W., Feng, Z., & Cong, Q. (2018). LSTM with sentence representations for document-level sentiment classification. *Neurocomputing*, *308*, 49–57. https://doi.org/10.1016/j.neucom.2018.04.045.

Rashid, T. A., Mustafa, A. M., & Saeed, A. M. (2018). *Automatic Kurdish Text Classification Using KDC 4007 Dataset* (pp. 187–198). https://doi.org/10.1007/978-3-319-59463-7_19.

Refaee, E. A., & Rieser, V. (2014). An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. *International Conference on Language Resources and Evaluation*. https://api.semanticscholar.org/CorpusID:6241685.

Rogers, A., Romanov, A., Rumshisky, A., Volkova, S., Gronas, M., & Gribov, A. (2018). RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian. *International Conference on Computational Linguistics*. https://api.semanticscholar.org/CorpusID:49221615.

Rushdi-Saleh, M., Martín-Valdivia, M. T., López, L. A. U., & Ortega, J. M. P. (2011). OCA: Opinion corpus for Arabic. *Journal of the American Society for Information Science and Technology*, *62*. https://api.semanticscholar.org/CorpusID:16310031.

Saeed, A. M., Hussein, S. R., Ali, C. M., & Rashid, T. A. (2022). Medical dataset classification for kurdish short text over social media. *Data in Brief*, *42*, 108089. https://doi.org/10.1016/j.dib.2022.108089.

Schouten, K., & Frasincar, F. (2016). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, *28*(3), 813–830. https://doi.org/10.1109/TKDE.2015.2485209.

She, X., Zhang, D., & on Hybrid CNN-LSTM Hybrid Model. (2018). Text Classification Based. *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, 185–189. https://doi.org/10.1109/ISCID.2018.10144.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.