QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (4), Autumn 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

Twitter Sentiment Analysis for Kurdish Language

Didam Mahmud
Information Technology, College of Commerce, Sulaimani University, Sulaymaniyah, Iraq
Didam.mahmud@univsul.edu.iq

Bawar Abid Abdalla
Software Engineering, Faculty of Engineering, Koya University, Koya, Iraq
bawar.abid@koyauniversity.org

Azhi Faraj
Information Technology, College of Commerce, Sulaimani University, Sulaymaniyah, Iraq
Azhi.faraj@univsul.edu.iq

ARTICLE INFO

Article History:
Received:18/9/2022
Accepted: 7/12/2022
Published:Autumn 2023

Keywords:
Kurdish text, Machine learning, Sentiment analysis, Stemming.

Doi:
10.25212/lfu.qzj.8.4.42

ABSTRACT

Sentiment analysis of text data has received a significant attention throughout Natural Language Processing stages. However, most of the focus has been on English language depriving many other languages from taking advantage of the state-of-the-art techniques most suitable to a particular language especially the Kurdish Sorani language. This paper is an attempt to bridge the gap between English and Kurdish language in sentiment analysis for social media text. For this purpose, firstly a new Kurdish sentiment analysis dataset was curated and annotated then we tried different combinations of machine learning algorithms including classical machine learning algorithms such as Random Forrest, KNN, SVM, Naive Bayes bias and Decision trees and compared the results to Deep Learning techniques namely ANN, LSTM and CNN. In our experiments Naïve Bayes achieved the best results achieving an 78% accuracy.

1.Introduction

The use of microblogging networks like Twitter has increased dramatically in recent years (Bayari, 2021). Sentiment analysis sometimes alternatively referred to as opinion mining is the computational study of people's opinions, sentiments, emotions, appraisals, and attitudes towards entities such as products, services,

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (4), Autumn 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

organizations, individuals, topics, and their attributes (Edmonds, 2013). A polarity classification task for distinguishing positive, negative, or neutral messages is used in sentiment analysis to quantify what people believe using textual qualitative data. (Nassr, Sael, & Benabbou, 2020)

Since, for the first time in human history, we have a massive volume of opinionated data recorded in digital forms, the inception and rapid growth of Sentiment Analysis coincide with those of social media on the Web, such as reviews, forum discussions, blogs, microblogs, Twitter, and social networks. Sentiment analysis has been one of the most active study fields in natural language processing since its inception in the early 2000s (Zhang , Wang, & Liu, 2018).

In the recent decade, sentiment analysis has sparked a lot of academic interest making it one of the most researched and experimented with topics of Natural Language Processing due to the importance of sentiment analysis in business and research. (Bahja, 2020) Yet Kurdish language to the best of our knowledge has no publicly available annotated dataset despite vast amounts of Kurdish “tweets” on twitter, giving a rich supply of viewpoints on a wide range of issues and topics. This has led to shallow research as stated in the literature review section of this paper. The reason for this might owe to difficulties processing the morphologically complex as well as a lack of tools and resources for extracting Kurdish feelings from text.

In this research we have built an annotated dataset for Kurdish Sorani language that are categorized into positive and negative polarities. Then the dataset has been annotated and a variety of machine learning algorithms and deep learning techniques were applied to determine the best achieving algorithm.

1.1 KURDISH SORANI LANGUAGE

Kurdish refers to Indo-European language family, it is one of the languages of the Iranian language group. Consequently, Kurdish has a lot of similarities with Persian, Pashto, Balochi and other Iranian languages. Similarly, there are a lot of parallels between Kurdish and Hindi, but it is less than the parallels among the Kurdish with

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (4), Autumn 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

the Iranian language. Furthermore, the similarities decline once we compare it with other languages from other groups of Indo-European languages (Ismail, 1977).

Although, there are a lot of similitudes between Kurdish and other languages such as Arabic and Turkish, the resemblances are not original. This is because their origins are not same (Ismail, 1977) (Anon, 2021). Sorani Kurdish is one of the main dialects of Kurdish along with Kurmanji Kurdish and Southern Kurdish (Edmonds, 2013). This dialect is mainly spoken by the Kurdish populations in the Kurdistan regional of Iran and Iraq. Unlike Kurmanji dialect for which a Latin-based script is used, Sorani Kurdish is written in the Arabic-based script of Kurdish.

Kurdish Sorani utilizes Unicode alphabet which is represented by 34 letters while وو is included as an individual letter. The letters are (ئ،ا،ب،پ،ت،ج،چ،ح،خ،د،ر،ڕ،ز،ژ،س،ش،ع،غ،ف،ڤ،ق،ک،گ،ل،ڵ،م،ن،ه،ە،و،ۆ،وو،ی،ێ), of which 26 are consonants and 8 are vowels (bolded) It is almost similar to Arabic and Persian alphabet letters (Edmonds, 2013) (Anon, 2021).

Kurdish Sorani utilizes subject-object-verb word order where the system of tense-aspect-modality and person marking is present (Haig & Matras, 2002). Moreover, Kurdish Sorani uses different pronominal enclitics for ergative-absolutive alignment (Esmaili & Salavati, 2013). For Example, (khwardnakanman hênanهێنان خواردنەکانمان) This sentence is in past, the man in italic is used to refer the subject while the underlined n is used to recognize the object. However, (deçine malman/ دەچنە ماڵمان) The man in italic is referred to ours, as well as the underlined n is indicating the present tense (Ahmadi & Masoud, 2020).

1.2 LITERATURE REVIEW AND RELATED WORK

With fast growth of social media and web applications, users began to provide comments, review feedback, and rating. These viewpoints can cover a wide range of topics, including, politics products, events, services, people, and news. To get a good estimate of what the user thinks and feels, all of this must be processed and evaluated. Prior to the availability of automated sentiment analysis tools, getting customer evaluations was a time-consuming and inefficient procedure. This is most

likely the reason for the high level of interest in this subject of study (Waters & Lester, 2010).

Kurdish language suffers from lack of standards, rules, alphabets, and syntax of grammar. As a result, Kurdish people express themselves in many ways through social media. Some users like to write their emotions and sentiments in Arabic script, while others prefer to write them in Latin letters. different Kurdish accents and grammar might also have been written, and some even use English letters. A specialized algorithm for Kurdish sentiment analysis is needed that only considers the presence of each word in negative or positive phrases (Abdulla & Hama, 2015)

Many studies have recently focused on analyzing social media emotions to obtain a sense of how people feel about current subjects of interest or specific products or services, However There is limited research regarding the sentiment analysis of Kurdish language feeds. The authors of (Abdulla & Hama, 2015)proposed sentiment analysis for Kurdish Language, The Social Network comments are categorized into positive and negative polarity, Kurdish text have been gathered from various social media platforms such as Twitter, Facebook, and Google+. This collected information is saved in a special database, then Naive Bayes classifier is used for the unigram feature on a Kurdish text dataset, the accuracy of sentiment analysis was 66%.

Various source of text data has been used in previous research for instance, (Gamallo & Garcia, 2014) They authors present a technique for identifying the polarity of English tweets, the training corpus contained 6,408 tweets with the polarity values Positive, Negative, Neutral, Objective, and Neutral-or-Objective. Naive-Bayes classifier method was applied. The experimental results showed that utilizing a binary classifier with only two polarity categories: positive and negative. The F-score obtained as a result of the experiment was 63%.

The authors of (Abbasi, Chen, & Salem, 2008) utilized local grammar to assess sentiment in material published in Arabic, English, or Urdu. Financial news was the focus of the target reviews. Their systems accuracy was 91%.

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (4), Autumn 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

Rushdi-Saleh et al. The authors in (Rushdi-Saleh, 2011)constructed a small Arabic opinion corpus with 500 negative and 500 positive reviews. To predict sentiment from the corpus, a handful of classifiers were trained. The classifiers and preprocessing approaches they used provided 90% accuracy using SVM.

Recently several researches been contacted regarding to sentiment analysis on twitter (Rakibul , Maisha , & Arifuzzaman, 2019), in (Rosenthal, Farra, & Nakov, 2017)used supervised learning approach, according to the authors, is based on label datasets that are trained to provide useful outputs. Applying the Naive Bayes method, maximum entropy, and support vector machines to monitor the learning process that provides effective sentiment analysis. In (Vadivukarassi, Puviarasan, & Aruna, 2017), the authors demonstrated that they could achieve a maximum accuracy of 82.1% using the Naive Bayes algorithm. In (Pal & Ghosh, 2017), the authors performed sentiment analysis using the K-Nearest Neighbor classifier and achieved an accuracy of 74.74%

2.THE DATASET

The dataset is also a contribution of this paper which includes 2000 rows, each row consists of text column (tweet text) and an emotion column which is the dependent variable. The emotion column contains one of two possible numbers 1 and 2 for positive and negative sentiments, respectively. There are two thousand rows in the dataset 1000 positive and 1000 negative, the tweets have been assigned to positive and negative classes manually by three experts. The dataset Text is in Kurdish Sorani. The dataset has been created using twitter which is a social networking website. In twitter, the users can register accounts and they will be able to express their emotions as well as they have abilities to read others posts and retweet them. We used vicinitas website to extract tweets from tweeter from Jan 4th 2021 to 8th Jan 2021 (Vicinitas, 2021), which the Covid-19 spread over the Kurdistan region of Iraq. In addition, we set a filter to select only the Kurdish Sorani tweets as well as remove all the retweets. furthermore, the dataset was modified manually to remove any nonrelevant tweets. Finally, the rows were assigned to positive or negative values by experts.

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (4), Autumn 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

3.METHODOLOGY

The metrics used in text classification include Accuracy, Precision, Recall and F1 score each applicable in different scenarios (Forman, 2003). For a balanced dataset accuracy is usually reported (Forman, 2003) which is the ratio of correctly predicted samples over all predictions. Because there are only two possible sentiments to be determined by the model an accuracy of 50% is considered random guess therefore our subsequent attempts were to increase the accuracy as much as possible.

In the data preprocessing step, we removed links, punctuations, and numbers, it is important to set the flag to regex Unicode for removing non-ascii punctuations while dealing with Kurdish. since in Kurdish language there are no capital and small letters, there was no need to apply lowercasing techniques. In regard to stemming there are no available stemmers for Kurdish language therefore it was only possible to apply F3 and F5 stemming which refer to taking the first three letters and first five letters respectively. The F5 stemming made a considerable impact on all our scenarios.

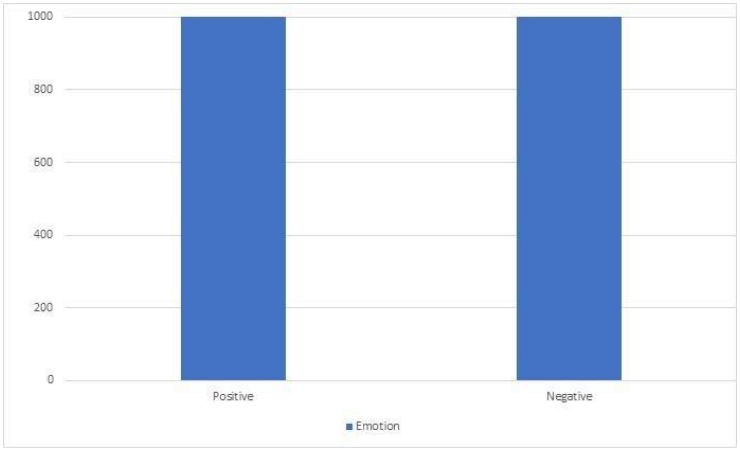

Figure 1: Distribution of positive and negative samples

To get more insight about the data, a frequency distribution of words was plotted for each emotion category, however without removing the stop words there will not be

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (4), Autumn 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

much information to obtain from such graphs because a list of generic words will be at the top of every category. For the stop words we used the following list [' ئێمە', 'ئێوە', 'ئەم', 'ئەو', 'ئەوان', 'ئەوەی', 'بۆ', 'بێ', 'بێجگە', 'بە', 'بەبێ', 'بەدەم', 'بەردەم', 'بەرلە', 'بەرەوی', 'بەرەوە', 'بەلای', 'بەپێی', 'تۆ', 'تێ', 'جگە','دوای', 'دوو', 'دە', 'دەکات', 'لەگەڵ', 'سەر', 'لێ', 'لە', 'لەبابەت', 'لەباتی', 'لەبارەی', 'لەبرێتی', 'لەبن', 'لەبەر', 'لەبەینی', 'لەدەم','لەری', 'لەرێگا', 'لەرەوەی', 'لەسەر', 'لەلایەن', 'لەناو', 'لەنێو', 'لەو', 'لەپێناوی', 'لەژێر', 'لەگەڵ', 'من', 'ناو', 'نێوان', 'هەر','هەروەها', 'و', 'وەک', 'پاش', 'پێ', 'پێش', 'چەند', 'کرد', 'کە'] (StopWords, 2021), However after removing the stop words a more meaningful representation was achieved as shown in (Figure 2). A sample of converted text after the preprocessing is show in table 1.

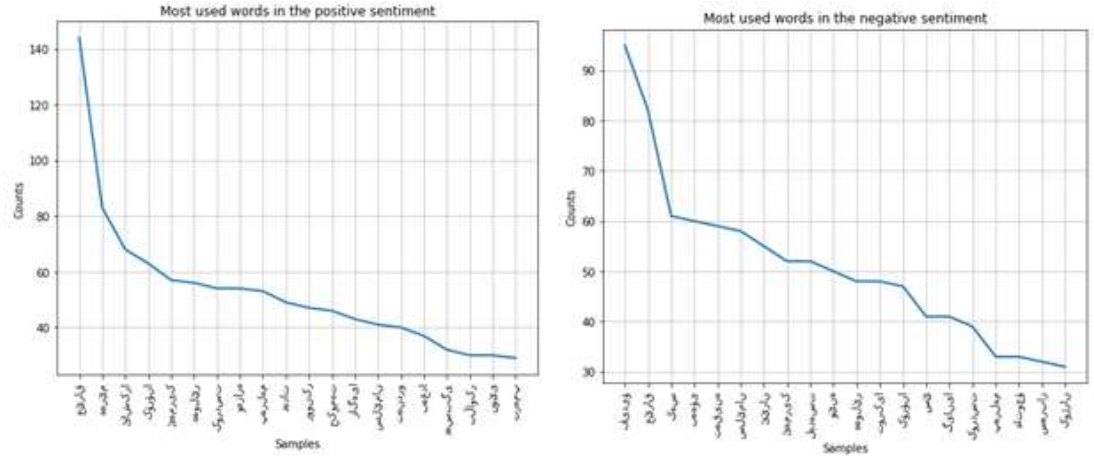

Figure 2: Top 20 used words after removing stop words in negative and positive sentiment

Original Tweet	Cleaned Tweet
تەندروس تەندروستی هەرێم لەبارەی کەمبوونەوەی ژمارەی دەدات تۆشبووانی کۆرۆناوە رووبەرووبونەوە https://t.co/07RvG3IweA	تەندروستی هەرێم کەمبوو ژمارەی تۆشبوو کۆرۆنا رووبەرووبونەوە دەدات

Table 1: removing noise from a tweet

For the models that does not have an embedding layer (all except CNN and LSTM) a bag of words representation of the tweets was created before feeding the data to

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (4), Autumn 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

the models, in this paper several machine learning algorithms were tried and for each algorithm several hyperparameter were tested out, we used Grid search of parameters to find the best performing hyperparameter. The best performing hyperparameters for logistic Regression we specified the max iteration on 200, L2 as the loss function, used Saga as the solver, as for KNN the number of neighbors were set to 5, distance as the weight and Makowski as the metric, For Naïve Bayes we opted for the Multinomial version, regarding Random Forrest classifier we set the max depth to 90, number of estimators to 500.

Convolutional Neural Networks require an embedding layer instead of a bag of words representation hence the embedding layer was developed with a 50k vocabulary size and 8 vector space size additionally the input length was set to 50. The embedding are fed into a 1D convolution layer with ReLU activation function then a max pooling layer. In order to feed the data to a fully connected network a flattening layer is added which is connected to a dense layer with ReLU activation function, a dropout layer of 20% and the output layer with two output nodes. We opted for adamax as the optimizer and accuracy as the metric.

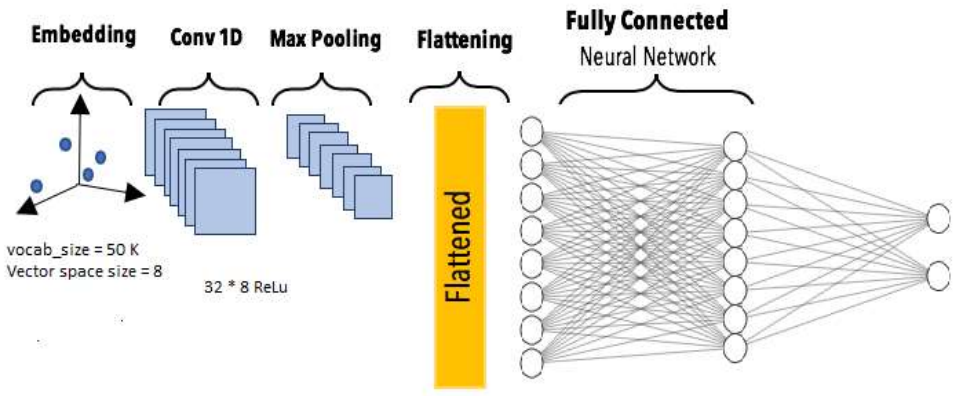

Figure 3: Proposed CNN Architecture

In addition to CNN, we also added a bidirectional LSTM to the architecture described in Figure 3.

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (4), Autumn 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

4.RESULTS AND DISCUSSION

In order to evaluate the proposed algorithms and to measure the effect of the aforementioned techniques such as the removal of stop words and stemming accuracy was used as the evaluation metric.

Logistic Regression Produced the best results with an accuracy of 78%, Naïve Bayes and Random Forrest were a close second with 77%, The deep learning techniques although are proven to work well with text data achieved worst results however this was expected due to the limited amount of text data available, we believe that with a bigger dataset our deep learning approaches will surpass the other algorithms. Figure 4 shows the results of the classification accuracy of each model.

The usage of stop words despite their usefulness for analysis in our research failed to improve the results, although the stop words in Kurdish language has not been unified and this might play a role in their ineffectiveness, the removal of stop words in other languages has not guaranteed an improvement in results as well. (Aro, 2019)used three different datasets and concluded that the removal of stop words has had a negative impact on predictions.

Another important step in sentiment analysis is stemming which refers to reducing the word to its root form. For stemming we used fixed length stemming which refers to taking a static number of characters from each word token, different lengths were investigated in this research and the best performing technique was F5 stemming which improved the accuracy of every single model. Figure 5 shows the results of the same models with F5 stemming. The biggest gain was noticed in CNN model with a 10% enhancement in the accuracy.

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (4), Autumn 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

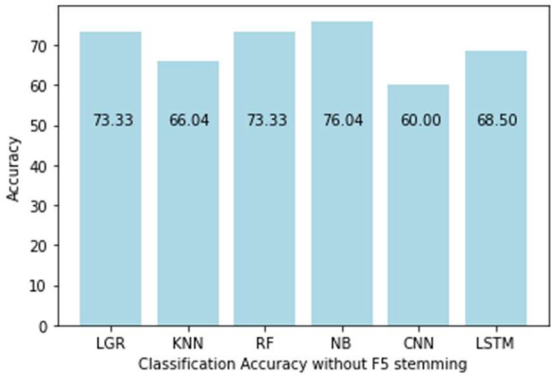

Figure 4: accuracy without stemming

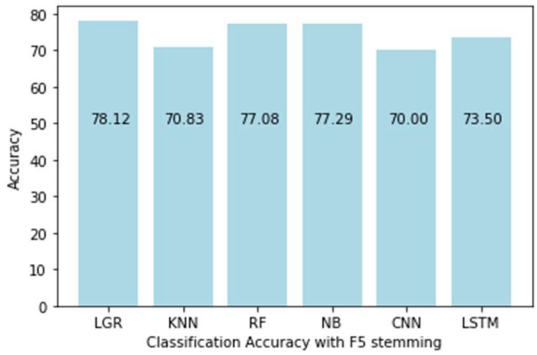

Figure 5: accuracy with F5 stemming

5.CONCLUSION

In this paper we recurred a Kurdish Sorani dataset for the purpose of binary sentiment analysis, we analyzed the lack of researches utilizing natural language processing technique for Kurdish language. A variety of machine learning algorithms were tested with varying hyperparameters. Additionally, Deep learning techniques were applied using CNN and LSTM. The techniques achieved state of the art results with 78% accuracy. The best performing model was based on Naïve bias. Our research highlighted the significance of F-5 stemming which improved the results of all the used models. The deep learning results were inconsistent due to the limited size of the dataset, as a future work we propose a bigger dataset and a different dataset for emotion analysis .

References:

Edmonds, A. J. (2013). *The Dialects of.* Ruprecht-Karls-Universität Heidelberg.

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM transactions on information systems (TOIS)*, 1-34.

Abdulla, S., & Hama, M. (2015). Sentiment analyses for Kurdish social network texts using Naive Bayes classifier. *Journal of University of Human Development*, 393-397.

Ahmadi, S., & Masoud, M. (2020). Towards machine translation for the Kurdish language. *arXiv preprint arXiv:2010.06041*.

Anon. (2021, 7 1). *Ethnologue Languages of the World*. Retrieved from https://www.ethnologue.com

Aro, T. O. (2019). Homogenous ensembles on data mining techniques for breast cancer diagnosis. *Daffodil International University*.

Bayari, R. a. (2021). Text mining techniques for cyberbullying detection: state of the art. *Adv. Sci. Technol. Eng. Syst. J*, 783_790.

Esmaili, K., & Salavati, S. (2013). Sorani Kurdish versus Kurmanji Kurdish: an empirical comparison. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 300-305).

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. . *J. Mach. Learn. Res.*, 1289-1305.

Gamallo, P., & Garcia, M. (2014). Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets. In *Semeval@ coling}* (pp. 171-175).

Haig, G., & Matras, Y. (2002). Kurdish linguistics: a brief overview. *STUF-Language Typology and Universals*, 3_14.

Ismail, Z. B. (1977). *Kurdish language history [tarikh al-alughat al-kurdiya].* Baghdad: al-Hawadth.

Nassr, Z., Sael, N., & Benabbou, F. (2020). Preprocessing arabic dialect for sentiment mining: State of art. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 323_330.

Rushdi-Saleh, M. a. (2011). OCA: Opinion corpus for Arabic. *Journal of the American Society for Information Science and Technology*, 2045-2054.

StopWords. (2021, 6 1). *StopWords*. Retrieved from https://stopwords.net/kurdish-ku

Vicinitas. (2021, 8 8). *Vicinitas*. Retrieved from https://www.vicinitas.io/free-tools/download-user-tweets

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (4), Autumn 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

Waters, J., & Lester, J. (2010). *The Everything Guide to Social Media: All you need to know about participating in today's most popular online communities.* Simon and Schuster.

Zhang , L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e12345.

الملخص:
حظي تحليل المشاعر للبيانات النصية باهتمام كبير خلال مراحل معالجة اللغة الطبيعية. ومع ذلك ، فإن معظم التركيز كان على اللغة الإنجليزية بالتالي احرمت العديد من اللغات الأخرى من الاستفادة من هذه التقنيات وخاصة اللغة السورانية الكردية. هذا البحث هي محاولة لسد الفجوة بين اللغة الإنجليزية واللغة الكردية في تحليل المشاعر لنصوص في وسائل التواصل الاجتماعي. لهذا الغرض ، أولاً ، تم تنسيق مجموعة بيانات جديدة لتحليل المشاعر الكردية ووضع تعليقات توضيحية عليها ، ثم جربنا مجموعات مختلفة من خوارزميات التعلم الآلي بما في ذلك خوارزميات التعلم الآلي الكلاسيكية مثل Random Forrest و KNN و SVM و Naive Bayes bias و أشجار القرار وقارننا النتائج مع Deep تقنيات التعلم وهي ANN و LSTM و CNN. في تجاربنا ، حقق Naïve Bayes أفضل النتائج محققًا دقة تبلغ 78٪.

پـوخـتـه:

لێکدانەوەی بیروبۆچونەکانی زانیاری نوسراو گرنگی و سەرنجی زۆری پێدراوە لە قوناغەکانی پرۆسیێس کردنی زمانە سروشتیەکان٠ زۆربەی سەرنج و گرنگی دراوە بە زمانی ئینگلیزی ، زمانەکانی تر بێبەش کراون لە سود وەرگرتن لەم تەکنیکە هونەریە نوێیانە بەتایبەتیش زمانی کوردی سۆرانی٠ئەم توێژینەوەیە هەوڵێکە دروستکردنی پردێک لە نێوان زمانی کوردی و زمانی ئینگلیزی لە لێکدانەوەی بیرو بۆچونەکانی تێکستی سۆشیاڵ میدیا٠

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (4), Autumn 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

بۆ ئەم مەبەستە پێش هەموو شتێک لێکدانەوەی تازەی زانیاری بە کوردی ڕێکخراوە و دروست کراوە پاشان هەوڵمان داوە چەند ئەلگۆریزمێکی جیاواز تێکەڵ بکەین وەکو Random Forrest KNN, SVM, NAÏVE Bayes, Decision Tree وە بەراوردکردنی ئەنجامەکان بە تەکنیکەکانی Deep Learning وەک ANN, LSTM, CNN لە تاقیکردنەوەکانماندا Naïve Bayes باشترین ئەنجامی هەبووە کە ڕێژە ی دروستیەکەی لە سەدا ٧٨ بووە.