Research Article

Sentiment Analysis of Opinions about Online Education in the Kurdistan Region of Iraq during COVID-19

Maryam Sami[1], Hossein Hassani[1]

1. University of Kurdistan Hewlêr (UKH), Erbil, Iraq

Sentiment analysis is widely used in various areas and has versatile applications. For example, it is used in market research, customer retention strategies, and product analysis, to name a few. Although some previous work on the topic exists for the Kurdish language, similar to other fields in Kurdish processing, it is not well-studied, and particularly it suffers from data inadequacy. In this paper, we present an overview of the research we conducted to analyze the sentiments of learners/educators toward online education during COVID-19 in the Kurdistan Region of Iraq. We collected the data from Tweets tweeted up to March 2022. After preprocessing, 511 items remained. The dataset is publicly available for non-commercial use under CC0 1.0 Universal license at https://github.com/KurdishBLARK/SentimentAnalysis/tree/main/OnlineEducation-during-COVID-19

Corresponding authors: Maryam Sami, maryam.sami@ukh.edu.krd; Hossein Hassani, hosseinh@ukh.edu.krd

1. Introduction

The spread of Coronavirus (COVID-19) in 2020 dramatically affected every aspect of life. As a precaution, governments had to set laws for people's safety. In almost all countries, the authorities enforced quarantine at different levels. That caused the world to descend into chaos, and like the rest of the world, people in the Kurdistan Region of Iraq (KRI) had to adapt to managing their lives in a new way. One of the concerns was education, and COVID-19 forced educational institutions to switch from face-to-face to online education. However, regardless of the readiness level of different regions and countries for such a dramatic shift, many wondered about its efficiency.

Sentiment analysis has been an instrument to study the efficiency of various educational methods for a while (Kechaou et al., 2011), but its usage in the analysis of the ``enforced'' online education to understand the opinion of students emerged rapidly during the COVID-19 pandemic (see (Mujahid et al., 2021; Waheeb et al., 2022)). We also attempted to use the technique to analyze the situation in the KRI and developed a dataset from tweets about online education in the region. We focused on the tweets written in Sorani Kurdish and tested different machine-learning algorithms to assess their accuracy in analyzing the sentiments.

In this paper, we present an overview of our research and the developed dataset developed regarding sentiment analysis of opinions about online education during the COVID-19 pandemic in the KRI. The dataset is available under CC0 1.0 Universal license at https://github.com/KurdishBLARK/SentimentAnalysis/tree/main/OnlineEducation-during-COVID-19. The rest of this paper is organized as follows. In Section 2 we present a summary of related work. Section 3 briefly presents the method we followed, Section 4 demonstrates the results of data collection and preparation, Section 5 illustrates the outcomes of the experiments of applying various algorithms on the dataset, and finally, Section 6 concludes the paper.

2. Related Work

The research about sentiment analysis has attracted the research community in the past decade and it seems to keep its status at least for the near future. Table 1 summarizes the related work and presents the topics, number of entries in the studied datasets, methods with the highest accuracy and the languages of the contents.

 doi.org/10.32388/SU5BR9

Reference	Subject of dataset	Entries	Best method(s)	Accuracy	Language
Neethu and Rajasree (2013)	Review tweets	1200	SVM, Ensemble & Maximum Entropy	90%	English
Barnaghi et al. (2016)	Tweets about about World Cup 2014	4162	Bayesian Logistic Regression	74.84%	English
Ramadhan et al. (2017)	Tweets about Jakarta Governor election	1356	Multinomial Logistic Regression	74%	English
Baid et al. (2017)	Book reviews	2000	Naïve Bayes	81.45%	English
Jagdale et al. (2019)	Camera reviews	3106	Naïve Bayes	98.17%	English
	Laptop reviews	1946	Naïve Bayes	90.22%	English
Jagdale et al. (2019)	Mobile reviews	1918	Naïve Bayes	92.85%	English
Jagdale et al. (2019)	Tablet reviews	1894	Naïve Bayes	97.17%	English
Jagdale et al. (2019)	TV reviews	1596	Naïve Bayes	90.16%	English
Jagdale et al. (2019)	Video surveillance	2597	Naïve Bayes	91.13%	English
Alomari et al. (2017)	Tweets in Jordanian	1800	SVM	88.72%	Arabic
Ahmad and Aftab (2017)	Self-deriving cars	7156	SVM	59.9%	Arabic
Ahmad et al. (2017)	Apple products	3884	SVM	71.2%	Arabic
Ahuja et al. (2019)	Tweets of various opinions	4242	Logistic Regression	57%	Arabic
Almouzini et al. (2019)	Tweets about depression	4542	Random Forest	82.39%	Arabic
Alhajji et al. (2020)	Tweets about governmental preventive measures to contain COVID-19	58000	1-gram Naïve Bayes	89%	Arabic
Al-Bayati et al. (2020)	Arabic book reviews	3315	Deep Learning	82%	Arabic
Vaziripour et al. (2016)	Tweets about online education	3480	Logistic Regression	89.9%	Arabic
	Tweets about politics	3000	Naïve Bayes	95%	Persian
Dashtipour et al. (2021)	Movie reviews	2010	Long Short-Term Neural Network	95.61%	Persian
Abdulla and Hama (2015)	Social media comments	15000	Naïve Bayes	66%	Kurdish
Saeed et al. (2022)	Medical sentiments	6756	N/A	N/A	Kurdish

doi.org/10.32388/SU5BR9

Table 1. A summary of literature review.

The review of the related work suggests the following points:

- Naïve Bayes, Support Vector Machine (SMV), Logistic Regression, and Random Forest have better performance and score higher accuracy than other algorithms. Therefore, we use those four algorithms in this research.
- The increase of the training set enhances the accuracy and performance of the sentiment analysis model, particularly, when the datasets are small.
- Term Frequency-Inverse Document Frequency (TF-IDF) and N-grams are common approaches for the feature extraction.
- Tweepy and other applications interacting with Twitter's API work better for Tweets retrieval.
- Research on sentiment analysis in Kurdish is extremely limited. We were able to only retrieve two papers regarding sentiment analysis in Kurdish at the time of preparing this paper.

We use the points mentioned above to design the method to conduct this research and as we described in the following sections.

3. Method

We obtain a developer's account from Twitter to collect the required data by setting the proper attributes that filter the data. We preprocess the data and clean it to prepare it for labeling. As we expect not to be able to collect a large amount of data, we select the most suitable algorithms that have shown appropriate performance on a small amount of data (see Section 2). This section explains the details of the method.

Tweets usually includes informal language containing slang words, abbreviations, and grammatical mistakes. Preprocessing steps must be applied to the data to transform the text into a format that a machine learning model can analyze. In the preprocessing phase, we do the following:

- Removing all irrelevant tweets, such as those made in a language other than Kurdish, tweets about ads, news, and topics unrelated to how the online education process is received in Kurdistan.
- Removing punctuation, emojis, symbols, URLs, stop words, numbers
- Removing all Arabic diacritics ???????????

doi.org/10.32388/SU5BR9

- Removing duplicates.
- Transliterating texts in Latin into Persian/Arabic script.

3.1. Labeling

We ask three native Kurdish experts to evaluate the data and manually label them. The three experts delete spams, unrelated tweets, and transliterate Tweets written in Latin to Persian/ Arabic Script, sorting the data into Negative or Positive by majority voting, and tagging them with either of the two classes. This process dictates that this research is a supervised machine-learning model. The preprocessing phase results in two datasets due to splitting the preprocessed dataset into training and testing datasets.

4. Data Collection and Preparation

We obtained a developer's account from Twitter and prepared a program to collect the data. We set the date to March 2022, the location to the KRI, and used Persian-Arabic and Latin for Kurdish terms along with English terms for the search parameters. Table 2 shows examples of terms we used for searching.

Persian-Arabic	Latin	English
خوێندنی ئلکترۆنی	Xwendny online	online_education
خوێندنی ئۆنالین	Xwendny electrony	OnlineLearning
خوێندنی کۆرونا	Xwendny kati corona	UniversityOnlineCourse
	Xwendny zamani corona	DistanceLearning
	Xwendn ba shewazy nwey electrony	

Table 2. Examples of search terms.

Using this technique, we collected 720 tweets, including spam, unrelated news, replies, retweets, and ads. After preprocessing, the resulting dataset consisted of 511 tweets in Total, 368 (5852 words) items being negative, and 143 (3535 words) labeled Positive. Table 3 shows samples of the labeled tweets.

Table 4.3: sample from the collected labeled tweets.

Positive	Negatice
خندنی ئۆنالین ومعاشی ئۆفالین	کۆرسی نوسنی کرەنتف، هەتم بۆت!! زۆر ئێکساتدم
خوتندنی ئۆنالین حاڵم خۆشە#onlinelearning	فترخوازانی زانکۆ خۆتان بۆ خوتندنی #ئۆنڵاین ئامادەکەن سبەی بڕیار لەسەر خوتندنی ئەمساڵی زانکۆکان دەدرتت! #StayAtHome
برادەرتک هەیە ئەڵتی پشتگیری لە خوتندنی ئۆنڵاین و هەموو بڕیارتکی فاشلی زانکۆ نەکات معاشەکەی ئەبڕن مەبەستشم @shkoagha نیە!	خوتندنی #ئۆنڵاین دەستپتدەکاتەوە

Table 3. Examples of labeled tweets.

The resulting data of this stage still included noises, such as emojis, URLs, numbers, Arabic diacritics, and English words. We used the KLPT toolbox (Ahmadi, 2020) to clean the data further and to stem it. Table 4 shows examples of this activity.

Before preprocessing	After preprocessing
تووخوا تتچوو و پارەی خوتندنی دوور چە لەکاتی پتندەمک؟ گەمەی بە عەسابم دەکا.	تووخوا چ پارە خوتن دوور چە لەکات ندەم؟ گەمە عەساب دەکا.
بۆچی فتربوونی دوور هەڵناوەشتننەوە و کانسلی ناکەن؟	بۆ فتر دوور هەڵناوەشتن کانسل کەن
ترسناکە!!! #فتربوونی دوور	ترسناکە فتر دوور

Table 4. Examples of preprocessed tweets.

5. Experiments

We assessed four methods for their accuracy in the sentiments classification: Naïve Bayes, Support Vector Machine (SVM), Random Forest, and Logistic Regression. We chose those algorithms because the literature has reported their performance on small datasets is reasonable.

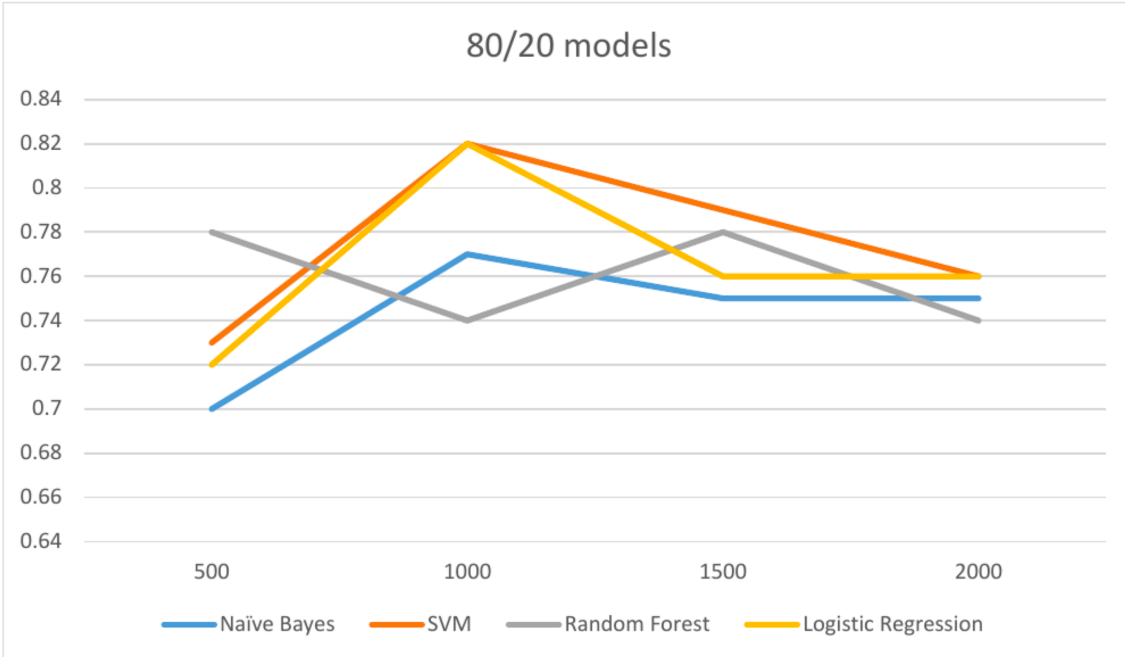

Figure 1. Classifiers evaluations for 80/20 models through features incrementation.

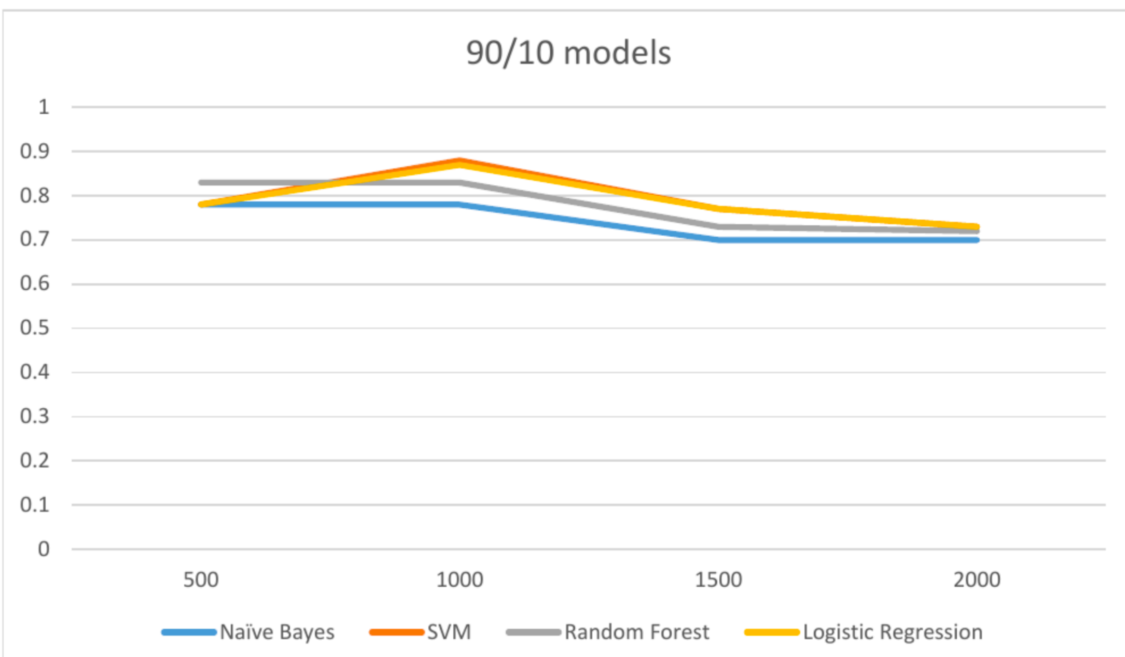

Figure 2. Classifiers evaluations for 90/10 models through features incrementation.

doi.org/10.32388/SU5BR9

6. Conclusion

The work presents sentiment analysis models for Kurdish (Sorani). We developed a dataset of Sorani tweets about online education during COVID-19, resulting in 512 tweets. We developed a language model and trained four algorithms, Naïve Bayes, SVM, Random Forest, and Logistic Regression, to classify the sentiments into positive and negative.

Acknowledgments

We would like to extend our gratitude to those who helped us in checking the relevance of the tweets' contents to the subject and assisted us in labeling the data.

References

- Abdulla, S. and Hama, M. H. (2015). Sentiment analyses for kurdish social network texts using naive bayes classifier. Journal of University of Human Development, 1(4):393–397.
- Ahmad, M. and Aftab, S. (2017). Analyzing the Performance of SVM for Polarity Detection with Different Datasets. International Journal of Modern Education and Computer Science, 9(10):29.
- Ahmad, M., Aftab, S., and Ali, I. (2017). Sentiment Analysis of Tweets using SVM. Int. J. Comput. Appl, 177(5):25–29.
- Ahmadi, S. (2020). KLPT – Kurdish language processing toolkit. In Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS), pages 72–84, Online, November. Association for Compu- tational Linguistics.
- Ahuja, R., Chug, A., Kohli, S., Gupta, S., and Ahuja, P. (2019). The Impact of Features Extraction on the Sentiment Analysis. Procedia Computer Science, 152:341–348.
- Al-Bayati, A. Q., Al-Araji, A. S., and Ameen, S. H. (2020). Arabic sentiment analysis (asa) using deep learning approach. Journal of Engineering, 26(6):85–93.
- Alhajji, M., Al Khalifah, A., Aljubran, M., and Alkhalifah, M. (2020). Sentiment Analysis of Tweets in Saudi Arabia regarding Governmental Preventive Measures to contain COVID-19.
- Almouzini, S., Alageel, A., et al. (2019). Detecting Arabic Depressed Users from Twitter Data. Procedia Computer Science, 163:257–265.
- Alomari, K. M., ElSherif, H. M., and Shaalan, K. (2017). Arabic Tweets Sentimental Analysis using Machine Learning. In Advances in Artificial Intelligence: From Theory to Practice: 30th

International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Arras, France, June 27-30, 2017, Proceedings, Part I 30, pages 602–610. Springer.

- Baid, P., Gupta, A., and Chaplot, N. (2017). Sentiment Analysis of Movie Reviews using Machine Learning Techniques. International Journal of Computer Applications, 179(7):45–49.
- Barnaghi, P., Ghaffari, P., and Breslin, J. G. (2016). Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment. In 2016 IEEE second international conference on big data computing service and applications (BigDataService), pages 52–57. IEEE.
- Dashtipour, K., Gogate, M., Adeel, A., Larijani, H., and Hussain, A. (2021). Sentiment analysis of persian movie reviews using deep learning. Entropy, 23(5):596.
- Jagdale, R. S., Shirsat, V. S., and Deshmukh, S. N. (2019). Sentiment Analysis on Product Reviews using Machine Learning Techniques. In Cognitive Informatics and Soft Computing: Proceeding of CISC 2017, pages 639–647. Springer.
- Kechaou, Z., Ammar, M. B., and Alimi, A. M. (2011). Improving e-learning with sentiment analysis of users' opinions. In 2011 IEEE global engineering education conference (EDUCON), pages 1032–1038. IEEE.
- Mujahid, M., Lee, E., Rustam, F., Washington, P. B., Ullah, S., Reshi, A. A., and Ashraf, I. (2021). Sentiment analysis and topic modeling on tweets about online education during covid-19. Applied Sciences, 11(18):8438.
- Neethu, M. and Rajasree, R. (2013). Sentiment Analysis in Twitter using Machine Learning Techniques. In 2013 fourth international conference on computing, communications and networking technologies (ICCCNT), pages 1–5. IEEE.
- Ramadhan, W., Novianty, S. A., and Setianingsih, S. C. (2017). Sentiment Analysis using Multinomial Logistic Regression. In 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC), pages 46–49. IEEE.
- Saeed, A. M., Hussein, S. R., Ali, C. M., and Rashid, T. A. (2022). Medical dataset classification for kurdish short text over social media. Data in Brief, 42:108089.
- Vaziripour, E., Giraud-Carrier, C., and Zappala, D. (2016). Analyzing the political sentiment of tweets in farsi. In Tenth International AAAI Conference on Web and Social Media.
- Waheeb, S. A., Khan, N. A., and Shang, X. (2022). Topic modeling and sentiment analysis of online education in the covid-19 era using social networks based datasets. Electronics, 11(5):715.

 doi.org/10.32388/SU5BR9

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.