Available online at https://www.asjp.cerist.dz/en/PresentationRevue/134

RIST

Information Processing at the Digital Age Journal

CERIST Natural Language Processing Challenge

March 29th, 2023

Modeling Sentiment Analysis Using Machine Learning Algorithms for Arabic covid-19 Tweets

Yousra F.G. Elhakeem[a]*, Safa Eltayeb[a], Mohammed Aldawsari[b], Omer Salih Dawood Omer[b]

[a]*Department of Software Engineering, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, P.O. Box 151, Alkharj 11942, Saudi Arabia*

[b]*Department of Computer Science, College of Arts and Science in Wadi Aldawasir, Prince Sattam bin Abdulaziz University, Wadi-Aldawasir, Saudi Arabia*

Abstract

During Covid-19 pandemic period, people worldwide turned to use social media network to express their opinions and general feelings. Social media platforms like Twitter have become widespread tools for broadcasting and distributing news and opinions. This paper presents our participation to CERIST Natural Language Processing Challenge, task1.c: Arabic sentiment analysis and fake news detection within covid-19. This complex task is further increased when dealing with dialects that do not have the structure of Modern Standard Arabic (MSA). We introduce an experiment of sentiment analysis of Arabic tweets within covid-19 using machine learning algorithms. The used Arabic dataset was provided by the challenge organizers and it contains 4,128 tweets labeled as Positive, Negative and Neutral for training and 1,034 tweets unlabeled for testing Hadj Ameur & Aliane, 2021. In this experiment the opinions are classified by various machine learning classifiers including Support Vector Machine (SVM), Logistic Regression (LR), Multinomial Naïve Bayes (NB) and K-Nearest Neighbors (KNN). The experimental results indicated that the highest accuracy (94%) was obtained using the Logistic-Regression and SVM among other with training times of 8609s.

Keywords: Sentiment analysis ; support vector machine (SVM) ; logistic regression (LR) ; KNnearest-neighbours (KNN) ; Multinomial Naive Bayes (NB).

**E-mail address: Dr.yousraelhakeem@gmail.com.*

1. Introduction

During the pandemic period, people express their feelings on social media which contain valuable information in the context of COVID-19 pandemics. Social networking media has been globally used to share news and opinions, Alamoodi et al., 2021, Lad et al., 2022, Nemes and Kiss, 2021. People can openly express their opinions on social media sites such as Facebook, Twitter, etc. During the COVID-19 lockdown, sentiment analysis technique was used widely to determining the opinions of people and thoughts related to the worldwide pandemic, Lad et al., 2022, Syarief et al., 2019, Pokharel, 2020. Sentiment analysis is an important task in natural language processing, it provides a mean for classifying opinions based on texts of a specific topic to positive, negative, and neutral. It also presents great ability to help public health organizations design an effort for accurate information in order to give people what they need to know about Covid-19, Srikanth et al., 2022, Ali, 2021. Twitter is the leading social media platform for Arab gulf countries to share their opinions, Manguri et al., 2020, Chakraborty et al., 2020. Most of Arabic tweets use the informal Arabic (dialects) which needs an effective method to analyses Arabic tweets to understand the global imputation of Covid-19 influence, Mansoor et al., 2020, Basiri et al., 2021. In this paper, we proposed to use several models to perform sentiment analysis, related to the COVID-19 pandemic. The proposed sentiment analysis models use various Machine Learning algorithms. This paper is organized as following: section 2 contains related works, while section 3 describes the proposed model, then, in section 4 experiment and tests are presented and finally, section 5 is the conclusion.

2. Related Work

During the Covid-19 outbreak, we witnessed a great increase in using social media platforms like twitter that became the favorite source to express public opinions. Several studies addressed sentiment analysis for covid-19 Tweets. Albahli,2022 proposed sentiment analysis model using Machine Learning and SMOTE for imbalanced dataset handling. The author used a dataset focused on the gulf countries, i.e., Oman, Qatar, Bahrain, Saudi Arabia, and United Arab Emirates (UAE). Their experiment results showed most of the opinions had a negative sentiment during COVID-19 pandemic.

Al-Ayyoub et al.,2019 proposed many methods using the Opinion Corpus for Arabic (OCA) which contains 500 Arabic movies reviews. They used several Machine Learning algorithms including Support Vector Machine (SVM) and Naïve Bayes (NB), which are, to predict classification of positive and negative reviews. Abo et al.,2019 describes a systematic mapping study (SMS) of 51 primary selected studies (PSS) handled with the approval of an evidence-based systematic method to ensure handling of all related papers. Al-Smadi et al., 2019 classified Arabic hotel reviews using Aspect-Based Sentiment Analysis (ABSA) by SVM, Decision Tree, Naïve Bayes, and K-NN (K Nearest-Neighbours) algorithm using WEKA Classifiers. Their result showed that K-NN performs better than the SVM, obtaining around 85.3% accuracy.

Gamal et al.,2019 proposed methods using Machine Learning algorithms including Naïve Bayes, Multinomial Naïve Bayes, Adaptive Boosting, Logistic Regression, Stochastic Gradient Decent (SGD), PA, RR, and SVM for sentiment analysis of social media platforms in the Arabic language. In Elnagar et al., 2020, authors performed deep learning method on text classification in Arabic language. Their research used single-label SANAD and multi-label NADiA datasets. Their experiment results showed that SANAD performed better with an accuracy of 96.4%, while the attention-GRU got an accuracy of 88.64%.

In Bhatia et al., 2022, authors discussed how COVID-19 allowed people to use social media rapidly in Arab countries, specifically people using Twitter as their main source of news. The authors addressed the spread of COVID-19 in Arab countries and also gave the solution to use the different approaches for sentiment analysis by using machine learning and deep learning approaches to get the sentiments on Arabic

tweets' dataset. Their model was evaluated and got 84% accuracy value using DNN algorithm. Ali et al.,2021 proposed a model for sentiment analysis in terms of finding correlation between tags to understand multi-tag learning algorithms.

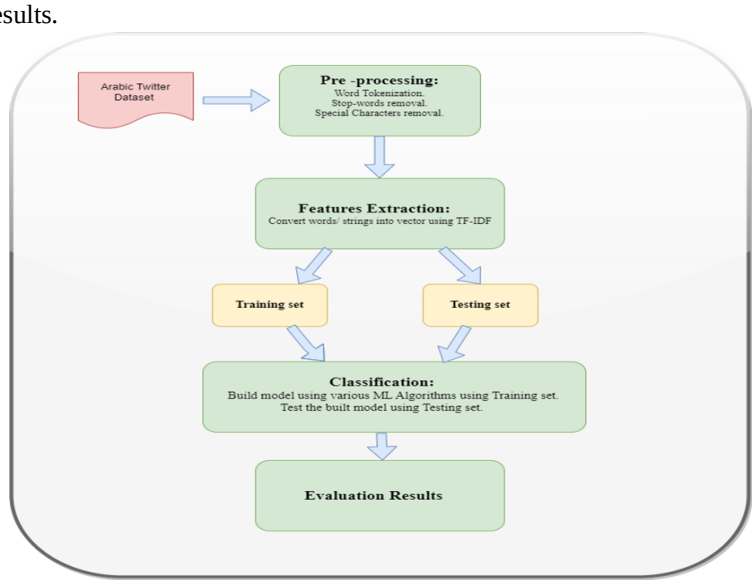
3. Proposed Model

In this paper, the proposed model worked on sentiment analysis for Arabic text in using Twitter dataset. We used word-cloud to provide quick insights at a glance of the dataset as showed in figure 1.

Fig. 1. Dataset Word-cloud

Python3 and many libraries were used for developing all of the codes for this sentiment analysis task. As shown in figure 2, the proposed model consists of many steps including:

1. Data Preprocessing: this step contains many processes consist of data-cleaning.
2. Extracting Features using TF-IDF.
3. Machine Learning Algorithm Predication.
4. Evaluation results.



4. Experiments

4.1. Platform and software

The experiments are conducted on a laptop with an i5 processor with 8GB RAM, using an Anaconda Jupyter notebook (python3).

4.2. Dataset and Processing Steps

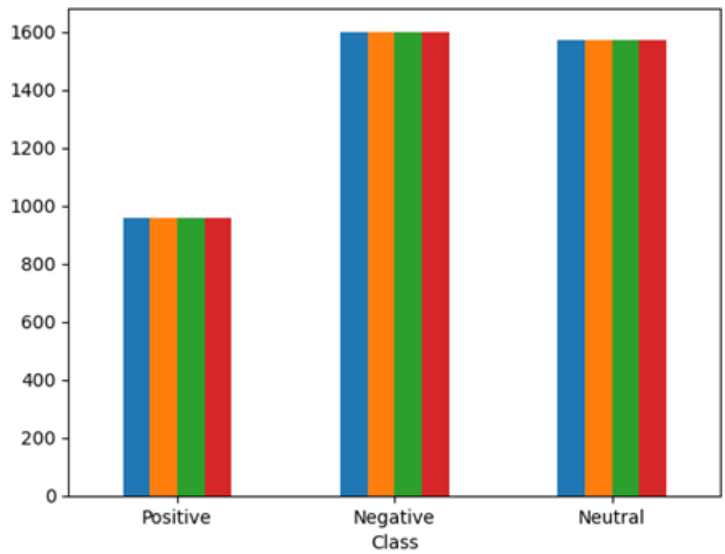


Fig. 2. Distribution of the Twitter Dataset

The labeled training dataset consists of 1600 Neutral, 1571 Positive, and 957 Negative tweets and the testing dataset contains 1034 opinions obtained from the organizer team of Arabic sentiment analysis and fake news detection within covid-19 task.

The experiment includes the following steps:

1. Cleaning/pre-processing dataset:

Pre-processing means prepared datasets for a purpose of training and testing. NLTK package in Python is used for removing stop words, tokenization, removing special characters, and stemming.

2. Extracting Features using TF-IDF.

Convert the dataset into vectors as preparing for machine learning modeling.

3. Applying the classification algorithms.

The following Machine Learning algorithms are utilized for classification purposes:

- Support Vector Machine
- Logistic-Regression
- K-Nearest Neighbor
- Multinomial Naïve Bayes

4. Evaluate output as positive, negative, neutral opinions from tweets.

4.3. Results

In our experiment, we trained more than one classifier (i.e., SVM, LR, NB and KNN). To extract the most useful features to train models with highest accuracy several methods of preprocessing were used. Four models were trained and tested, finally the results are compared. Precision, recall and f- measure scores of implemented classifiers are summarized in figure 4 and in Table 1, 2, 3 and 4. When evaluating the results, it is showed that Logistic-Regression and SVM have highest accuracy scores than other classifiers with 94%.

Table 1. SVM Scores

Class	Precision	Recall	F1-score	Accuracy
Negative	0.94	0.94	0.94	
Neutral	0.90	0.98	0.94	94%
Positive	0.98	0.90	0.94	

Table 2. K-Neighbors

Class	Precision	Recall	F1-score	Accuracy
Negative	0.92	0.85	0.89	
Neutral	0.93	0.94	0.94	92%
Positive	0.90	0.93	0.94	

Table 3. Logistic Regression

Class	Precision	Recall	F1-score	Accuracy
Negative	0.94	0.92	0.93	
Neutral	0.90	0.98	0.94	94%
Positive	0.97	0.90	0.94	

Table 4. SVM Scores

Class	Precision	Recall	F1-score	Accuracy
Negative	0.93	0.82	0.87	
Neutral	0.87	0.95	0.90	89%
Positive	0.89	0.88	0.89	

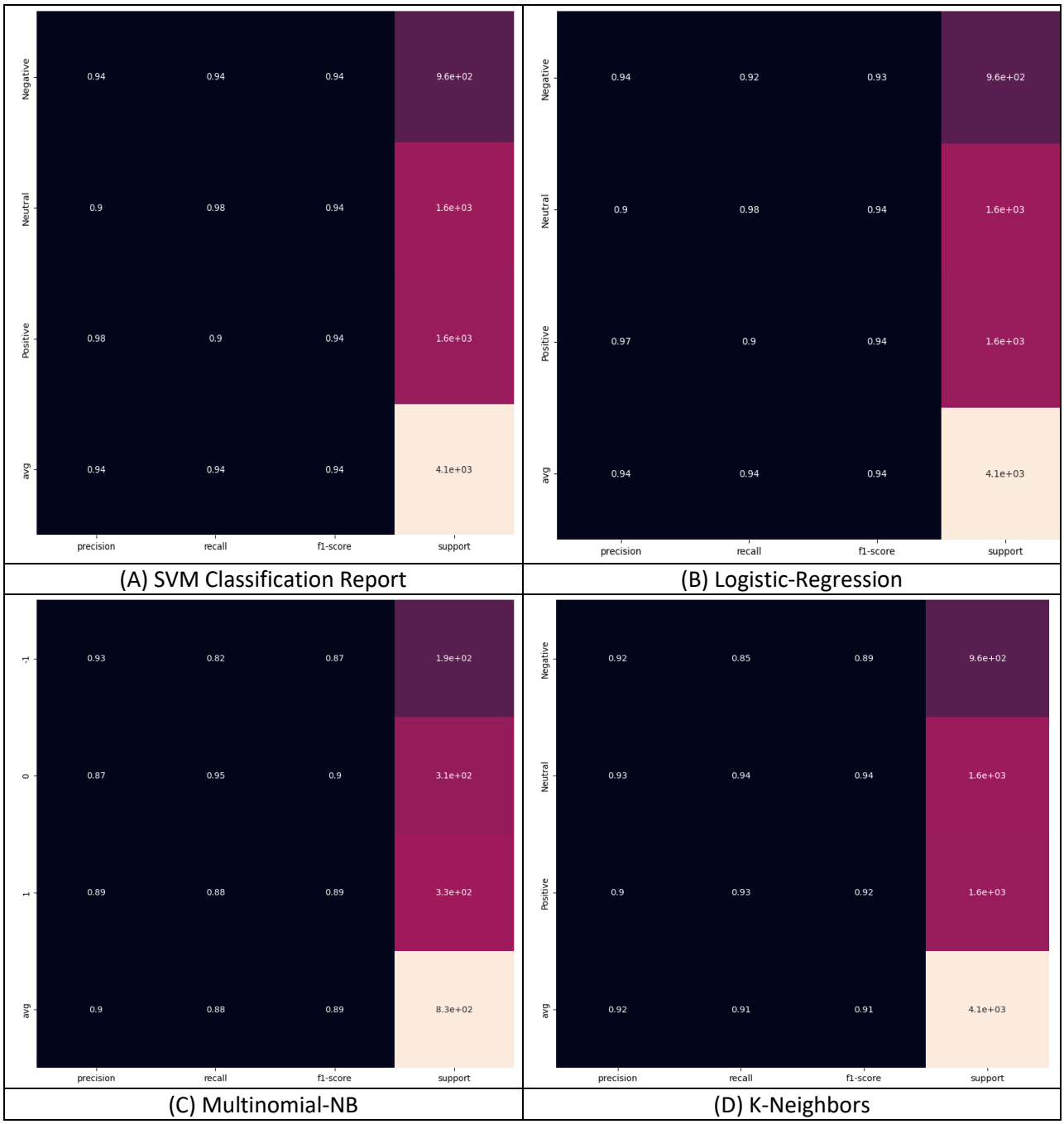


Fig. 3. Classification scores

5. Conclusion

Since the COVID-19 pandemic began people rapidly used twitter to express their thought and opinions. This paper investigated the opinions related to Covid-19 using sentiment analysis in Arabic twitter review. Various Machine learning algorithms are used and the experiments results show that Logistic-Regression and SVM have the highest scores with 94% accuracy.

References

Abo, M. E. M., Raj, R. G., Qazi, A., Zakari, A., 2019. Sentiment analysis for arabic in social media network: A systematic mapping study. arXiv preprint arXiv:1911.05483.

Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., Alaa, M., 2021. Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. Expert systems with applications, 167, 114155.

Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., Al-Kabi, M. N., 2019. A comprehensive survey of arabic sentiment analysis. Information processing & management, 56(2), 320-342.

Albahli, S., 2022. Twitter sentiment analysis: An Arabic text mining approach based on COVID-19. Frontiers in Public Health, 10, 966779.

Ali, M. M., 2021. Arabic sentiment analysis about online learning to mitigate covid-19. Journal of Intelligent Systems, 30(1), 524-540.

Al-Smadi, M., Al-Ayyoub, M., Jararweh, Y., Qawasmeh, O., 2019. Enhancing aspect-based sentiment analysis of Arabic hotels' reviews using morphological, syntactic and semantic features. Information Processing & Management, 56(2), 308-319.

Basiri, M. E., Nemati, S., Abdar, M., Asadi, S., Acharrya, U. R., 2021. A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets. Knowledge-Based Systems, 228, 107242.

Bhatia, S., Alhaider, M., Alarjani, M., 2022. Sentiment analysis for Arabic Tweets on Covid-19 using computational techniques. In 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 559-566). IEEE.

Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., Hassanien, A. E., 2020. Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. Applied Soft Computing, 97, 106754.

Elnagar, A., Al-Debsi, R., Einea, O., 2020., Arabic text classification using deep learning models. Information Processing & Management, 57(1), 102121.

Gamal, D., Alfonse, M., El-Horbaty, E. S. M., Salem, A. B. M., 2019. Implementation of machine learning algorithms in Arabic sentiment analysis using n-gram features. Procedia Computer Science, 154, 332-340.

Hadj Ameur M.S, Aliane, H., 2021. AraCOVID19-MFH: Arabic COVID-19 multi-label fake news & hate speech detection dataset. Procedia Computer Science 189, 232-241.2021.

Lad, S., Mane, G., Padwal, A., Dixit, M., 2022. Machine learning based sentiment analysis of Twitter data: AIP Conf. Proc. 31 October 2022; 2494 (1): 050007.

Nemes, L., Kiss, A., 2021. Social media sentiment analysis based on COVID-19. Journal of Information and Telecommunication, 5(1), 1-15.

Manguri, K. H., Ramadhan, R. N., Amin, P. R. M., 2020. Twitter sentiment analysis on worldwide COVID-19 outbreaks. Kurdistan Journal of Applied Research, 54-65.

Mansoor, M., Gurumurthy, K., Prasad, V.R., 2020. Global sentiment analysis of COVID-19 tweets over time. arXiv preprint arXiv:2010.14234.

Pokharel, B. P., 2020. Twitter sentiment analysis during covid-19 outbreak in nepal. Available at SSRN 3624719.

Srikanth, J., Damodaram, A., Teekaraman, Y., Kuppusamy, R., Thelkar, A. R.,2022. Sentiment analysis on COVID-19 Twitter data streams using deep belief neural networks. Computational intelligence and neuroscience, 2022.

Syarief, M. G., Kurahman, O. T., Huda, A. F., Darmalaksana, W.,2019. Improving Arabic stemmer: ISRI stemmer. In 2019 IEEE 5th International Conference on Wireless and Telematics (ICWT) (pp. 1-4). IEEE.