Sentimental Analysis of COVID-19 Vaccine Tweets using BERT+NBSVM

Areeba Umair[1][0000-0003-3638-7369], Elio Masciari[1], Giusi Madeo[2], and Muhammad Habib Ullah[1]

[1] University of Naples, Federico II, 80125 Naples, Italy
[2] IC Rende Commenda, Rende (CS), Italy

Abstract. The development of the vaccine for the control of COVID-19 is the need of hour. The immunity against coronavirus highly depends upon the vaccine distribution. Unfortunately, vaccine hesitancy seems to be another big challenge worldwide. Therefore, it is necessary to analysis and figure out the public opinion about COVID-19 vaccines. In this era of social media, people use such platforms and post about their opinion, reviews etc. In this research, we proposed BERT+NBSVM model for the sentimental analysis of COVID-19 vaccines tweets. The polarity of the tweets was found using TextBlob(). The proposed BERT+NBSVM outperformed other models and achieved 73 % accuracy, 71 % precision, 88 % recall and 73 % F-measure for classification of positive sentiments while 73 % accuracy, 71 % precision, 74 % recall and 73 % F-measure for classification of negative sentiments respectively. Thus, these sentimental and spatial analysis helps in world-wide pandemics by identify the people's attitudes towards the vaccines.

Keywords: Sentimental analysis · Vaccine · COVID-19 · Vaccine hesitancy

1 Introduction

The COVID-19, caused by coronavirus, started spreading an infectious disease in December, 2019 in Wuhan, China [4], [14]. The preventive measures of social distancing and wearing masks were observed in different countries. However, the long-term solution of this disease was the development of vaccines [6], [30]. Moreover, after the vaccines were developed, the acceptance of vaccines among the general public was next milestone [16]. Most of the population from all over the world was not willing to get themselves vaccinated because of the its side-effects and other misinformation [12], [2], [17]. To overcome this situation, it is a good strategy if the agencies put their efforts in understanding what people are thinking about vaccines and design their strategies accordingly [1], [28].

The increasing number of social media platforms users made it easy for researchers to find and extract the user-generated freely data [31]. Such kind of data can easily be used for public sentiments analysis [10]. This data can be very helpful in investigating the people's behaviour during this disease, during lockdown and for the vaccine campaign [15], [34]. Twitter is considered as the most popular social media app which has been used worldwide for sharing the feeling, opinion and ideas [7] and [25]. Tweets can be useful for analysing the people's feedback on any trending topic [18]. Sentimental analysis is the most famous method by which people's feelings are extracted [9]. It uses machine learning and deep learning for this purpose [15], [13]. For performing the sentimental analysis, the polarity of the tweet is found [22].

In this study, we used tweets related to COVID-19 vaccine and used them for extracting the sentiments of people towards vaccination of COVID-19. This research can bring fruitful results for government and policy makers for designing the vaccination campaign according to the sentiments of people. We used freely available twitter data from Kaggle website and found out the polarity of the tweets. At the end of the method, we used BERT+NBSVM model for classification of positive and negative sentiments.

The objectives of this research paper are:

- Using freely available twitter data about COVID vaccines and categorize the text into different sentiment classes.
- To categorize the tweets based on their polarity values using python script.
- To propose BERT+NBSVM model for positive and negative tweet classification.

2 Methods

Machine learning and natural language processing are used for sentimental analysis.

2.1 Sentimental Classification Framework

Our proposed sentimental classification framework consists of three stages as we can see in figure 1.

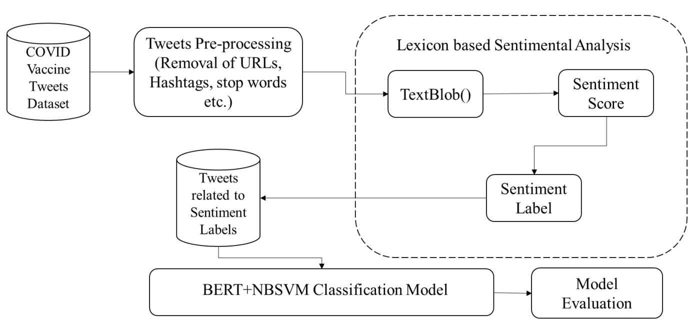

Fig. 1. Our proposed Sentimental Classification Framework

First, the dataset is collected and pre-processing is performed.

Secondly, the sentiment polarity is extracted using TextBlob() function.

Third, polarity values are used to classify the positive and negative tweets with the help of BERT+NBSVM model.

2.2 Collection and pre-processing of data

In this study, we used freely available twitter data. The dataset contains tweets about COVID-19 vaccines. The dataset is then further processed by removing URLs, hashtags, and stop-words using python script. In table 1, column 1 shows the dummy tweets, column 2 shows tweets after hashtag removal, column 3 shows tweets after URLs removal.

Table 1. Comparison of tweets before and after pre-processing

Dummy samples	**After removing Hashtags**	**After removing URLs**
Fever after first dose #PfizerBioNTech https://t.co/xffiee77	Fever after first dose PfizerBioNTech https://t.co/xffiee77	Fever after first dose PfizerBioNTech
Vaccine scheduling available online https://t.co/jgeeityc	Vaccine scheduling available online https://t.co/jgeeityc	Vaccine scheduling available online
Any update on booster dose?? https://t.co/hdrryuugy	Any update on booster dose https://t.co/hdrryuugy	Any update on booster dose

2.3 Finding Values of Sentiments Polarity

The sentiment analysis depends upon the polarity of the sentence. The polarity shows that either the given text is neutral, negative or positive. We categorized the tweets into seven classes of sentiments [27] based on the polarity values. The classes includes neutral, weakly positive, mild positive, strongly positive, weakly negative, mild negative and strongly negative. We used principles of [27] as given in Table 2.3 to fix the polarity range of each class. We find the polarity using TextBlob() library of Python, which returns polarity between -1 to +1.

We find the polarity values (between [-1 to +1]) of the given tweets using the TextBlob() library function of Python. The working principle of TextBlob() can be seen in Figure 2.

2.4 Combining BERT and Naive Bayes-SVM for Sentimental Classification

In this research, we combined the Bidirectional encoder representation of transformers (BERT) with hybrid of Naive Bayes and Support Vector Machine (NB-SVM). BERT is the transformer based model which used attention mechanism.

Polarity	Sentiment Class
0	Neutral
> 0 and ≤ 0.3	Weakly Positive
> 0.3 and ≤ 0.6	Mild Positive
> 0.6 and ≤ 1	Strongly Positive
> -0.3 and ≤ 0	Weakly Negative
> -0.6 and ≤ -0.3	Mild Negative
> 1 and ≤ -0.6	Strongly Negative

Table 2. Rules for sentimental classes

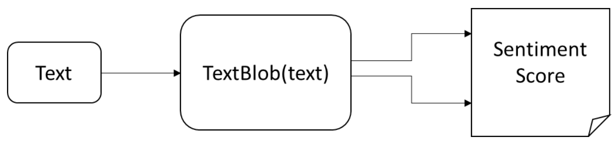

Fig. 2. How TextBlob() works?

In transformer, encoder and decoder both are used, while in BERT model, only encoder layers of transformers are used [20] [32]. The two famous architectures of BERT are base and large models. Both of the models have four differences between them [20].

Naive Bayes and Support vector machines are the machine learning algorithms, the former works good on short sentimental tasks while the later on longer documents. The hybrid of NB and SVM uses the variances of SVM and ratio of log NB for better accuracy [19].

BERT+NBSVM system architecture The combination of deep learning and classical machine learning results in the BERT+NB-SVM model, which is estimated on DTM (document term frequency) features. The DTM is used to compute the NB log-count ratios. These ratios helps to calculate the word probability of positive and negative classes in a document. The system architecture of BERT+NB-SVM is shown in Figure 3. It can be seen from the figure that, the left side shows the process of training while the right side shows the classification.

The following steps are adopted for training and classification:

- The training dataset was used in fine tuning of BERT model
- The NB log-count ratios are used for SVM model training
- While prediction, final score is calculated as the weighted sum of the fitted NB-SVM model and best fine-tuned BERT model.
- The best fine tuned model indicates the model with best performances with different epochs and batch sizes.

BERT+NBSVM Model Training To train the model, we used pre-training and fine tuning. As a loss function, Adam optimizer is used to train the model

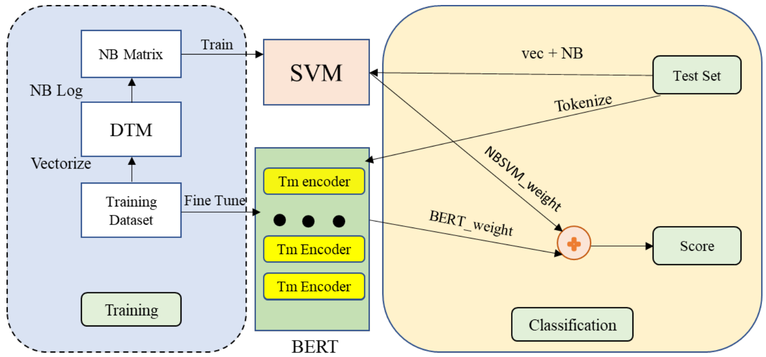

Fig. 3. System architecture of BERT+ NBSVM

and grid search was used for parameter tuning. These best weight for BERT model is 0.87 and NB-SVM is 0.08. Two classification models, (positive and negative). The precision, recall and F_1 score are used, as shown in equation 1, equation 2 and equation 3) respectively.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$FMeasure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

Where:
TP occurs when both item and the result are positive.
TN occurs when both item and the result are negative
FP occurs when item is negative while model is giving positive result.
FN occurs when item is positive while model is giving negative result.

State-of-the-art We performed the experiments with state-of-the-art in order to evaluate our proposed model. We used K-nearest neighbour (KNN) algorithm, Support Vector Machine (SVM) algorithm, Random Forest (RF) algorithm, Naive Bayes (NB) algorithm and DT (Decision Tree) algorithm because of their being mostly used in literature [4], [15] and [24].

Decision tree and random forest have ability to learn from uses [33]. Random forest works good on non-linear datasets. It chooses randoms samples and features from the dataset [23]. While, decision tree works on decision rules from the entire dataset. It works well on small dataset [5]. Naive Bayes uses the principles of probability for its working. It considers all the features statistically independent [5] [3]. Naive Bayes uses Bayesian theorem and calculates the probability of the items as we can seen in equation 4:

$$P(H|X) = P(X|H)P(H)/P(X) \tag{4}$$

The simplest machine learning algorithm, KNN, looks for the most similar item among its neighbours. It requires alot of time, as it needs to search from the entire dataset. Therefore, it is good for the small datasets. Moreover, it is the algorithm, that does not follow test-train mechanism. User provides the number of neighbours during search [3]. The equation 5) is used for find the most similar item.

$$di = \sqrt{[(xi - x)^2 + (yi - y)^2]} \tag{5}$$

SVM is the effective algorithm when the dataset contains high dimensional feature space. It works by generating the hyperplane. The hyperplane thus helps in classification. It combines features from different sources and make one feature to train the model. The higher the separation of hyperplane, the more accurate the classification. Linear function, polynomial function and radial basis function are the kernels functions used in SVM [8].

BERT works on the principles of masked languages (MLM) by using word representation model. It has [SEP] and [CLS] as two special tokens. BERT takes input as [CLS] and then transfers it to the upper layer. At that step, the self attention is applied. The output from this step is transferred to the feed-forward network. The vector C, output of the model, is used for classification and translation. The probability of sentimental classes can be calculated by following equation 6 [29].

$$P = softmax(CW^T) \tag{6}$$

3 Results and discussion

3.1 Sentiment Polarity

The polarity value helps to find the sentiment of the text. Table 3, shows the polarity values for each text with the sentiments category, described in section 2.3

Table 3. Categorization of sentiments on sample data

Data Sample	Polarity	Category
Fever after first dose PfizerBioNTech	-0.5	Mild Negative
Vaccine scheduling available online	0.7	Strongly Positive
Any update on booster	0	Neutral

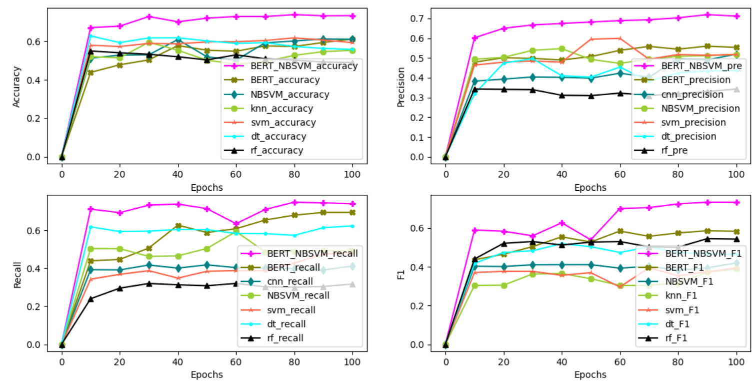

Fig. 4. Graph showing the results of experiments for positive sentiments

3.2 Sentimental Classification

The results of the experiments are shown in the Figure below:

Figure 4 shows the sub-graphs depicting the classification accuracy, precision, recall and F1 score of our proposed BERT+NBSVM model in comparison with BERT, NBSVM, Decision tree, KNN, random forest and SVM for the classification of positive sentiments. The classification results of positive tweets classification show that our proposed approach outperformed all other state of the art models. The proposed BERT+NBSVM showed the best accuracy.

Figure 5 shows the sub-plots of accuracy, precision, recall and F1 score for our proposed BERT+NBSVM model in comparison with BERT, NBSVM, Decision tree, KNN, random forest and SVM for the classification of negative sentiments. The BERT+NBSVM showed best performance among all other state of the art neural network and machine learning models that have been used in literature. Deep learning has attracted the attention due to its prediction performance in the social media domain. Out of all baseline neural network models, BERT+NBSVM outperformed all others.

Among machine learning models, the performance of SVM was high among other baseline algorithms because SVM does not show any effect of hyper-parameters related to data [26]. KNN and decision trees were found with similar accuracy and they show a significant effect in the classification [11]. Random forest shows the intermediate performance in both of the scenarios of our study, because random forest draws observation strategies randomly and requires a hyper-parameter tuning for good performance [21].

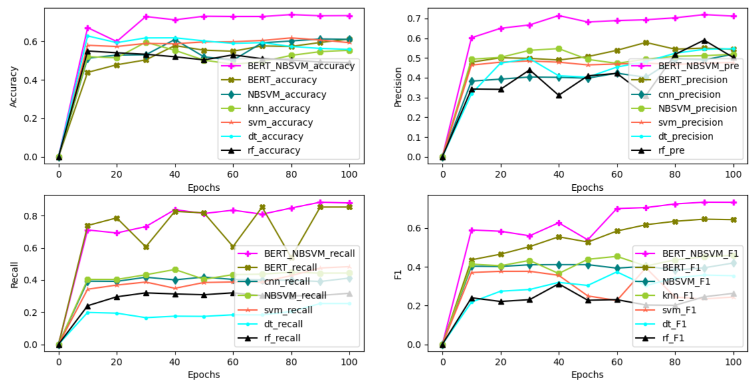

Fig. 5. Graph showing the results of experiments for negative sentiments

4 Conclusion

Twitter based sentimental analysis for extraction people's response towards any general issue or topic is a very fruitful and efficient way for policy makers. Vaccine hesitancy is a hurdle in the control of COVID-19 disease and is emerged as a bigger challenge worldwide. In this research, people's reaction during COVID-19 vaccination campaigns are analyzed using twitter data. We categorized the data into seven categories of sentiments using their polarity. We proposed the BERT+NBSVM classification models (positive and negative) for sentiment classification. Hence, such kind of research can help the policy makers and government to understand what people are thinking about their campaigns and initiatives. So, they can educate people timely about any misinformation regarding any campaigns and thus can save the lives of citizens from any disease or epidemic.

References

1. Examining Australian public perceptions and behaviors towards a future COVID-19 vaccine. medRxiv pp. 1–9 (2020). https://doi.org/10.1101/2020.09.29.20204396
2. Knowledge and attitudes of us adults regarding COVID-19. Int. J. Emerg. Med. **13**(1), 1–6 (2020). https://doi.org/10.1186/s12245-020-00309-6
3. Abdulrahman, N., Abedalkhader, W.: KNN Classifier and Naive Bayse Classifier for Crime Prediction in San Francisco Context. Int. J. Database Manag. Syst. **9**(4), 1–9 (2017). https://doi.org/10.5121/ijdms.2017.9401
4. Adamu, H., Lutfi, S.L., Malim, N.H.A.H., Hassan, R., Di Vaio, A., Mohamed, A.S.A.: Framing twitter public sentiment on Nigerian government COVID-19 palliatives distribution using machine learning. Sustain. **13**(6) (2021). https://doi.org/10.3390/su13063497

5. Almanie, T., Mirza, R., Lor, E.: Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots. Int. J. Data Min. Knowl. Manag. Process **5**(4), 1–19 (2015). https://doi.org/10.5121/ijdkp.2015.5401
6. Chou, W.Y.S., Budenz, A.: Considering Emotion in COVID-19 Vaccine Communication: Addressing Vaccine Hesitancy and Fostering Vaccine Confidence. Health Commun. **35**(14), 1718–1722 (2020). https://doi.org/10.1080/10410236.2020.1838096, https://doi.org/10.1080/10410236.2020.1838096
7. Das, S., Dutta, A.: Characterizing public emotions and sentiments in COVID-19 environment: A case study of India. J. Hum. Behav. Soc. Environ. **31**(1-4), 1–14 (2020). https://doi.org/10.1080/10911359.2020.1781015, https://doi.org/10.1080/10911359.2020.1781015
8. Dqj, L.X., Hqj, L.X., Hpsor, D.H., Suhglfw, H.H.V., Ixwxuh, W.K.H., Vxssruw, D.Q.G., Dwlrqv, R., Wkh, S., Dqg, I., Exvlqhvvhv, V., Wr, D., Hpsor, I., Survshfw, H.H., Wkhvh, Z.: Liu, J., Wang, T., Li, J., Huang, J., Yao, F., & He, R. In: IEEE Int. Conf. Syst. Man Cybern. pp. 4056–4062 (2019)
9. Flesca, S., Furfaro, F., Masciari, E.: On the minimization of xpath queries. J. ACM **55**(1), 2:1–2:46 (2008). https://doi.org/10.1145/1326554.1326556, https://doi.org/10.1145/1326554.1326556
10. Flesca, S., Masciari, E.: Efficient and effective web change detection. Data Knowl. Eng. **46**(2), 203–224 (2003). https://doi.org/10.1016/S0169-023X(02)00210-0, https://doi.org/10.1016/S0169-023X(02)00210-0
11. George, J., Skariah, S.M., Xavier, T.A.: Role of contextual features in fake news detection: a review. In: 2020 international conference on innovative trends in information technology (ICITIIT). pp. 1–6. IEEE (2020)
12. Green, M.S., Abdullah, R., Vered, S., Nitzan, D.: A study of ethnic, gender and educational differences in attitudes toward COVID-19 vaccines in Israel – implications for vaccination implementation policies. Isr. J. Health Policy Res. **10**(1), 1–12 (2021). https://doi.org/10.1186/s13584-021-00458-w
13. H. Manguri, K., N. Ramadhan, R., R. Mohammed Amin, P.: Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks. Kurdistan J. Appl. Res. pp. 54–65 (2020). https://doi.org/10.24017/covid.8
14. Huang, H., Peng, Z., Wu, H., Xie, Q.: A big data analysis on the five dimensions of emergency management information in the early stage of COVID-19 in China. J. Chinese Gov. **5**(2), 213–233 (2020). https://doi.org/10.1080/23812346.2020.1744923, https://doi.org/10.1080/23812346.2020.1744923
15. Jelodar, H., Wang, Y., Orji, R., Huang, H.: Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. arXiv **24**(10), 2733–2742 (2020)
16. Kourlaba, G., Kourkouni, E., Maistreli, S., Tsopela, C.G., Molocha, N.M., Triantafyllou, C., Koniordou, M., Kopsidas, I., Chorianopoulou, E., Maroudi-Manta, S., Filippou, D., Zaoutis, T.E.: Willingness of Greek general population to get a COVID-19 vaccine. Glob. Heal. Res. Policy **6**(1), 1–10 (2021). https://doi.org/10.1186/s41256-021-00188-1
17. Lazarus, J.V., Ratzan, S.C., Palayew, A., Gostin, L.O., Larson, H.J., Rabin, K., Kimball, S., El-Mohandes, A.: A global survey of potential acceptance of a COVID-19 vaccine. Nat. Med. **27**(2), 225–228 (2021). https://doi.org/10.1038/s41591-020-1124-9, http://dx.doi.org/10.1038/s41591-020-1124-9

18. Luo, Y., Xu, X.: Comparative study of deep learning models for analyzing online restaurant reviews in the era of the COVID-19 pandemic. Int. J. Hosp. Manag. **94**(December 2020), 102849 (2021). https://doi.org/10.1016/j.ijhm.2020.102849, https://doi.org/10.1016/j.ijhm.2020.102849
19. Muhammad, A.N., Bukhori, S., Pandunata, P.: Sentiment analysis of positive and negative of youtube comments using naïve bayes–support vector machine (nbsvm) classifier. In: 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE). pp. 199–205. IEEE (2019)
20. Pota, M., Ventura, M., Catelli, R., Esposito, M.: An effective bert-based pipeline for twitter sentiment analysis: A case study in Italian. Sensors (Switzerland) **21**(1), 1–21 (2021). https://doi.org/10.3390/s21010133
21. Probst, P., Wright, M.N., Boulesteix, A.L.: Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery **9**(3), e1301 (2019)
22. Raheja, S., Asthana, A.: Sentimental analysis of twitter comments on COVID-19. Proc. Conflu. 2021 11th Int. Conf. Cloud Comput. Data Sci. Eng. pp. 704–708 (2021). https://doi.org/10.1109/Confluence51648.2021.9377048
23. Rangnekar, R.H., Suratwala, K.P., Krishna, S., Dhage, S.: Career Prediction Model Using Data Mining And Linear Classification. In: Fourth Int. Conf. Comput. Commun. Control Autom. pp. 1–6 (2018), https://ieeexplore.ieee.org/abstract/document/8697689/
24. Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., Choi, G.S.: A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. PLoS One **16**(2), 1–23 (2021). https://doi.org/10.1371/journal.pone.0245909, http://dx.doi.org/10.1371/journal.pone.0245909
25. Samuel, J., Rahman, M.M., Ali, G.G.N., Samuel, Y., Pelaez, A., Chong, P.H.J., Yakubov, M.: Feeling Positive about Reopening? New Normal Scenarios from COVID-19 US Reopen Sentiment Analytics. IEEE Access **8**, 142173–142190 (2020). https://doi.org/10.1109/ACCESS.2020.3013933
26. Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., Brenning, A.: Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecological Modelling **406**, 109–120 (2019)
27. Singh, M., Jakhar, A.K., Pandey, S.: Sentiment analysis on the impact of coronavirus in social life using the BERT model. Soc. Netw. Anal. Min. **11**(1), 1–11 (2021). https://doi.org/10.1007/s13278-021-00737-z, https://doi.org/10.1007/s13278-021-00737-z
28. Sv, P., Ittamalla, R., Deepak, G.: Analyzing the attitude of Indian citizens towards COVID-19 vaccine e A text analytics study (January) (2020)
29. Umair, A., Masciari, E.: Using high performance approaches to covid-19 vaccines sentiment analysis. In: 2022 30th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP). pp. 197–204. IEEE (2022)
30. Umair, A., Masciari, E., Habib Ullah, M.H.: Sentimental analysis applications and approaches during covid-19: A survey. In: 25th International Database Engineering & Applications Symposium. p. 304–308. IDEAS 2021, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3472163.3472274, https://doi.org/10.1145/3472163.3472274
31. Umair, A., Sarfraz, M.S., Ahmad, M., Habib, U., Ullah, M.H., Mazzara, M.: applied sciences Spatiotemporal Analysis of Web News Archives for Crime Prediction (2020). https://doi.org/10.3390/app10228220

32. Yadav, N., Singh, A.K.: Bi-directional Encoder Representation of Transformer model for Sequential Music Recommender System. ACM Int. Conf. Proceeding Ser. pp. 49–53 (2020). https://doi.org/10.1145/3441501.3441503
33. Yi, F., Yu, Z., Xu, H., Guo, B.: Talents Recommendation with Multi-Aspect Preference Learning. Green, Pervasive, Cloud Comput. **11204**, 409–423 (2018). https://doi.org/10.1007/978-3-030-15093-8$_2$9, $http : //link.springer.com/$10.1007/978 $-$ 3 $-$ 030 $-$ 15093 $-$ 8_29
34. Zhou, B., Cheng, C., Ma, G., Zhang, Y.: Remaining Useful Life Prediction of Lithium-ion Battery based on Attention Mechanism with Positional Encoding. IOP Conf. Ser. Mater. Sci. Eng. **895**(1), 0–9 (2020). https://doi.org/10.1088/1757-899X/895/1/012006