International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 07 | July - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Optimization Accuracy of Tweets for Coronavirus Pandemic using BERT Based CNN Model

Abhishek Kumar
M. Tech. Scholar
Department of Information Technology
Technocrats Institute of Technology
Bhopal, India

Dr. Neetesh Kumar Gupta
Technocrats Institute of Technology
Bhopal, India

Abstract:- It assures illustrating the results of the survey and displaying series of data that is derived from the primary collection of data. In this context, it assures descriptive form of analysis of the collected primary data. It also assures summarising the collected data and information with descriptive form of statistics. It interprets the results without any bias regarding the primary data collected from survey due to its quantitative nature. In this paper coronavirus pandemic using hybrid bidirectional encoder representations from transformers (BERT) with convolution neural network (CNN) is present. At the startof training of CNN, filters are initialized with random values and as the model gets trained by back-propagation algorithm using the errors in estimating the actual output, the weights of the filters are modified to identify specific patterns in inputs. We show that combining CNN with BERT is better than using BERT and CNN on its own, and we emphasize the importance of utilizing pre-trained coronavirus models for downstream tasks. The hybrid model is implemented python platform and calculates accuracy, precision, recall and F1-score.

Keywords:- CNN, BERT, Accuracy, Recall, F1-score

I. INTRODUCTION

Moreover, it involves analyzing potential features that are required for suitably incorporating the solutions within the organizational standards. Additionally, it also engages analysing new form of product attributes, which will be necessary from different providers of UCaaS during the upcoming post-pandemic situation in Indian context. This study enhances both "Primary and secondary" collection of data, thereby delivering a "mixed method of data collection" [1, 2]. It enables the study for an in-depth analysis of the research context while concentrating on both real-time as well as existing data and information collection for bringing out most suitable outcomes of the thesis. In this context, for collecting the "primary data", a questionnaire is delivered, that comprises of close-ended questions asked to employees of large organizations based in India. Additionally, "Personal Interview", is further carried out and an "online survey", is conducted for marking the feasibility of collecting adequate data and information regarding reasons as well as plans for implementation of UCaaS within organizational standards [3].

The collection of primary-data enables the researcher for resolving specific context of the research issues, therefore performing proper research while gaining best form of accuracy attributes. It also allows higher standards of control over the research and collection of real-time data through the employees along with assuring no bias in the analysis of the collected data. It assures up-to-date information collection and the study assures direct addressing of the thesis context [4].

In this aspect, it will involve an in-depth analysis along with summarising potential existing form of data and information, which is further collated for increasing the best possible outcomes of the research [5, 6]. It will also help to analyse and interpret the collected data while bridging out gaps along with potential deficiencies. It will also bring out a proper understanding of various additional information and data, which requires being properly collected. Therefore, this type of collection of data will assure improving understanding related to the research problem. It will also add proper support to the primary data, that is already collected for analysing the context of the research in a more effective and efficient manner.

This research implements "*Simple Random sampling and Deliberate sampling"* for collection of primary data and incorporates "*Purposive Sampling*", for collection of secondary data as well as information [7, 8]. In this context, for collection of the primary form of data and information it assures incorporation of individual form of knowledge workers in big organizations in Indian context. It also assures top form of organizations, which is further equipped with experienced workers. In this context, it also evaluates qualitative form of methods for assuring best possible evaluation of research findings. For collecting the potential "*secondary data*" and information, this thesis utilises potential data from relevant texts, conference papers, articles, and journals along with potential magazines.

II. RELATED WORK

Xiongwei Zhang et al. [1], twitter is a virtual informal organization where individuals share their posts and feelings about the ongoing circumstance, for example, the Covid pandemic. It is viewed as the main streaming

information hotspot for AI research with regards to investigation, expectation, information extraction, and feelings. Opinion investigation is a text examination technique that has acquired further importance because of interpersonal organizations' development. Thusly, this paper presents a constant framework for opinion expectation on Twitter streaming information for tweets about the Covid pandemic. **J. Samuel et al. [2],** the proposed framework means to find the ideal AI model that acquires the best exhibition for Covid feeling examination expectation and afterward involves it progressively. The proposed framework has been formed into two parts: fostering a disconnected feeling examination and demonstrating an internet based expectation pipeline. The framework has two parts: the disconnected and the internet based parts. For the disconnected part of the framework, the authentic tweets' dataset was gathered in span 23/01/2020 and 01/06/2020 and sifted by #COVID-19 and #Coronavirus hashtags. Two component extraction techniques for literary information examination were utilized, n-gram and TF-ID, to remove the dataset's fundamental elements, gathered utilizing Covid hashtags. **R. Ali et al. [3],** five normal AI calculations were performed and thought about: choice tree, strategic relapse, k-closest neighbors, arbitrary backwoods, and backing vector machine to choose the best model for the web-based expectation part. **G. Barkur et al. [5],** it assures illustrating the results of the survey and displaying series of data that is derived from the primary collection of data. In this context, it assures descriptive form of analysis of the collected primary data. It also assures summarising the collected data and information with descriptive form of statistics. It interprets the results without any bias regarding the primary data collected from survey due to its quantitative nature**. A. Hager et al. [7],** close by the Coronavirus pandemic, another crisis has showed itself as mass anxiety and furor idiosyncrasies, filled by lacking and every now and again mixed up information. There is in this way a tremendous need to address and better fathom COVID-19's illuminating crisis and really take a look at general assessment, so reasonable illuminating and procedure decisions can be executed. In this assessment article, we recognize public inclination related with the pandemic using Coronavirus express Tweets and R quantifiable programming, close by its viewpoint examination groups. **N. D. Younis et al. [8],** show pieces of information into the headway of fear assessment long term as COVID-19 pushed toward high levels in the United States, using enlightening text based examination upheld by important literary information representations. Besides, we give a systemic outline of two fundamental AI (ML) grouping techniques, with regards to text based examination, and look at their viability in ordering Coronavirus Tweets of changing lengths. We notice areas of strength for an exactness of 91% for short Tweets, with the Naïve Bayes strategy. **K. H. Manguri et al. [9],** likewise see that the strategic relapse grouping technique furnishes a sensible exactness of 74% with more limited Tweets, and the two techniques showed moderately more fragile execution for longer Tweets. This exploration gives experiences into Coronavirus dread opinion movement, and layouts related strategies, suggestions, impediments and valuable open doors.

III. CNN

2D Convolution Neural Networks (CNN) [10-11] are a special type of Neural Networks. These work on two-dimensional input (example images) whereas Artificial Neural Network (ANN) [12] works on one-dimensional vectors. Each layer in CNN has a certain number of filters which convolves over input 2-D signal. In initial layers, these filters detect edges and corners and in deeper layers, they get trained to recognize more complex features pertaining to the objects in the image. In CNN forward propagation is guided by a convolution operation on input images using a kernel and explained with equation (1):

$$(I * K)[i, j] = \sum_{p=0}^{m-1} \sum_{q=0}^{n-1} I[i - p, j - q]K[p, q]$$

At the start of training of CNN, filters are initialized with random values and as the model gets trained by back-propagation algorithm using the errors in estimating the actual output, the weights of the filters are modified to identify specific patterns in inputs. The convolution layers are followed by max-pooling layers which reduce the size of the 2-D signal to retain only important features and to speed up computations. After couple of convolution and max pool layers, a flattened layer is included to transform the 2-D signal to 1-D signals similar to those used in ANN. The last layer of the model is a softmax layer which has nodes equal to the number of output classes. This layer indicates the probability for all the classes under consideration. The maximum probability node is assigned the value 1 and the rest are assigned 0 values. Thus the output is in form of a 1-D vector of 0's and 1's. The node with value 1 represents the class to which the input image belongs to.

IV. PROPSOED METHODOLOGY

It comprises of different layers of neuron, which computationally associated for insignificant handling. In CNN, Convolution layer is a fundamental part that assumes a fundamental part for outcome in picture handling undertakings like division and grouping. The motivation behind the convolutional layer is to distinguish remarkable neighborhood designs like lines, edges and any further visual components.

The boundaries utilized for explicit channel tasks are gone about as convolutions that learn while preparing the model. Portion depicts the numerical tasks that duplicate the nearby neighbors of a given pixel by a little cluster of learned boundaries; this part activity helps in extraction of visual highlights, for example, edges tones and so on.

It emphasizes an unsupervised form of research approach assuring no particular algorithm is maintained. It helps in moving towards generation of a new theoretical framework associated with the outcomes of the research. In this aspect, it is essential to note that it allows the researcher proper flexibility for bringing out a vast area of different kinds of analytic options. It is also useful for enabling the researcher to examining as well as analysing constructionist form of methodological position.

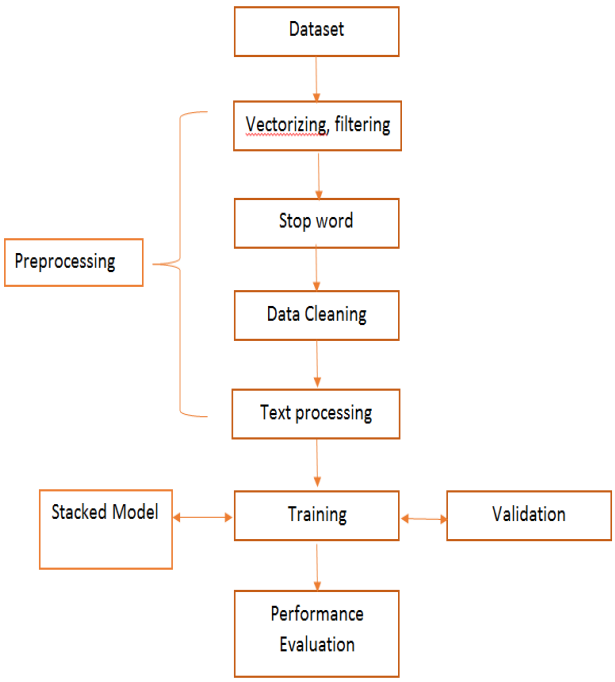

Fig. 1: Flow Chart of Proposed Methodology

This extraction cycle is performed by utilizing channels and each channel is a matrix molded component that can move over the given picture. The worth of the given picture and the moving network are added in view of the loads of the channel. The convolutional layer can apply many quantities of channels subsequently to create various element maps.

The convolutional layers are trailed by the pooling layer that limits the element map continuously and spatially.

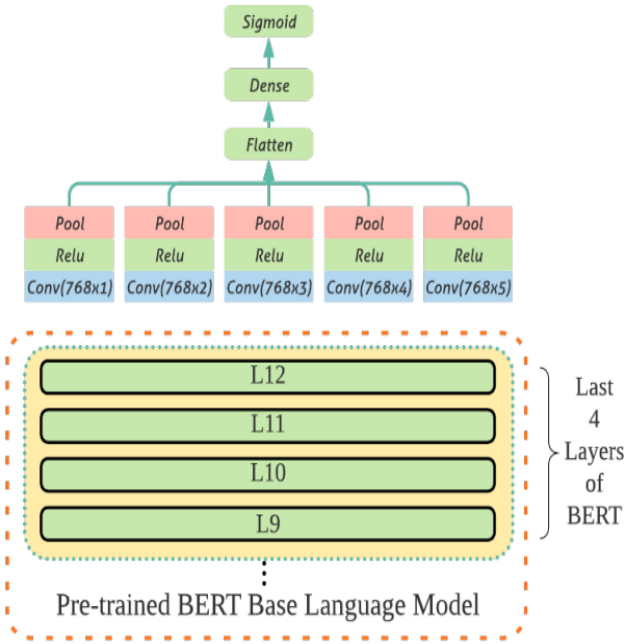

Fig. 2: BERT-CNN model structure

Hence the pooling layer is utilized for limiting the components of element maps proficiently and stays vigorous with the shape and position of the recognized semantic highlights of the given picture.

Generally max pooling capabilities are utilized for the pooling layer for include map. The convolutional layers and pooling layers are utilized over and over or on the other hand for a few times.

BERT:- BERT is cutting edge language model, which can be calibrated, or utilized straightforwardly as an element extractor for different text based undertakings. In our analyses, three pre-prepared language-explicit. Subsequent to setting the most extreme arrangement length of every message test.

It assures proper understanding along with suitable interpretations of qualitative data that is collected in this research. It also assures generation of best possible form of themes, which is supported through secondary data. The themes are further described and interpreted in the most adequate manner for assuring a proper understanding of the research consequences.

Like other researches, this thesis also complies with specific "*ethical considerations*" during its conduction process. Firstly, this study is conducted through adoption of proper honesty and integration. In this context, it also assures objectivity and maintains openness during the conduction of the study.

Step 1:
Collect the dataset, this dataset contains 50000 covid tweets.

Step 2:
Performing EDA and getting insights of the dataset.

Step 3: Processing

- Remove Url
- Remove Emoji
- Remove Duplicates
- Decontraction
- Sepereate alphanumeric
- Unique Character
- Lowering
- Remove Punctuation
- Remove Stop Words

Step 4:
Preparing is dataset for the training purpose by ONE HOT ENCODING and VECTORIZING.

Step 5:
Creating a Bert Based CNN Model and fitting the data to it, let it train. After completion, use the model for testing.

Step 6:
Evaluation of the model, testing the model on the test set and measuring is the performance in terms of precision, recall and F1-Score. The Deep Hybrid Model performed very well.

Step 7: Processing

- Remove Url
- Remove Emoji
- Remove Duplicates
- Decontraction
- Sepereate alphanumeric
- Unique Character
- Lowering
- Remove Punctuation
- Remove Stop Words

```
Model: "model_1"
_________________________________________________________________________________________________
Layer (type)                 Output Shape              Param #   Connected to
=================================================================================================
input_3 (InputLayer)         [(None, 128)]             0
_________________________________________________________________________________________________
input_4 (InputLayer)         [(None, 128)]             0
_________________________________________________________________________________________________
tf_bert_model (TFBertModel)  TFBaseModelOutputWit 109482240   input_3[0][0]
                                                              input_4[0][0]
_________________________________________________________________________________________________
dense_1 (Dense)              (None, 3)                 2307      tf_bert_model[0][1]
=================================================================================================
Total params: 109,484,547
Trainable params: 109,484,547
Non-trainable params: 0
_________________________________________________________________________________________________
```

Fig. 3: Model summary of Proposed Methodology

V. SIMULATION RESULT

The simulation result for tweets data is shown in below. The fig. 4 represent the date vs tweets count in coronavirus pandemic.

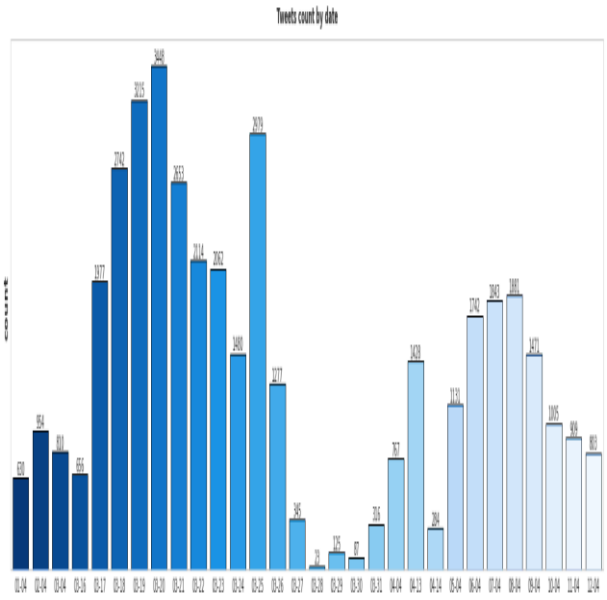

Fig. 4: Tweets Count by date

Fig. 5 represent by country vs tweets count. This fig. is shown in 25 country tweets data and clearly that London is the highest tweets and Mumbai is the lowest tweets.

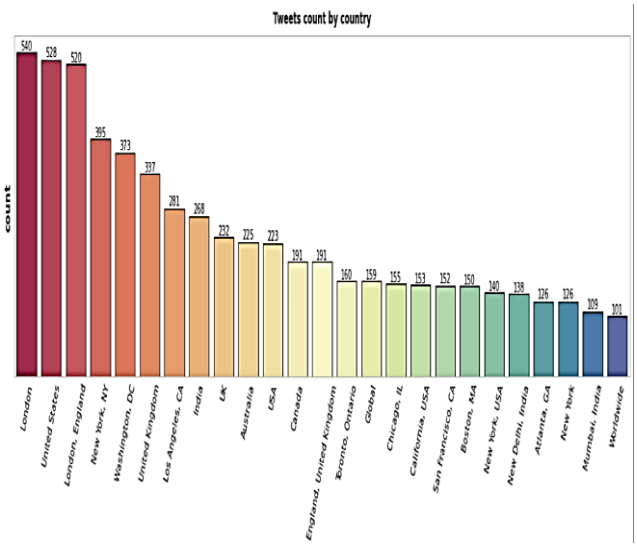

Fig. 5: Tweets count by country

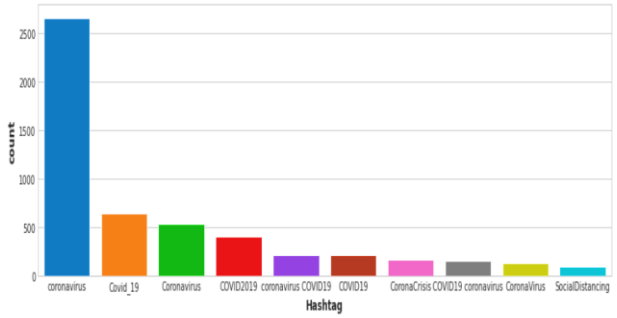

Fig. 6: Top Hastags

Different types of hashtag's are representing in fig.6. Coronavirus is the most popular hashtag's compared to other.

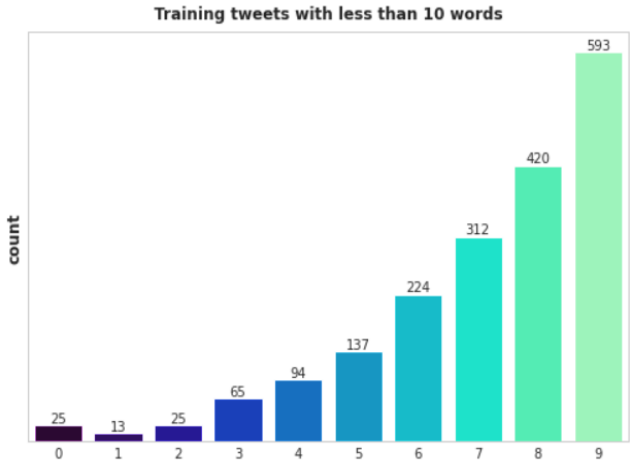

Fig. 7: Training tweets with less than 10 words

Training tweets with less than 10 words vs count is represent in fig. 7. Highest nine words are count and one words is lowest is shown in fig.

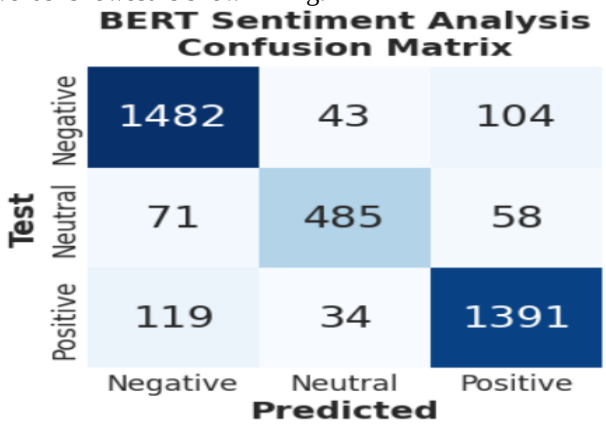

Fig. 8: Conclusion Matrix

Fig. 8 represents the confusion matrix of proposed methodology. The confusion matrix is divided into three part i.e. positive, negative and neutral. 1391 is the highest positive and 1482 is the negative tweets in coronavirus pandemic.

VI. CONCLUSION AND FUTURE WORK

Due to the on-going pandemic of COVID-19, the digital transformation has happened at a rapid pace and in this case, UCaaS is subjected to incorporation in most of the businesses across the world. Several companies caught flat-footed by having people who can work from anywhere. Additionally, it taught every industry a helpful lesson that medical emergencies or pandemic can happen again and people will not have any plan to get through this issue; however, possible solutions can help to fight against these types of issues. Because of COVID-19 pandemic, the majority of industries have significantly adopted UCaaS and that is why this particular study is going to understand growth of UCaaS by focusing on global conditions and a special reference to India. Due to COVID, adoption of UCaaS has been noticed significantly in Asia Pacific, especially in India.

REFERENCES

[1] Xiongwei Zhang, Hager Saleh, Eman M. G. Younis, Radhya Sahal and Abdelmgeid A. Ali, "Predicting Coronavirus Pandemic in Real-Time Using Machine Learning and Big Data Streaming System", Hindawi, 2020.

[2] J. Samuel, G. G. M. N. Nawaz Ali, M. M. Rahman, E. Esawi, and Y. Samuel, "Covid-19 public sentiment insights and machine learning for tweets classification," *Information*, vol. 11, no. 6, p. 314, 2020.

[3] R. Ali, A. Bharathi, and K. Saritha, "COVID-19 outbreak: tweet based analysis and visualization towards the influence of coronavirus in the world," *Gedrag en Organisatie*, vol. 33, no. 2, 2020.

[4] H. Wang, Z. Wang, Y. Dong et al., "Phase-adjusted estimation of the number of coronavirus disease 2019 cases in Wuhan, China," Cell Discovery, vol. 6, no. 1, pp. 1–8, 2020.

[5] G. Barkur and G. B. K. Vibha, "Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: evidence from India," Asian Journal of Psychiatry, vol. 51, Article ID 102089, 2020.

[6] M. Bhat, M. Qadri, Noor-ul-Asrar Beg, M. Kundroo, N. Ahanger, and B. Agarwale, "Sentiment analysis of social media response on the Covid19 outbreak," Brain, Behavior, and Immunity, vol. 87, pp. 136-137, 2020.

[7] A. Hager, E. M. G. Younis, A. Hendawi, and A. A. Ali, "Heart disease identification from patients' social posts, machine learning solution on Spark," Future Generation Computer Systems, vol. 111, pp. 714–722, 2020.

[8] N. D. Younis, N. Y. Hashim, Y. M. Mohialden, M. A. Mohammed, T. Sutikno, and A. H. Ali, "A real-time big data sentiment analysis for iraqi tweets using spark streaming," Bulletin of Electrical Engineering and Informatics, vol. 9, no. 4, pp. 1411–1419, 2020.

[9] K. H. Manguri, R. N. Ramadhan, and P. R. M. Amin, "Twitter sentiment analysis on worldwide COVID-19 outbreaks," Kurdistan Journal of Applied Research, vol. 5, no. 3, pp. 54–65, 2020.

[10] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, and C. U. Lehmann, "An "infodemic": leveraging high-volume twitter data to understand public sentiment for the COVID- 19 outbreak," Open Forum Infectious Diseases, vol. 7, no. 7, 2020.

[11] D. Hashim, "A spark-based big data analysis framework for real-time sentiment prediction on streaming data," *Software: Practice and Experience*, vol. 49, no. 9, pp. 1352–1364, 2019.

[12] N. Coletta, "An ensemble classification system for twitter sentiment analysis," Procedia Computer Science, vol. 132, pp. 937–946, 2018.

[13] S. Das, R. K. Behera, M. Kumar, and S. K. Rath, "Real-time sentiment analysis of twitter streaming data for stock prediction," *Procedia Computer Science*, vol. 132, pp. 956–964, 2018.

[14] L. R. Rath, S. D. Shetty, and S. Deepak Shetty, "Streaming big data analysis for real-time sentiment based targeted advertising," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 1, p. 402, 2017.

[15] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *Proceedings of the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, Springer, Vancouver, Canada, October 2017.

[16] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection," *IEEE Access*, vol. 6, pp. 33789–33795, 2018.

[17] B. Basheri, D. Wang, S. Cheng, and X. Xie, "Modeling and analysis for vertical handoff based on the decision tree in a heterogeneous vehicle network," *IEEE Access*, vol. 5, pp. 8812–8824, 2017.

[18] Z. Wang, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient k NN classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143–148, 2016.