iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.53088

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Design of Efficient Model to Predict Duplications in Questionnaire Forum using Machine Learning

Dr. G. R. Bamnote[1], Ms. Deepti Ingole[2]
Department of Computer Science & Engineering, PRM Institute of Technology & Research, Badnera

Abstract: ***Detection of duplicate sentences from a corpus containing a pair of sentences deals with identifying whether two sentences in the pair convey the same meaning or not. This detection of duplicates helps in deduplication, a process in which duplicates are removed. Traditional natural language processing techniques are less accurate in identifying similarity between sentences, such similar sentences can also be referred as paraphrases. Using Quora and Twitter paraphrase corpus, we explored various approaches including several machine learning algorithms to obtain a liable approach that can identify the duplicate sentences given a pair of sentences. This paper discusses the performance of six supervised machine learning algorithms in two different paraphrase corpus, and it focuses on analyzing how accurately the algorithms classify sentences present in the corpus as duplicates and non-duplicates.***

Keywords: Machine learning, Support Vector Machine, Logistic Regression, K-Nearest Neighbor

I. INTRODUCTION

Social media platforms are a great success as can be witnessed by the number of the active user base. In the age of internet and social media, there has been a plethora of social media platforms, for example, we have Facebook, for user interaction, LinkedIn, for professional networking, WhatsApp for chat and video calling, Stack Overflow for technical queries, Instagram for photo sharing. Along the line, Quora is a Question & Answer platform and builds around a community of users to share knowledge and express their, opinion and expertise on a variety of topics. Question Answering sites like Yahoo and Google Answers existed over a decade however they failed to keep up the content value [32] of their topics and answers due to a lot of junk information posted; thus their user base declined. On the other hand, Quora is an emerging site for the quality content, launched in 2009 and as of 2019, it is estimated to have 300 million active users1 . Quora has 400,000 unique topics2 and domain experts as its user so that the users get the first-hand information from the experts in the field. With the growing repository of the knowledge base, there is a need for Quora to preserve the trust of the users, maintain the content quality, by discarding the junk, duplicate and insincere information. Quora has successfully overcome this challenge by organizing the data effectively by using modern data science approach to eliminate question duplication

A. Research Problem

As for any Q&A, it has become imperative to organize the content in a specific way to appeal users to be an active participant by posting questions and share their knowledge in respective domain of expertise. In keeping the users ' interest, it is also essential that users do not post duplicate questions and thus multiple answers for a semantically similar question, this is avoided if semantically duplicate questions are merged then all the answers are made available under the same subject. Detecting semantically duplicate questions and finding the probability of matching also helps the Q&A platform to recommend questions to the user instead of posting a new one. Given our focus of study, we defined the following two research questions: RQ1: How can we detect duplicate questions on Quora using machine learning and deep learning methods? RQ2: How can we achieve the best possible prediction results on detecting semantically similar questions ? Research questions one and two have been studied on the first dataset released by Quora3 , however our aim is to achieve the higher accuracy on this task.

II. LITERATURE SURVEY

The previous work to detect duplicate question pairs using Deep learning approach [1], shows that deep learning approach achieved superior performance than traditional NLP approach. They used deep learning methods like convolutional neural network(CNN), long term short term memory networks (LSTMs), and a hybrid model of CNN and LSTM layers. Their best model is LSTM network that achieved accuracy of 81.07% and F1 score of 75.7%. They used GloVe word vector of 200 dimensions trained using 27 billion Twitter words in their experiments.

©IJRASET: All Rights are Reserved

The method proposed in [17] makes use of Siamese GRU neural network to encode each sentence and apply different distance measurements to the sentence vector output of the neural network. Their approach involves a few necessary steps. The first step was data processing, which involves tokenizing the sentences in the entire dataset using the Stanford Tokenizer4 . This step also involved changing each question to a fixed length for allowing batch computation using matrix operations. The second step involves sentence encoding, where they used both recurrent neural network(RNN) and gated recurrent unit (GRU). They initialized the word embedding to the 300-dimensional GloVe vectors [27]. The next step was determining the distance measure [21] that are used in combining the sentence vectors to determine if they are semantically equivalent. There were two approaches for this step, the first being calculating distances between the sentence vectors and running logistic regression to make the prediction. The paper has tested cosine distance, Euclidean distance, and weighted Manhattan distance. The problem here is that it is difficult to know the natural distance measure encoded by the neural network. To tackle this issue, they replaced the distance function with a neural network, leaving it up to this neural network to learn the correct distance function. They provided a row concatenated vector as input to the neural network and also experimented using one layer and two- layer in the neural network. The paper utilized data augmentation as an approach to reduce overfitting. They also did a hyperparameter search by tuning the size of the neural network hidden layer (to 250) and the standardized length of the input sentences (to 30 words) which led to better performance. In the literature [30], authors have used word ordering and word alignment using a long-short-term-memory(LSTM) recurrent neural network [10], and the decomposable attention model respectively and tried to combine them into the LSTM attention model to achieve their best accuracy of 81.4%. Their approach involved implementing various models proposed by various papers produced to determine sentence entailment on the SNLI dataset. Some of these models are Bag of words model, RNN with GRU and LSTM cell, LSTM with attention, Decomposable attention model. LSTM attention model performed well in classifying sentences with words tangentially related. However, in cases were words in the sentences have a different order; the decomposable attention model [26] achieves better performance. This paper [26] tried to combine the GRU/LSTM model with the decomposable attention model to gain from the advantage of both and come up with better models with better accuracy like LSTM with Word by Word Attention, and LSTM with Two Way Word by Word Attention. In the relevant literature [31], the authors have experimented with six traditional machine learning classifiers. They used a simple approach to extract six simple features such as word counts, common words, and term frequencies(TF-IDF) [28] on question pairs to train their models. The best accuracy reported in this work is 72.2% and 71.9% obtained from binary classifiers random forest and KNN, respectively. Finally, we reviewed the experiments by Quora's engineering team [20]. In production, they use the traditional machine learning approach using random forest with tens of manually extracted features. Three architectures presented in their work use LSTM in combination with attention, angle, and distances. The point noted from this literature is that Quora uses the word embedding from its Quora Corpus whereas all other selected baselines from the literature review used GloVe [27] pre-trained word to vectors from the glove project5 .

III. METHODOLOGY

To implement a model that classifies the question pairs and tweet pairs as duplicate or non-duplicate, we considered several machine learning approaches to find out the best suited approach that can classify the Quora question pairs and Twitter tweet pairs more accurately. We incorporated algorithms that follow a supervised learning approach in which we have a set of targets which has to be predicted by the model with the help of various variables.

1) *Machine Learning Algorithms:* This is a binary classification problem. We used supervised learning algorithms. Initially, we cleaned the strings by removing leading and trailing white spaces and then converted them to lower case. Feature extraction is carried after this from the data. These features include word count, character count, word share, and TF-IDF share. The algorithms we incorporated are Random Forest, Logistic Regression, Decision Trees, K-Nearest Neighbors, Naive Bayes, and Support Vector Machine. Except for the Naive Bayes model, we have normalized all feature values before performing the fitting of the model.
2) *Logistic Regression:* Logistic Regression is a classification methodology. It is used for predicting an outcome from independent variables. It predicts the probability for the outcome to fall into any of the available class. Logistic Regression predicts the probability of occurrence of an event by fitting data to a logistic function.
3) *Support Vector Machine:* An algorithm which outputs an optimal hyperplane to separate labeled training data. We implemented Grid Search cross-validation to find the best parameters (C, kernel), which are (1000, sigmoid). We set max iteration to 500 to avoid out of memory issue. SVM deals with finding an optimal hyperplane and reduction of the errors in classification based on it.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

4) *K-Nearest Neighbor:* It is an algorithm which uses the average of k-nearest data points to predict the testing data value. It predicts the label by finding the neighbor class which is nearest with the help of distance measures. The distance here is usually 2D Euclidean distance. We performed a Grid Search cross-validation to find best parameters (n neighbors, weights), which are (8, uniform).
5) *Random Forest:* Random Forest is performed using an ensemble of decision trees to reduce overfitting of one-tree model. The training algorithm for Random Forest applies the general technique of bagging, which means selecting a random sample of training data and fitting trees to these samples. Finally, we average the result of all trees to make prediction.

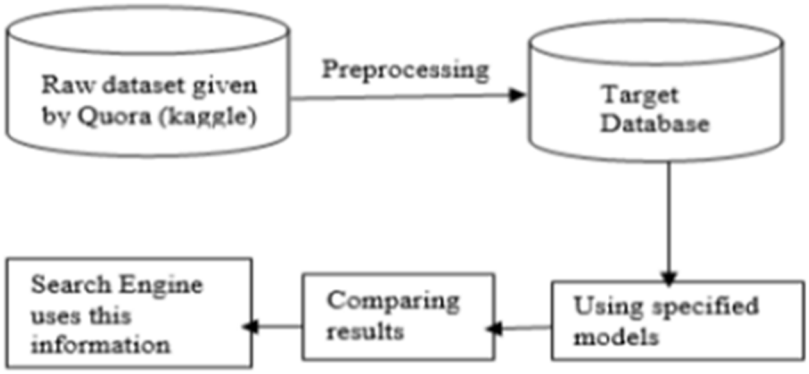


Fig 1:- Identification of duplicate Questionnaire using machine techniques

6) *Feature Visualization Feature:* Visualization deals with the representation of the features obtained from the experiments performed on the two available datasets. This helps in creating a better understanding of various unknown statistics about the dataset. The detection of duplicate sentences can be made easy with this proper visualization of the extracted features. In this section, we closely observe and represent the characteristics of the Twitter and Quora corpus. We visualize the features such as word frequency, number of characters in a sentence, number of words in a sentence, word share ratio among sentence pair, and TF-IDF share ratio among the sentences in a sentence pair for both the corpus.

A. Quora Paraphrase Corpus

A Quora corpus consists of question pairs in which each question will be related to certain domain which is posted by the users [6]. The specialty of Quora dataset is that each question should be semantically correct; it means that the questions asked in Quora should be grammatically correct unlike Twitter because in Quora, the users ask questions to get appropriate answers from other users; this can be made sure only if the question asked is understandable to others. The Feature visualization of the Quora corpus is interpreted in the following. The section deals with visualization of the features extracted from both the questions in question pair. The most frequently occurring words and phrases present in question1 and question2 of the corpus will be represented using bigger fonts in word cloud for both questions in the question pair. The word cloud generated includes important informations showing the presence of words such as India, make and phrases such as “best way” in common to both questions in a pair which helps in analyzing the similarity. The total characters present in the Quora question pairs and the probability of the occurrence of characters in both duplicate and non-duplicate question pairs are found; the number of characters present is in an average of 40 in the Quora corpus. The character count also indicates that there are no questions having more than 200 characters. The word count that represents the total number of words in the questions is diagrammatically visualized. It finds the average word count per question for both duplicate and non-duplicate pairs. The average word count in a question is 8 and the maximum number of words in a question is 40. The word share ratio represents the ratio of the number of the words shared in common between two questions among the duplicate and non-duplicate question pairs. The word share ratio was diagrammatically visualized and we observed that the word share ratio increases the chances for a sentence pairs to be a paraphrase or duplicate increases. In TF-IDF, Term Frequency represents the number of times a token occurs in a document which makes it important to know how frequent a word is repeated in a document to understand the importance of that particular word in document. Inverse Document Frequency finds out the number of documents in which a word is occurring. Therefore, the TF-IDF share ratio finds the importance of words in the question pairs. The ratio is represented in the above graph. From Fig. 2, we can observe that, as the TF-IDF share ratio increases, the chances for a sentence pairs to be a paraphrase or duplicate increase.

IV. SYSTEM REQUIREMENT

1) *Hardware:* The system requires a computer with sufficient processing power to run the ML algorithm for object detection and identification. The exact hardware requirements will depend on the size of the dataset and the complexity of the model, but a high-end CPU and GPU are generally recommended for best performance.
2) *Software:* The system requires Python and the necessary machine learning libraries (e.g. PyTorch, OpenCV) to be installed on the computer. The system also requires Gradio to be installed in order to create and deploy the user interface.
3) *GPU:* To train your ML model, you likely required a powerful GPU with sufficient memory to handle the large dataset and complex model architecture.
4) *CPU:* For running the Gradio web app, you would require a CPU that can handle multiple requests and run the inference on the trained model in real-time.
5) *Memory:* Sufficient memory would be required for storing the dataset, pre-trained models, and any intermediate data generated during training.
6) *Disk Space:* Sufficient disk space would be required for storing the dataset, trained models, and any other related data.

V. RESULTS

5 Result A total of six machine learning techniques were carried out on the Twitter and Quora corpus, and we have compared the performance of both the datasets on these algorithms separately. The results of classification for various models depicting the machine learning algorithms are shown in Tables 3 and 4. The output statistics we obtained shows that while performing duplicate detection on both considered datasets, all six algorithms didn't result in equal performance. In Twitter corpus, Logistic Regression yielded better accuracy than all other techniques and this Logistic Regression method had achieved higher accuracy's compared to the existing approaches in this dataset. Logistic Regression is an algorithm which is highly suitable for correlated data. The dataset constitute classes that are linearly separable which makes them suitable to work efficiently with Logistic Regression. One of the major advantages of Logistic Regression is that it can be regularized with respect to data in order to avoid overfitting. The specialty of Twitter dataset is that the tweets may not convey a proper semantic meaning. Therefore, this dataset also includes sentences which don't have a proper grammatical meaning. The Quora dataset yielded better result on Random Forest algorithm. The Quora dataset contains sentences with a proper semantic meaning. The advantages of Random Forest is that it can handle the missing values and since we have more trees in Random Forest the classifier will not overfit the model. SVM had yielded comparatively less results with respect to other algorithms in both datasets. Choosing a better kernel function for SVM was a difficult part and another disadvantage of SVM we incurred was that it took long training time on large datasets like Quora. Naive Bayes algorithm yielded a better result compared to SVM but it was the second worst performer among the six algorithms. KNN had a better performance equally in Quora and Twitter dataset which shows that distance measure-based mechanisms like KNN are also efficient in identifying the similarity between texts.

Table 1:- Results for Quora dataset analysis

Parameters\Models	Logistic Regression	SVM	KNN	Random Forest
Precision	0.79	0.66	0.78	0.78
Recall	0.79	0.70	0.78	0.78
F1score	0.78	0.66	0.71	0.71
Accuracy	78.6	66	78.0	78.1

VI. CONCLUSION

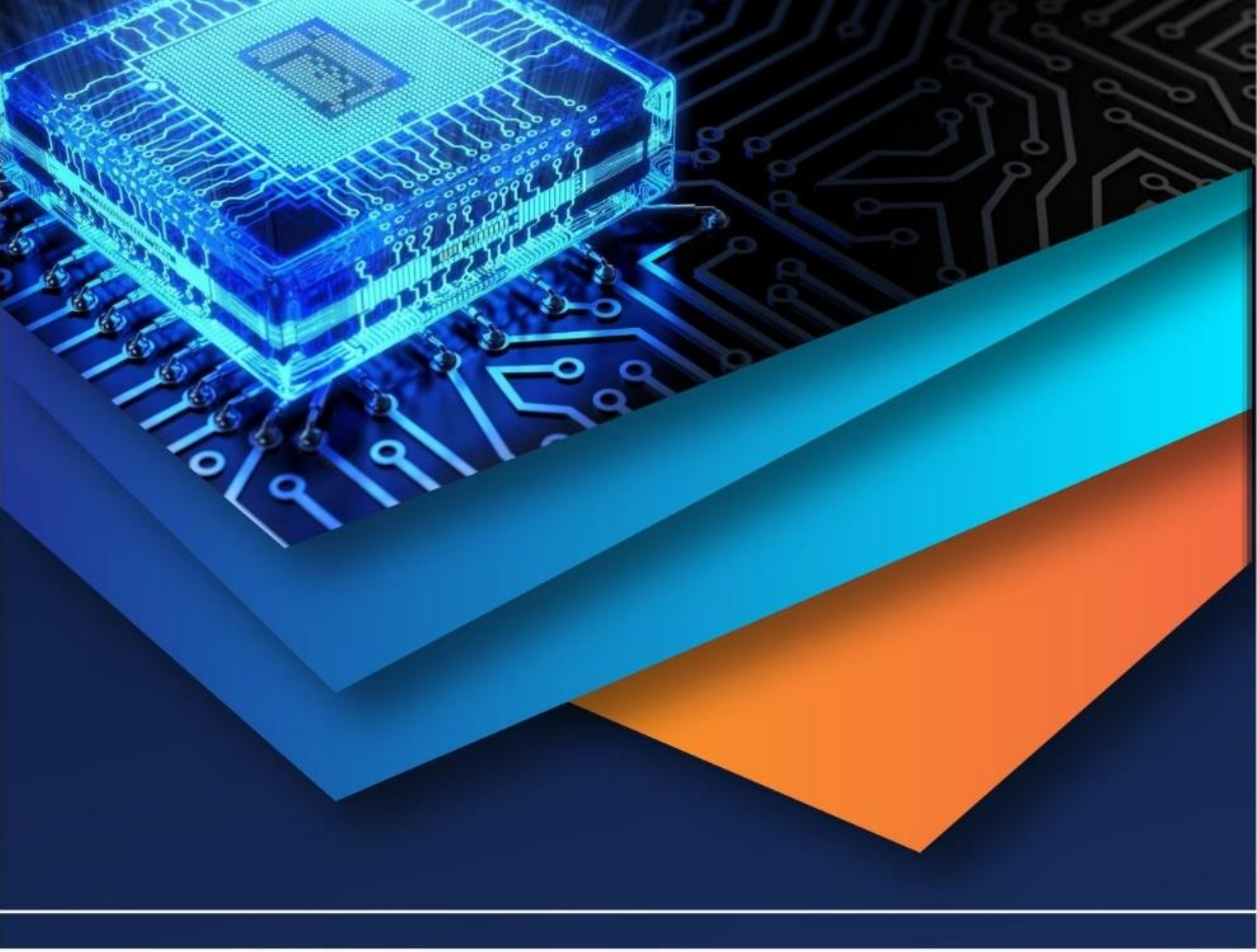
The duplicate detection carried out on the Quora and Twitter corpus suggests that machine learning algorithms work well in detecting the duplicate sentences among a sentence pair. The algorithms considered for this work included Logistic Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbor, Naive Bayes, and Random Forest. Among these considered algorithms no algorithm failed for paraphrase detection in both corpora. The Random Forest and K-Nearest Neighbor algorithms performed equally well in both datasets. The Logistic Regression performed best among all algorithms for Twitter corpus; similarly, Random Forest provided best result for Quora corpus.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

REFERENCES

[1] Travis Addair. 2017. Duplicate question pair detection with deep learning. Stanf. Univ. J (2017).
[2] Steven Bird, Ewan Klein, and Edward Loper. [n.d.]. Natural language processing with Python: analyzing text with the natural language toolkit.
[3] T Chen and C Guestrin. 2016. XGBoost : Reliable Large-scale Tree Boosting System. arXiv.
[4] François Chollet et al. 2015. Keras.
[5] Y M Chou, Y M Chan, J H Lee, C Y Chiu, and C S Chen. 2018. Unifying and merging well-trained deep neural networks for inference stage. IJCAI Int. Jt. Conf. Artif. Intell (2018), 2049–2056.
[6] E Dadashov, S Sakshuwong, and K Yu. 2017. Quora Question Duplication. , 9 pages.
[7] Raihana Ferdous et al. 2009. An efficient k-means algorithm integrated with Jaccard distance measure for document clustering. In 2009 First Asian Himalayas International Conference on Internet. 1–6.
[8] Y Freund and R E Schapire. 1996. Experiments with a New Boosting Algorithm. Proc. 13th Int. Conf. Mach. Learn (1996).
[9] J H Friedman. 1999. Greedy Function Approximation : A Gradient Boosting Machine 1 Function estimation 2 Numerical optimization in function space.
[10] F Gers. 2001. Long short-term memory in recurrent neural networks. Neural Comput (2001).
[11] P Geurts, D Ernst, and L Wehenkel. 2006. Extremely randomized trees. Mach. Learn 63, 1 (2006), 3–42.
[12] I J Goodfellow, D Warde-Farley, M Mirza, A Courville, and Y Bengio. 2013. Maxout Networks.
[13] G Guo, H Wang, D Bell, Y Bi, and K Greer. 2010. KNN Model-Based Approach in Classification.
[14] K He, X Zhang, S Ren, and J Sun. 2015. Delving deep into rectifiers. Proc. IEEE Int. Conf. Comput. Vis (2015).
[15] G Hinton. 2014. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. , 1929-1958 pages.
[16] T K Ho. 1995. Random decision forests. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR (1995).
[17] Y Homma, S Sy, and C Yeh. 2016. Detecting Duplicate Questions with Deep Learning. 30th Conf. Neural Inf. Process. Syst. (NIPS 2016), no. Nips (2016), 1–8.
[18] G Huang, Z Liu, L Van Der Maaten, and K Q Weinberger. 2017. Densely connected convolutional networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition (2017).
[19] S Ioffe and C Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.
[20] N Jiang, Lili, Chang, and Dandekar Shuo. 2019. , 4-4 pages.
[21] J O Josephsen. 1956. Similarity Measures for Text Document Clustering. Nord. Med 56, 37 (1956), 1335–1339.
[22] B Karlik and A Vehbi. 2011. Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks. Int. J. Artif. Intell. Expert Syst 1 (2011), 111–122.
[23] M J Kusner, Y Sun, I K Nicholas, and Q W Kilian. 2015. From Word Embeddings To Document Distances Matt, Vol. 63130. St. Louis, 1 Brookings Dr., St. Louis, MO.
[24] Kanti V Mardia. 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika 57, 3 (1970), 519–530.
[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. GoogleNewsvectors-negative300.bin.gz - Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).
[26] A Parikh, O Tckstrm, D Das, and J Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. Proc. 2016 Conf. Empir. Methods Nat. Lang. Process (2016), 2249–2255.
[27] J Pennington, R Socher, and C Manning. 2014. Glove: Global Vectors for Word Representation. Proc. 2014 Conf. Empir. Methods Nat. Lang. Process (2014), 1532– 1543.
[28] S Robertson. 2004. Understanding inverse document frequency: On theoretical arguments for IDF. J. Doc (2004).
[29] Philip H Swain and Hans Hauska. 1977. The decision tree classifier: Design and potential. IEEE Transactions on Geoscience Electronics 15, 3 (1977), 142–147.
[30] A Tung and E Xu. 2017. Determining Entailment of Questions in the Quora Dataset. , 8 pages.
[31] S Viswanathan, N Damodaran, and A Simon. 2019. Advances in Big Data and Cloud Computing, Vol. 750. Springer.
[32] G Wang, K Gill, M Mohanlal, H Zheng, and B Y Zhao. 2013. Wisdom in the social crowd: An analysis of Quora. WWW 2013 - Proc. 22nd Int. Conf. World Wide Web (2013), 1341–1351.
[33] Jiannan Wang, Guoliang Li, and Jianhua Fe. 2011. Fast-join: An efficient method for fuzzy token matching based string similarity join. In 2011 IEEE 27th International Conference on Data Engineering. 458–469.
[34] David I Warton, Stephen T Wright, and Yi Wang. 2012. Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution 3, 1 (2012), 89–101.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)