iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.44750

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 10 Issue VII July 2022- Available at www.ijraset.com

An Enhanced Question Pair Similarity Using Machine Learning Approaches

Dr. Sunil Bhutada[1], Dr. K. Kranthi Kumar[2], Pranavnath Pendota[3], Hemanth Pendam[4], Chaturya Katragadda[5]
[1, 2, 3, 4, 5] *Sreenidhi Institute of Science and Technology*

Abstract: ***Quora is a place to gain and share knowledge—about anything. It's a platform to ask questions and connect with people who contribute unique insights and quality answers. This empowers people to learn from each other and to better understand the world.***
Currently, Quora uses a Random Forest model to identify duplicate questions. Tackling this natural language processing problem by applying advanced techniques to classify whether question pairs are duplicates or not, so will make it easier to find high quality answers to questions resulting in an improved experience for Quora writers, seekers, and readers.
We by enhancing the features level by level in each system of total 3 got the XG Boost algorithm as the best model in order to solve such problem, not only in the case of Quora but also with the Stack overflow, medium etc. [1]
Keywords: Question Pair Similarity, Random Forest Algorithm, Extreme Gradient Boosting, Logistic Regression, Naïve bayes, Natural Language Processing, Natural Language Toolkit, Parts of Speech Tagging.

I. INTRODUCTION

Quora is popularly known for the question and answers forum where everyone can write the questions and anyone can respond to those questions. preprocessing that we have done here are if there are any duplicates present in the records as well as if there are any null values that are present in the text then those records get removed in order to make the text consistent and unique. In the system 1, we are taking features as Bag of Words only, but not any other mathematically formula generated features.[2]
In the system 2, we took 7 extra basic features along with the Bag of Words features which are used in the system 1. Finally in the system 3, we added 15 more advanced features which plays a key role in the incremental of accuracy of the models compare to the other two previously designed models. Models used are Random Forest, XG Boost, Logistic Regression, Naïve Bayes[2]. We have observed that the Random Forest, XG Boost are having accuracies in an increasing order from 70% to 79% reach. Where as in the Logistic Regression, the accuracy got increased from the system 1 to system 2, but I got decreased from the system 2 to system 3, which is supposed to be increased in that case because of the goodness of features which is proved. In the Naïve Bayes, there is a slight increase after the system 1, but in the case of system 2 and in the system 3 the accuracy neither increased nor decreased.[3]
XG Boost algorithm got an accuracy of 79% which was the highest compared to all. Random forest also performed well but not as best as XG Boost. Logistic Regression and the Naïve Bayes algorithms are concluded that they are not as par as with the level of XG Boost and the Random Forest. [3]

II. LITERATURE SURVEY

Many research papers have been published by many people, explaining different types of features, techniques that are being used in the process of finding duplicates, and syncing the same answer for the both.
The authors clarify that the XG Boost model yields accuracy high efficiency. Sultana R. et al. discussed overcrowding in health care units caused by population rise over the last 10 years. Data mining models namely boosting and decision trees, originating from the tree-based approach were used for prediction activities. The latest method of gradient-boosting machines was elucidated for enhancing productivity and multiply speediness.[4]
Classifying duplicate questions can be a tricky task since the variability of language makes it difficult to know the actual meaning of a sentence with certainty. This task is similar to the paraphrase identification problem, which is a thoroughly researched Natural Language Processing (NLP) task. [5]
Feature engineering has been the center of focus for most of the traditional methods developed by different practitioners. The common features used are bag of words (BOW), term frequency and inverse document frequency (TF IDF), unigrams and bigrams. Support Vector Machine (SVM), used with different feature extraction techniques such as BOW or n-gram vectors, is one of the main methods in text categorization. [6] – [8]

Recently, deep learning approaches have achieved very high performance across several Natural Language Processing (NLP) tasks especially in Semantic Text Similarity.[9]

Duplicate question detection is a binary classification problem on various length strings. The challenging part of the problem is to represent sentences as numerical inputs such that the learning algorithms can work on it. A widely used method involves hand engineered feature generation. This method, combined with tree-based models such as random forests, is common in industry. This is the current approach that Quora takes (Dandekar, 2017) and this method can be used together with bag-of-word based models to enhance the performance. [10]

In this work, Daoud classifies and arranges queries in the database concerning their logical structure considering their form and range of application. This approach is specific to a particular language i.e. Arabic. Chen, M et al. give an overview of a prediction method of transient stability for power systems based on XG Boost. The authors discuss that more features can be taken into account in the XG Boost-based method, which is preferred in complex power systems, particularly in the power system penetrated by new energy sources, compared to the traditional-method. [11]

III. PROPOSED SYSTEM

Main goal of our system is to find the different accuracies of 4 models mentioned below in 3 different environments (or systems) and get the best model out from them. [3]

Initially we go with Bag of Words features only in order to see how the model is performing without the external features which are wrote manually. Next Implementing the same models using the externally added features those are 7 basic features. Finally in the third system, we added 15 advanced features along with these 7 features which are basic along with the Bag of Features. Advanced and Basics are the terms that are just used to describe the intensity of the features on how they are impacting the model performances [1]. Below diagram Fig.1 represents the brief overview of the proposed system. We've divided the system into 3 parts: [7]

1) *Proposed System 1* Only Bag of Words features are used in the feature Engineering part.
2) *Proposed System 2* Along with the Bag of Words features, we use 7 basic features.
3) *Proposed System 3* Along with the Bag of Words features and 7 basic features, we also use 15 advanced features, and we do more preprocessing here. [12]

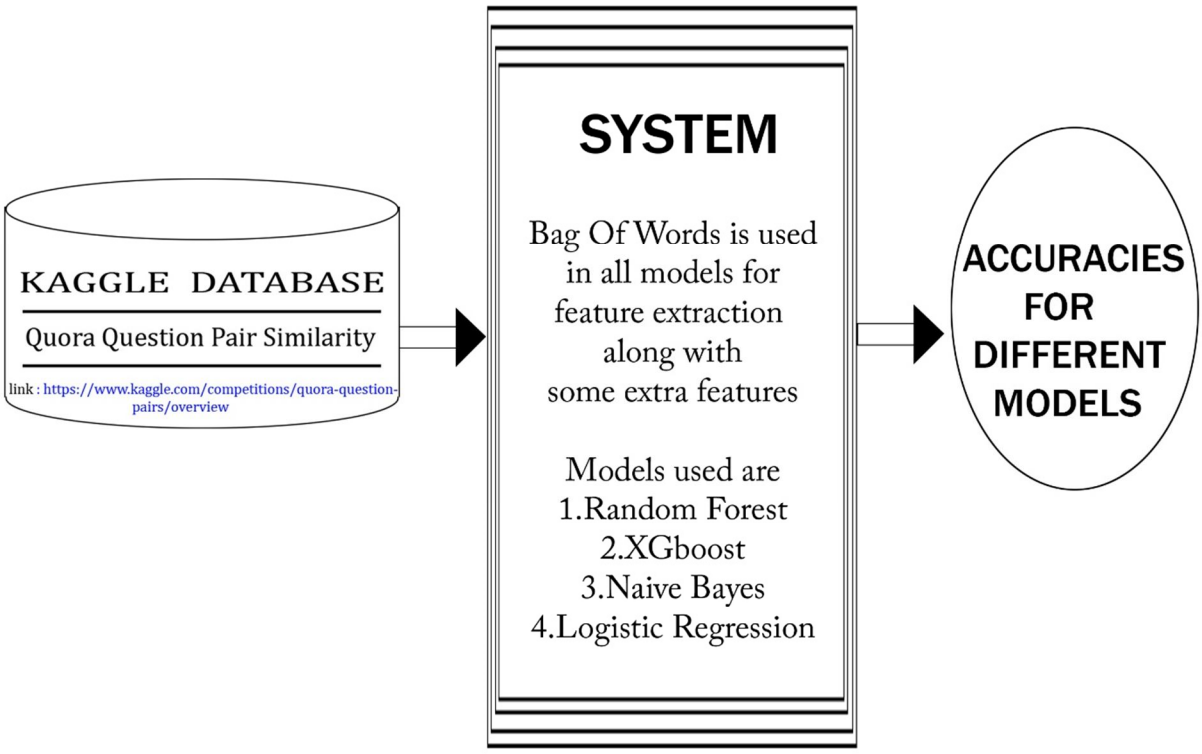


Fig. 1 Overview of the complete system

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

IV. ARCHITECHTURE

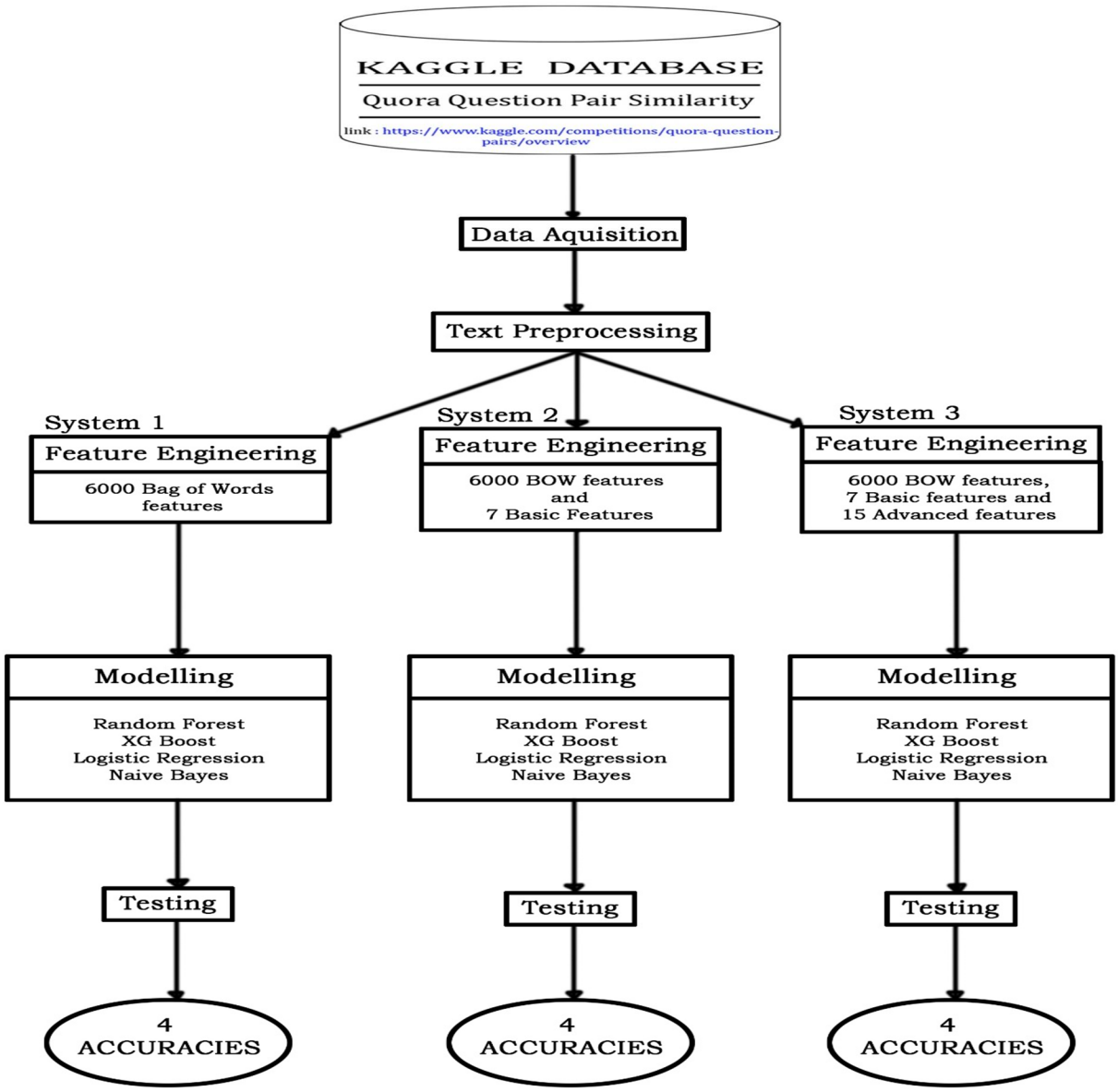


Fig. 2 Architecture of the systems

A. Feature Engineering

features which are used in the Quora question pair similarity problem are in the order of increasing. That is, In the system 1, we are taking features as Bag of Words only, but not any other mathematically formula generated features. Coming to the system 2, we took 7 extra basic features along with the Bag of Words features which are used in the system 1, they are q1_len, q2_len, q1_num_words, q2_num_words, word _common, word _total, word _share.[2]. Finally in the system 3,as shown in fig.2 we added 15 more advanced features which plays a key role in the incremental of accuracy of the models compare to the other two previously designed models, which are cwc_ min, cwc_ max, csc _min, csc _max, ctc _min, ctc _max , last _word _eq, first _word _eq, abs _len _diff, mean _len, longest _substr _ratio, fuzz _ratio, fuzz _partial _ratio, token _sort _ratio, token _set _ratio.[3] The reason behind the line that accuracy got increased by adding such features is the property of that particular feature. Features that we've used incrementally are having the PDF curves with less overlapping and can be written in the form of IF and ELSE.[11]

B. Experimental Setup

1) *Random Forest:* Random Forest comes under the Bagging process, which comes under the Ensemble Learning technique. Given Training Dataset, Random Forest algorithm performs the Row sampling with replacement and feature sampling with replacement. After the generation of each sample, that particular Dataset is given to the weak learners (models). So, if we take the forest proposed system as an example, then we had 6000 features, using this 6000features all the decision trees that are in the Random Forest get created and by using the Majority voting the resultant predicted output get generated from the model. The model that we're created using the training data as well as the basic features and the advanced features, in all the three cases the model is performing with a good accuracy of average 70%. [10]
2) *XG Boost:* XG Boost is a boosting technique, where boosting is also a type of Ensemble technique[1].As we know, ensemble techniques are having the weak learners. Here, the weak learners are Decision trees same as in the case of random Forest, once the Residue is calculated then that loss is used to construct the next decision tree. So, the output that we are getting with an accuracy which is greater than the Random Forest in such a case where we've added the external features such as basic and advanced features.[8]
3) *Logistic Regression:* Our problem comes under the binary classification problem i.e., the problem of classifying two questions whether those are the duplicates of each other or not with the number of 1's and 0's. This model uses Sigmoid Function, where it makes the output get generated between 0 and 1. The Optimizer used is Gradient Descent, which is used to reduce the Loss function and increase the cost function, in order to make the model more accurate in predicting the output of output problem. Here, as we are taking 6000 dimensions, 6000-D plane i.e., a hyper plane, which separated the whole 24000 Datapoints into two categories of classes 0's and 1's.[8]
4) *Naïve Bayes:* Naïve Bayes classifier is not just a single classifier, it represents the classifiers that are Naïve and which uses the Bayes theorem in its algorithms. So, one of the types of Naïve Bayes classifier that we are using here us Gaussian Naïve Bayes classifier, which initially assumes all the inputs are in a Gaussian way of looking i.e., in the normal distribution. If we assume Decision tree like structure as a weak learner, then the working of probabilities in the bayes theorem which is used is as follows: let say there is a feature in the root node, let assume that we know it is already true in that Datapoint, then what is the probability that the child feature of that root node is true is what we are solving using the Bayes theorem.[7]

V. IMPORTANT SECTIONS

A. Data Acquisition

Kaggle has a Quora question Pair Similarity dataset, which contains the two questions and their respective duplicative nature, in each row. Likewise, we are having 4 lakh records in it as a .csv file. So, the dataset that we had used has been taken from the Kaggle. There are a total of 7 columns in the dataset [3].

- Id this is the row id
- Qid1 this is the id for question1
- Qid2 this is the id for question2
- Question1 this column contains the actual question1
- Question2 this column contains the actual question1
- Is_ duplicate this column tells whether the questions in both question1 and question 2 are duplicates or not.

B. Text Preprocessing

In Text preprocessing, there are few important things that has to be done to the text before entering into the feature engineering part in the process of Natural Language Problem [1]. So, preprocessing that we have done here are if there are any duplicates present in the records as well as if there are any null values that are present in the text then those records get removed in order to make the text consistent and unique. Text Preprocessing that we've used here are [2]

Replace certain special characters with string equivalents,
De-contracting words,
Removing HTML tags,
Remove punctuations.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

VI. TESTING

A. *Accuracy Scores*

From fig.3, We have observed that the Random Forest[2], XG Boost are having accuracies in an increasing order from 70% to 79% reach. Where as in the Logistic Regression, the accuracy got increased from the system 1 to system 2, but I got decreased from the system 2 to system 3, which is supposed to be increased in that case because of the goodness of features which is proved.[3]. In the Naïve Bayes, there is a slight increase after the system 1, but in the case of system 2 and in the system 3 the accuracy neither increased nor decreased.[5]

		Accuracies for Different Features		
		Onlt BOW	BOW + 8 basic features	BOW + basic + advanced features
MODELS	Random Forest	75.2	76.6	78.6
	XGBoost	72.8	76.4	**79.2**
	Logistic Regression	70.3	73.8	71.5
	Naive Bayes	55.03	59.6	59.6

Fig. 3 Comparison of all models Accuracies

B. *Prediction VS Actual 'y' values*

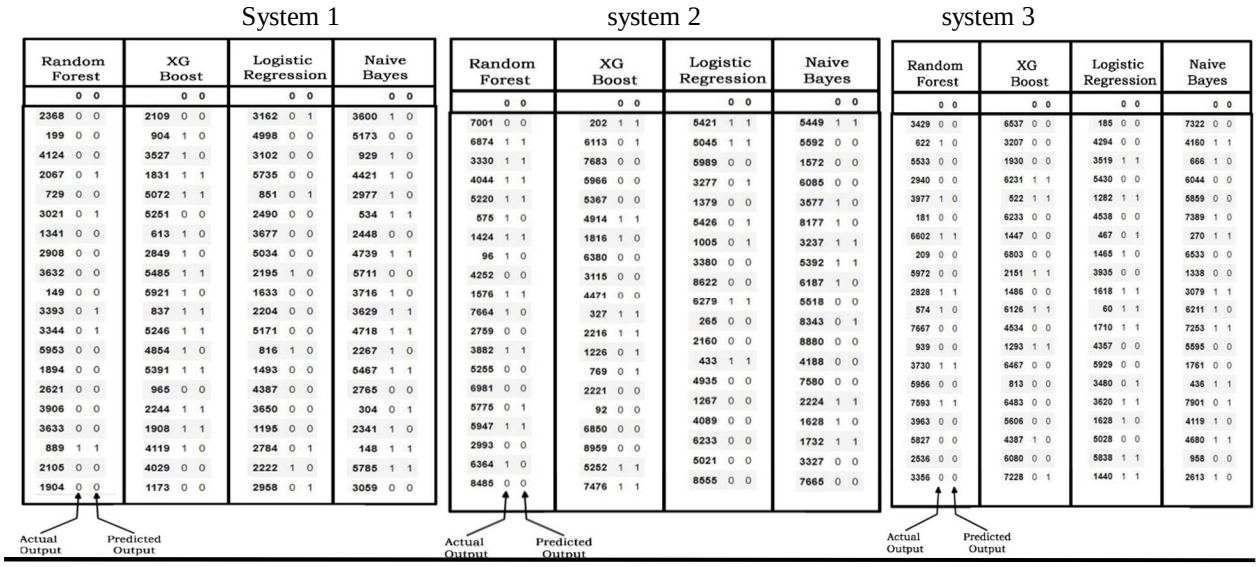


Fig.4 Actual And Predicted Y-Values For All Systems

In the Proposed System 3, after testing is done on all the models, we get the predicted outputs as y _pred. From the testing dataset we are having the actual values i.e., y _actual. As shown in the above figure Fig.4 we have collated the actual and predicted y-values of all systems [3]

Here we are taking the samples of both the y _actual and y _pred with a sample size of 20, In order to see how the model is predicting the values, if we compare the predicted values with the actual values then we got the above, is the way to describe the accuracy of that particular model.[4]

C. Visualization of Model Accuracies

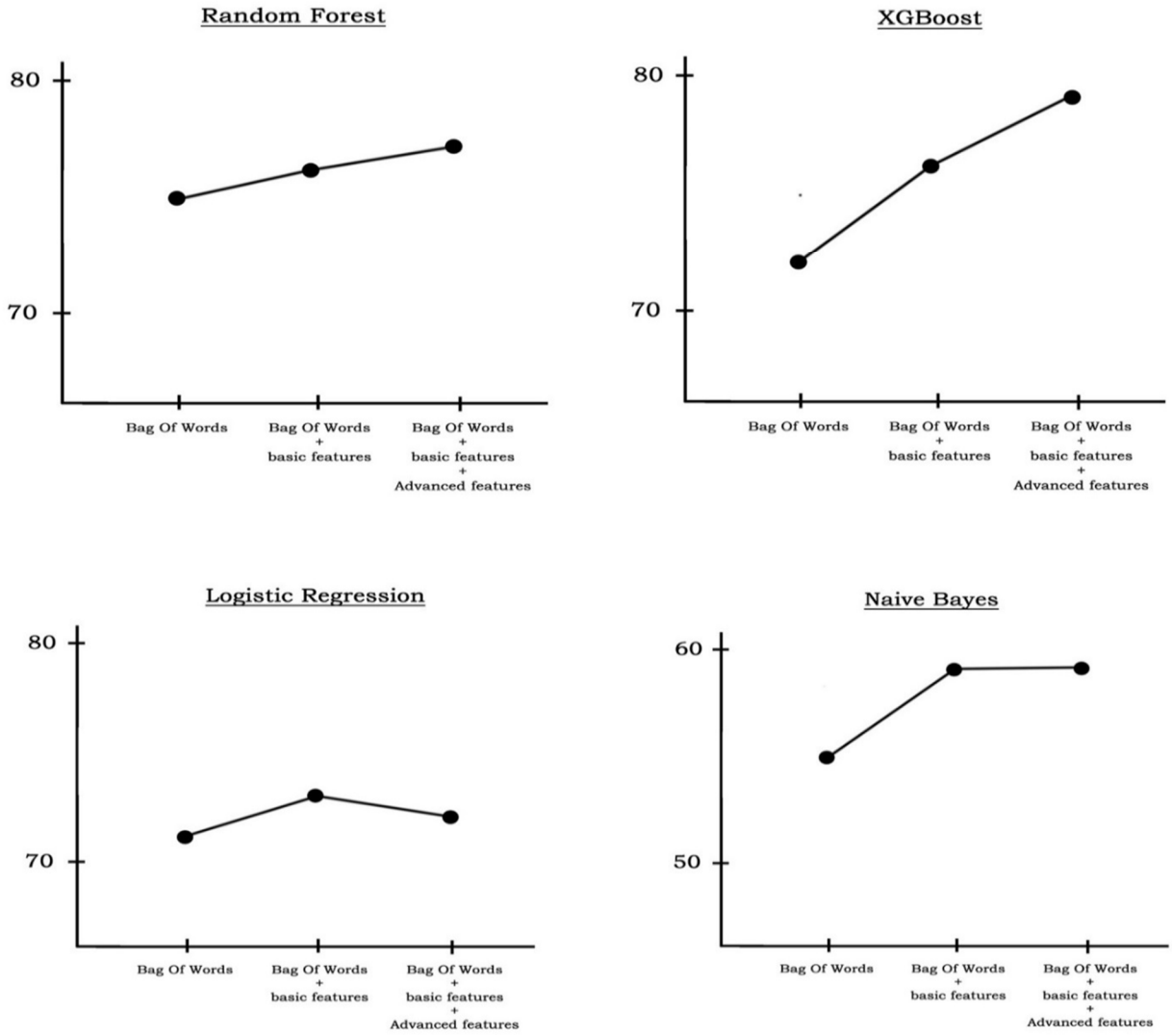


Fig. 5 Visualization of all model accuracies

From the above Fig.5, it is clear that the graph of the XG Boost was strictly increasing, even the graph for the Random Forest was, but the slope of it was less compared to the XG Boost, which results in concluding the XG Boost[12] as the best model. Coming to the Linear Regression[11], it is decreasing from the certain point, there is no guarantee that it will decrease in the next step, but at the same time there is no guarantee that it will increase. In the Naïve Bayes[10], the graph was monotonically increasing, that is it is continuously showing the same accuracy, even though there is change in the system, so its definitely not comparable to the XG Boost which finally treated as the best model according the accuracies[4].

VII. CONCLUSION

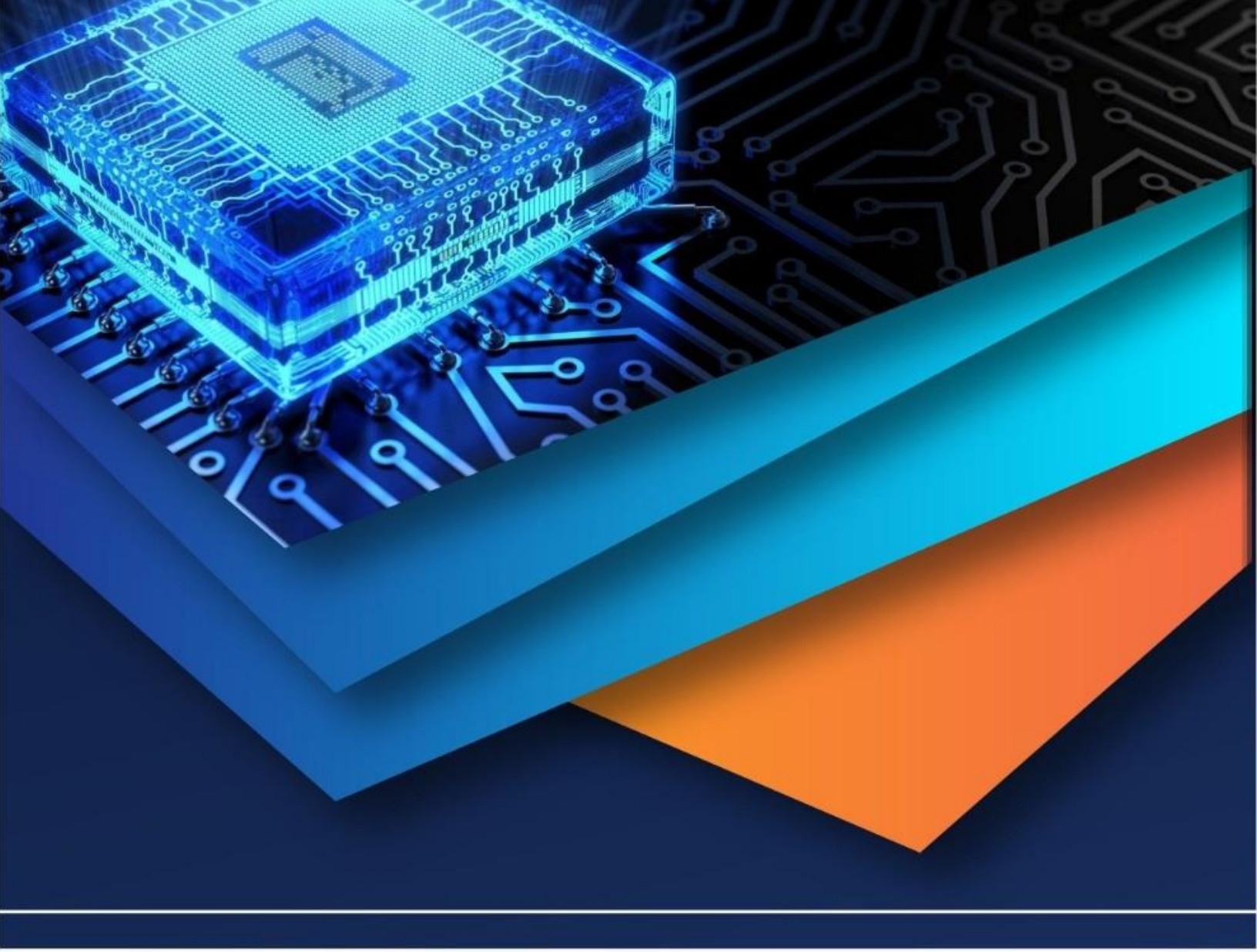
The case study that we have did on the natural language processing-based project gave the conclusions on many things regarding the models and the text preprocessing the many others. XG Boost is the best model that we had got by the means of accuracy, and reason for it is the iterative nature of the XG Boost [7], which is not present in any of the models that we've done. Random Boost also works well, but slightly less compared to the XG Boost model as the Random Forest Decision trees are not recorrected as in the case of XG Boost with help of Residue that has being calculated recurrently in that algorithm. Coming to the feature engineering, we have discovered some advanced features which played a key role in increasing the accuracy of the models.[11]

VIII. FUTURE SCOPE

In the applications like Quora, Stack overflow, medium, etc., which are all the questions and answer platforms that are being used for the many years and made them up to a level where there are today is at its peak. But, still they are developing. Quora one such is having one problem of question pair similarity that we've solved using various models by taking various cases in order to get the best model out of it. This case study will help in the future if someone will work on the problem question pair similarity, in the form of which features to use?[4] how to build models on NLP based problems? how to get the best model? [5]

REFERENCES

[1] W. Zhu, T. Yao, J. Ni, B. Wei, and Z. Lu, ''Dependency-based Siamese long short-term memory network for learning sentence representations,'' PLoS ONE, vol. 13, no. 3, Mar. 2018, Art. no. e0193919.
[2] B. N. Patro, V. K. Kurmi, S. Kumar, and V. P. Namboodiri, ''Learning semantic sentence embeddings using sequential pair-wise discriminator,'' CoRR, vol. abs/1806.00807, pp. 1–15, Mar. 2018.
[3] J. Mueller and A. Thyagarajan, ''Siamese recurrent architectures for learning sentence similarity,'' in Proc. 13th AAAI Conf. Artif. Intell., 2016, pp. 2786–2792.
[4] M. Tsubaki, K. Duh, M. Shimbo, and Y. Matsumoto, ''Non-linear similarity learning for compositionality,'' in Proc. 13th AAAI Conf. Artif. Intell., 2016, pp. 2828–2834.
[5] B. Rychalska, K. Pakulska, K. Chodorowska, W. Walczak, and P. Andruszkiewicz, ''Samsung Poland NLP team at SemEval-2016 task 1: Necessity for diversity; combining recursive autoencoders, WordNet and ensemble methods to measure semantic similarity,'' in Proc. 10th Int. Workshop Semantic Eval. (SemEval), San Diego, CA, USA, 2016, pp. 602–608.
[6] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, ''Distributed representations of words and phrasesand their compositionality,'' CoRR, vol. abs/1310.4546, pp. 1–9, Oct. 2013.
[7] P. Sravanthi and B. Srinivasu, ''Semantic similarity between sentences,'' Int. Res. J. Eng. Technol., vol. 4, no. 1, pp. 156–161, 2017.
[8] R. Johnson and T. Zhang, ''Supervised and semi-supervised text categorization using LSTM for region embeddings,'' in Proc. ICML, 2016, pp. 1–9.
[9] D. Tang, B. Qin, and T. Liu, ''Document modeling with gated recurrent neural network for sentiment classification,'' in Proc. Conf. Empirical Methods Natural Lang. Process., Lisbon, Portugal, Sep. 2015, pp. 1422–1432.
[10] C. Zhou, C. Sun, Z. Liu, and F. Lau, ''A C-LSTM neural network for text classification,'' Nov. 2015, arXiv:1511.08630. [Online]. Available: https://arxiv.org/abs/1511.08630
[11] K. S. Tai, R. Socher, and C. D. Manning, ''Improved semantic representations from tree-structured long short-term memory networks,'' in Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process., Beijing, China, vol. 1, Jul. 2015, pp. 1556–1566.
[12] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, ''Skip-thought vectors,'' CoRR, vol. abs/1506.06726, pp. 1–11, Jun. 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY