iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 10 **Issue:** VI **Month of publication:** June 2022

**DOI:** https://doi.org/10.22214/ijraset.2022.45072

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 VI June 2022- Available at www.ijraset.com*

# Detecting Duplicate Questions in Online Forums Using Machine Learning Techniques

Kratika Sharma[1], Satya Kiranmai Tadepalli[2]
*Assistant Professor, Department of Information Technology, Chaitanya Bharathi Institute of Technology (A), Hyderabad, Telangana, 500075*

***Abstract:*** ***Online forums like Quora are systems for collecting, sharing information, and discussing between users on a selected subject matter. Users in on-line forums can ask questions about a subject, then other users who're experts on that question would solution the query. However, due to the fact customers can ask questions in diverse methods, every now and then they ask questions that other users have previously requested. Consequently, a version is needed to detect the semantic similarity of questions in online boards. On this study, we're the usage of machine learning algorithms to discover the semantic similarity of questions. To seize the semantic similarity among questions, we're the usage of word embedding. This word embedding vector is used as an enter for neural network, and then the output is compared with other machine learning algorithms.***
***Index Terms: Online forums, machine learning algorithms, Natural language processing.***

## I. INTRODUCTION

Quora, stack overflow and reddit are Question-and-answer website where questions are asked, answered and edited by internet users either factually or in the form of opinions. Duplicate questions on this site are not uncommon, particularly as the number of questions asked grows. This poses an issue because, if treated independently, duplicate questions may prevent a user from seeing a high quality response that already exists and responders are unlikely to answer the same question twice. Identifying duplicate questions addresses these issues. It reduces the answering burden for responders and makes it possible to direct users to the best responses, improving the overall user experience. We aimed to present a comprehensive set of machine learning models, and to study their performance on the dataset. We used simple linear models as our baseline. We built and tested Support Vector Machines (SVM), Naïve Bayes, Decision tree, logistic regression, Random Forests; 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 and this method can be used together with bag-of-word based models to enhance the performance. The dataset will be in csv format (csv stands for comma separated values) which is CSV is a standard for storing tabular data in text format. Analyzing the outputs produced by algorithms. Graphs generated based on the statistics of the algorithms used as graphical representation is the most efficient representation to show the statistics of a dataset graphical representations using python only. Identifying duplicate questions addresses these issues. It reduces the answering burden for responders and makes it possible to direct users to the best responses, improving the overall user experience doing which will make it easier to find high quality answers to questions resulting in an improved experience for Quora, stack overflow, writers, seekers and readers. The dataset will be in csv format (csv stands for comma separated values) which is CSV is a standard for storing tabular data in text format. Analyzing the outputs produced by algorithms. Graphs generated based on the statistics of the algorithms used as graphical representation is the most efficient representation to show the statistics of a dataset graphical representations using python only.

## II. RELATED WORK

Detecting semantically equivalent sentences or questions has been a long-standing problem in natural language processing and understanding. As Dey et al. [5] demonstrate traditional machine learning algorithms such as Support Vector Machines (SVMs) using hand-picked features and extensively preprocessed data perform well on the SemEval-2015 dataset. They argue that the performance of deep learning methods is heavily limited by the small, noisy datasets that they are trained on. Bogdanova et al. found that pairing a convolutional neural network (CNN) with a cosine-similarity distance measure was more effective than traditional methods of using Jaccard similarity or SVMs in identifying duplicate questions in a Stack Exchange dataset. Sanborn and Skryzalin [7] compared the use of recurrent neural networks (RNNs) and recursive neural networks with traditional machine learning methods and found that recurrent neural networks performed the best on the SemEval-2015 dataset.

Nonetheless, deep learning techniques have made considerable progress in recent years. Most deep learning methods for detecting semantic equivalence rely on a "Siamese" neural network architecture [3] that takes to two input sentences and encodes them individually using the same neural network. The resulting two output vectors are then compared using some distance metric. This approach is used successfully by both Bogdanova et al. [1] and Sanborn-Skryzalin [7].

To date, the only published results on the Quora dataset come from Wang et al. [8]. Observing that the encoding procedure in Siamese networks does not provide any interaction between the two input sequences, they instead propose a bilateral multi-perspective matching LSTM model. Their "matching aggregation" approach performs better than the Siamese CNNs and LSTMs that they tested.

## III. RESEARCH METHODOLOGY

To achieve this fundamental like to check whether the pair of questions are similar or not using the algorithms we would divide this problem into three parts: Taking a dataset consisting of questions in paired format and pre-processing them for various operations performed by algorithms[16].

The algorithms which would be used are Logistic Regression, decision tree, random forest, naive bayes algorithm, support vector machine. The dataset will be in csv format (csv stands for comma separated values),which is CSV is a standard for storing tabular data in text format. Analyzing the outputs produced by algorithms. To see whether which algorithm and/or its feature gives best accuracy and output in terms of algorithmic loss. Graphs generated based on the statistics of the algorithms used as graphical representation is the most efficient representation to show the statistics of a dataset graphical representations using python only.

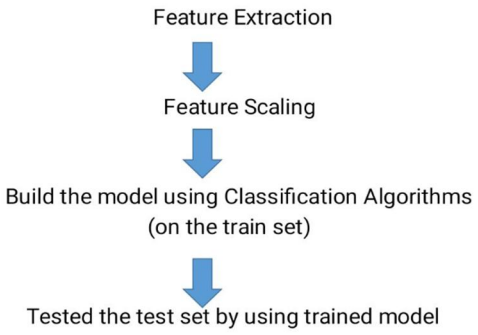



Fig. 1 System Data flow

To get the best decision model with highest accuracy, we tuned many parameters. First we tried to get the best parameters from both bag of words and n-gram techniques which give the best recognition rate on four dataset. The classifier has given about 100% accuracy in classifying the fake news texts

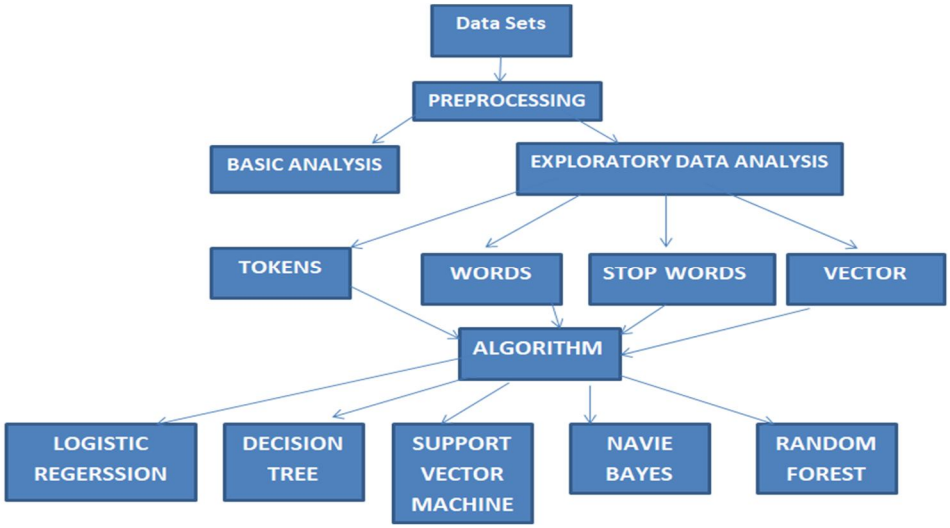



Fig. 2 System Architecture

## IV. RESULTS AND DISCUSSION

We illustrate experimental results of different approaches on the dataset in this part. And the average values of measures are presented. We have obtained the result as the following fig.

| SNo | Algorithm | Accuracy Score |
|---|---|---|
| 1 | Logistic Regression | 0.72480 |
| 2 | Decision Tree | 0.71585 |
| 3 | Random Forest | 0.75255 |
| 4 | Naïve Bayes | 0.73440 |
| 5 | SVM | 0.76015 |

Fig. 3 Final Result

As a result, we have 63.08% of non-duplicate pairs and 36.92% duplicate pairs.

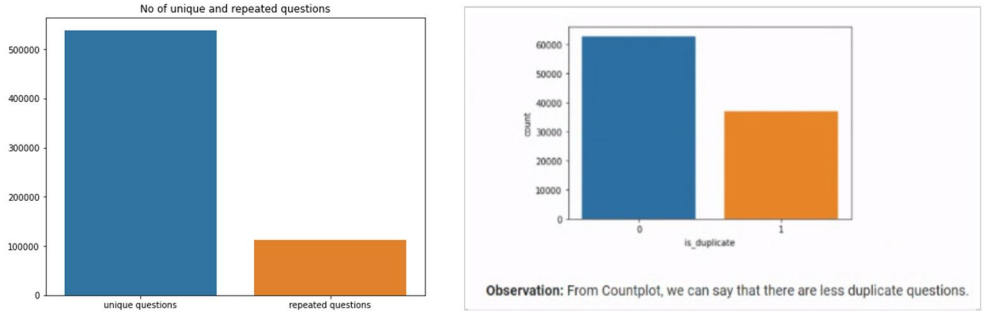



Fig. 4,5. Graphs of repeated questions


Fig. 6 The word cloud Question-1


Fig. 7 The word cloud Question-2

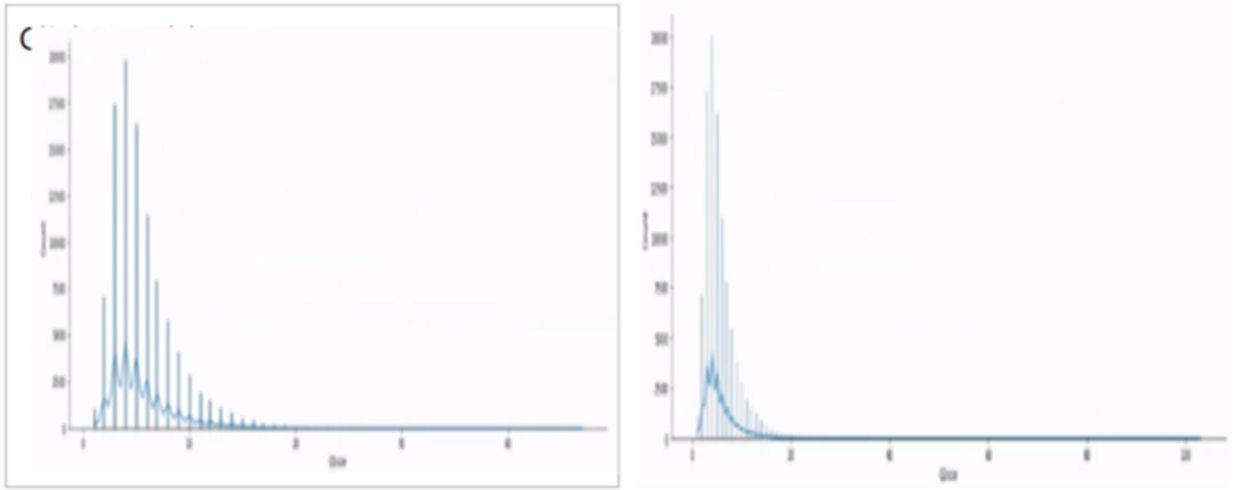


Fig. 8 Feature Comparison of size

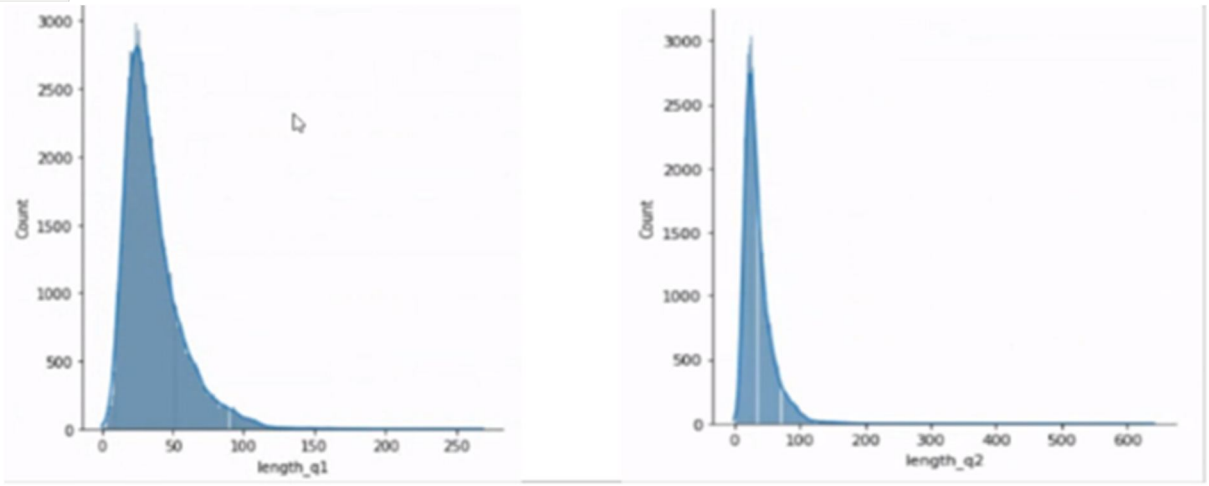


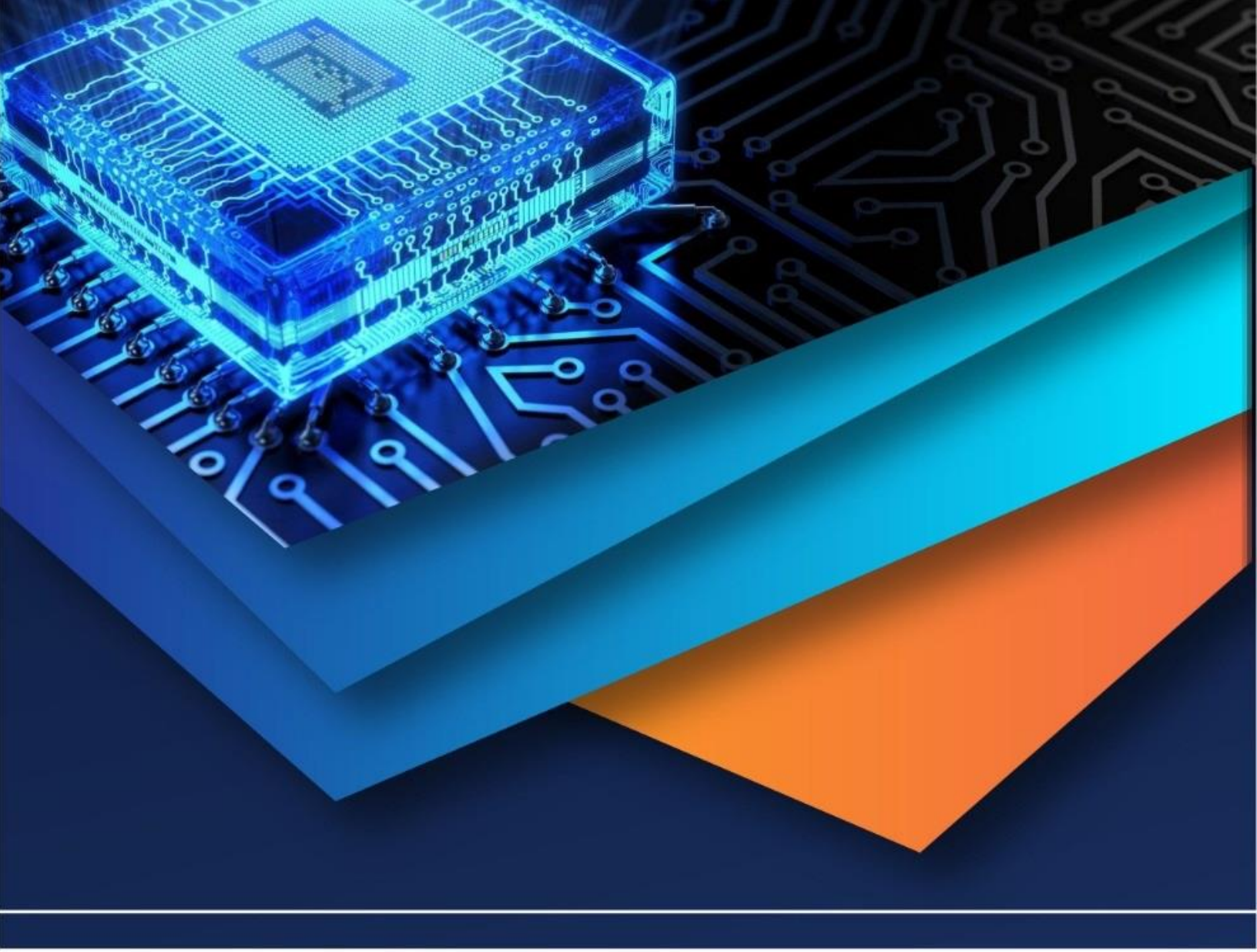

Fig. 9 Feature Comparison of length

## V. CONCLUSION AND FUTURE SCOPE

We tested a large number and variety of machine learning models to solve the duplicate question problem posed by the Quora dataset. Our best performing model was that of Support Vector Classifier then Random Forest, Naive Bayes, Logistic Regression and Decision tree respectively. We believe the Quora dataset is a useful resource to further explore the task of Natural Language understanding with ML techniques.

## REFERENCES

[1] Dasha Bogdanova, Cicero dos Santos, Luciano Barbosa, and Bianca Zadrozny. Detecting semantically equivalent questions in online user forums. Proceedings of the 19th Conference on Computational Natural Language Learning, 1:123–131, 2015.
[2] Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Chrisopher D. Manning, and Christopher P. Potts. A fast united model for parsing and understanding. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics., 1, 2016.
[3] Jane Bromley, James W. Bentz, Leon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Sackinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. International Journal of Pattern Recognition and Artificial Intelligence, 7:669–688, 1993.
[4] Toma´s Brychcın and Luk ˇ a´s Svoboda. Uwb at semeval-2016 task 1: Semantic textual similarity ˇ using lexical, syntactic, and semantic information. Proceedings of SemEval, pages 588–594, 2016.
[5] Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. A paraphrase and semantic similarity detection system for user generated short-text content on microblogs. In COLING, volume 16, pages 2880–2890, 2016.
[6] Richard Socher Jeffrey Pennington and Christopher D. Manning. Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pages 1532–1543, 2014.
[7] Adrian Sanborn and Jacek Skryzalin. Deep learning for semantic similarity. 2015.
[8] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral multi-perspective matching for natural language sentences. 2017.
[9] W. E. Zhang, Q. Z. Sheng, J. H. Lau, and E. Abebe, "Detecting duplicate posts in programming qa communities via latent semantics and association rules," in 26th International Conference on World Wide Web (WWW), Geneva, Switzerland, 2017, pp. 1221–1229.
[10] B. Xu, D. Ye, Z. Xing, X. Xia, G. Chen, and S. Li, "Predicting semantically linkable knowledge in developer online forums via convolutional neural network," in 31st International Conference on Automated Software Engineering (ASE), 2016, pp. 51–62.
[11] Y. Zhang, D. Lo, X. Xia, and J.-L. Sun, "Multi-factor duplicate question detection in Stack Overflow," Journal of Computer Science and Technology, vol. 30, no. 5, pp. 981–997, Sep 2015.
[12] M. Ahasanuzzaman, M. Asaduzzaman, C. K. Roy, and K. A. Schneider, "Mining duplicate questions in Stack Overflow," in 13th International Conference on Mining Software Repositories (MSR), 2016, pp. 402–412.
[13] Y. Mizobuchi and K. Takayama, "Two improvements to detect duplicates in Stack Overflow," in 24th International Conference on Software Analysis, Evolution and Reengineering (SANER), 2017, pp. 563–564.
[14] Sharma, Kratika & Tadepalli, Kiranmai. (2021). PREDICTION OF CARDIOVASCULAR DISEASES USING GENETIC ALGORITHM AND DEEP LEARNING TECHNIQUES. INTERNATIONAL JOURNAL OF EMERGING TRENDS IN ENGINEERING AND DEVELOPMENT. 3. 10.26808/rs.ed.i11v3.01.
[15] Sharma, Kratika & Goyal, Ajay. (2013). Very high resolution image registration based on two step Harris-Laplace detector and SIFT descriptor. 2013 4th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2013. 1-5. 10.1109/ICCCNT.2013.6726632.
[16] Sharma, Kratika & Goyal, Ajay. (2013). Classification based survey of image registration methods. 2013 4th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2013. 1-7. 10.1109/ICCCNT.2013.6726741.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY