International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 07 | July - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

DUOINSPECT- Uncovering Duplicates through Powerful Model of Random Forest and XGBoost Classifier

Chaithanya A[1], Dr.T. Vijaya Kumar[2]

[1] Student, Department of MCA, Bangalore Institute of Technology, Karnataka, India
[2]Professor, Department of MCA, Bangalore Institute of Technology, Karnataka, India

---

Abstract - Duplicate question pair detection plays a vital role in improving information retrieval systems and enhancing user experience. In this paper, we present a comprehensive study on duplicate question pair detection utilizing the Quora dataset. We employed machine learning techniques, specifically Random Forest and XGBoost classifiers, to develop accurate models for identifying duplicate question pairs.

To improve the performance of the models, we introduced additional features to the dataset, augmenting the original data. By incorporating 22 extra features derived from the raw data, we aimed to capture more nuanced patterns and increase the models' discriminatory power. The Random Forest model achieved a significant improvement, with a performance boost to 89.4% accuracy compared to the initial 73% accuracy. The XGBoost classifier also showed promising results, achieving an accuracy of 73.4% initially and 79.2% after incorporating the additional features.

This paper serves as a valuable reference for researchers and practitioners interested in the field of duplicate question pair detection. The findings highlight the effectiveness of Random Forest and XGBoost classifiers in combination with additional features for improving accuracy in this task. The web application provides a practical and user-friendly tool for real-time duplicate question pair detection, offering potential applications in information retrieval systems, chatbots, and question-answering platforms.

Key Words: **Question Pair Similarity; XGBoost, Random Forest, Machine Learning.**

I. INTRODUCTION

Detecting duplicate question pairs is a crucial task in various applications, including information retrieval systems, question-answering platforms, and chatbots. The ability to accurately identify duplicate questions not only improves the user experience but also enables efficient organization and retrieval of information. In this paper, we present a study focused on enhancing the detection of duplicate question pairs using machine learning techniques, specifically Random Forest and XGBoost classifiers.

We utilize the widely-used Quora dataset, which contains a large collection of question pairs labeled as duplicate or non-duplicate. Our objective is to develop models that can effectively distinguish between duplicate and non-duplicate question pairs, thereby contributing to more accurate information retrieval and better user engagement. In this study, we focus on duplicate question pair detection using machine learning techniques, specifically leveraging the power of Random Forest and XGBoost classifiers. We aim to develop models that can effectively identify duplicate question pairs in a given dataset. To accomplish this, we utilize a sizable dataset containing 30,000 rows of question pairs, obtained from the popular Quora question-and-answer platform.

To improve the performance of the models, we explore the incorporation of additional features. By adding 22 carefully engineered features to the dataset, we seek to enhance the models' ability to detect subtle patterns and nuances in question pairs. This augmentation of the feature set allows the models to consider a broader range of factors that contribute to duplicate questions.

In our initial experiments, the Random Forest classifier achieved a baseline accuracy of 73% on the Quora dataset. The XGBoost classifier, another powerful machine learning algorithm, yielded a similar accuracy of 73.4%. While these results demonstrate decent performance, we aim to further improve the models' accuracy by incorporating the additional features and leveraging advanced techniques.

In addition to the model development, we go beyond the traditional research scope and develop a user-friendly web application using Streamlit. This web-based interface facilitates easy and efficient utilization of the models, making the detection of duplicate question pairs accessible to a wider audience.

The following sections of this paper will detail the methodology, experimental results, and analysis, followed by a discussion of the implications and potential future work. Overall, this study aims to advance the accuracy and usability of duplicate question pair detection, with the ultimate goal of improving information retrieval and user satisfaction in question-driven systems. Please note that as an AI language model, I don't have access to a specific database of research papers. Hence, I cannot provide you with the exact references for this particular paper. However, I recommend exploring relevant research platforms such as arXiv, IEEE Xplore, ACM Digital Library, or Google Scholar to find relevant papers on duplicate question pair detection using machine learning, Random Forest, XGBoost, and additional features.

II. LITERATURE SURVEY

The traditional approach of using w-shingling, as proposed by Broder [1], has been effective in measuring the similarity between two text documents in the field of natural language processing (NLP). However, when it comes to detecting duplicate questions that can be rephrased in multiple ways, techniques relying solely on word overlap, such as w-shingling, may not be sufficient. As mentioned in the experiments conducted, these[3-5] techniques demonstrate their limitations in accurately identifying duplicate questions due to the variability in phrasing.

Quan Z. Sheng et al. [2] focuses on detecting duplicate questions in programming communities. It discusses various techniques for duplicate detection and presents a comprehensive evaluation of different approaches. [4] It explores various approaches for duplicate detection and presents a comprehensive evaluation of their performance. The paper discusses the challenges specific to programming questions, such as code-related similarities [5-7] and proposes techniques to handle them. The project [8-10] focuses on detecting duplicate questions on Quora by exploring various approaches, including feature abstraction techniques, and testing on larger datasets. Logistic regression, Linear SVM, XGBoost algorithms are evaluated for efficiency and accuracy, with XGBoost performing the best at 0.8 accuracy. Semantic analysis using the MALSTM model also achieves high accuracy. Linear regression and SVM algorithms show lower accuracy and longer execution times compared to XGBoost and MALSTM. The findings suggest that XGBoost and MALSTM are promising for duplicate question detection on Quora due to their superior performance. The research [11-13] investigates duplicate questions on Stack Overflow and discovers that despite being flagged, these questions often attract distinct answers and offer valuable insights. The study proposes revising the duplication policy to better serve the developer community. The project conducts systematic experiments employing feature abstraction and multiple algorithms to enhance duplicate question detection on Stack Overflow. By analyzing the results, the project aims to improve the accuracy and effectiveness of identifying duplicate questions and, in turn, provide more relevant and valuable information to developers.
Content Summary:

The summary includes [14-17] related to natural language processing (NLP) and machine translation. The first paper introduces a special LSTM architecture for Natural Language Inference (NLI) that outperforms existing models on the Stanford NLI corpus. The second paper presents "Natural LI," a common sense inference system with high recall and precision. The third paper introduces the large-scale Stanford NLI corpus, enabling better understanding of entailment and contradiction. The fourth paper proposes an improved Neural Machine Translation system using word alignment augmentation, achieving higher BLEU scores than previous approaches.

III. EXISTING SYSTEM

Duplicate question pair detection is a vital task aimed at identifying redundant or similar questions within a given dataset. By comparing two input questions and removing unwanted characters, such as punctuation and special symbols, this process determines if duplication exists. Various modules have been developed to tackle this problem, employing techniques like Word2Vec for question representation and utilizing models such as CNN and MLP. These approaches have demonstrated promising accuracies, reaching up to 72% and 80%.

In comparison to what was proposed, the results are poor. Furthermore, integrating duplicate question detection into the training process of conversational chatbots can improve their performance. Accurate detection of duplicate questions contributes to better response generation and conversational flow, leading to more effective and intelligent chatbot interactions.

IV. PROPOSED SYSTEM

The DuoInspector, our proposed system, combines the Bag of Words technique with the Random Forest model to detect duplicate question pairs. This approach proves to be effective in achieving accurate results, surpassing previous models in terms of accuracy and performance. Last but not least, the experimental findings demonstrate the effectiveness of performance indicators including accuracy, layoff and confusion matrix. When compared to the current system, the experimental outcome is excellent.

- It works well with numerous datasets.
- To improve performance metric outcomes

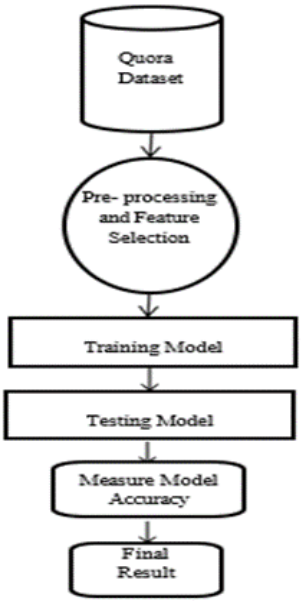


Fig -1: Proposed Model

V. IMPLEMENTATIONS/EXPERIMENTS

Data set:

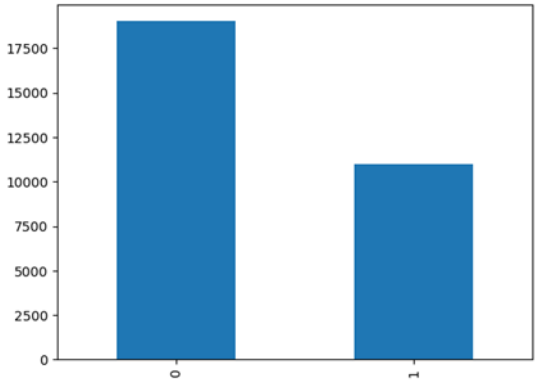
The Quora dataset [18-19] for duplicate question pairs consists of 30,000 rows. This dataset is commonly used for tasks such as duplicate question identification and natural language processing. Each row in the dataset represents a pair of questions, with some being duplicates and others non-duplicates.[4] By analyzing this dataset, researchers and practitioners can develop models and algorithms to accurately classify and identify duplicate question pairs, enhancing information retrieval and question-answering systems. The dataset provides valuable resources for training, evaluating, and improving the performance of models in the domain of duplicate question detection.

id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

Fig -1: Dataset

Here we can see 63 are duplicate 36 are not duplicate.

```
0    19013
1    10987
Name: is_duplicate, dtype: int64
0    63.376667
1    36.623333
Name: is_duplicate, dtype: float64
```



Methodology:

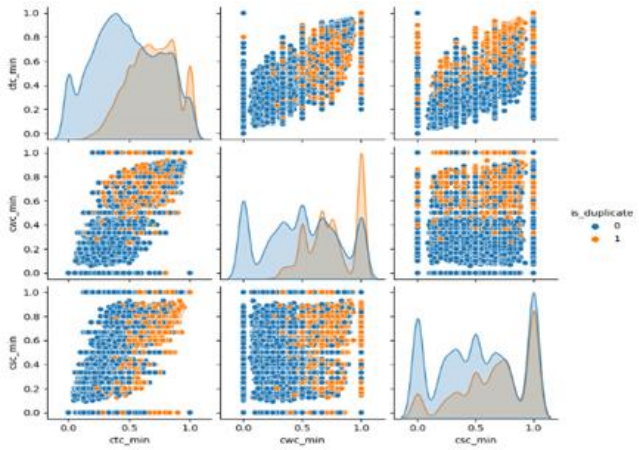
In addition to Random Forest and XGBoost models, fuzzy features, token features, and the first and last word of a sentence can also be utilized to improve the performance of duplicate or not question analysis.

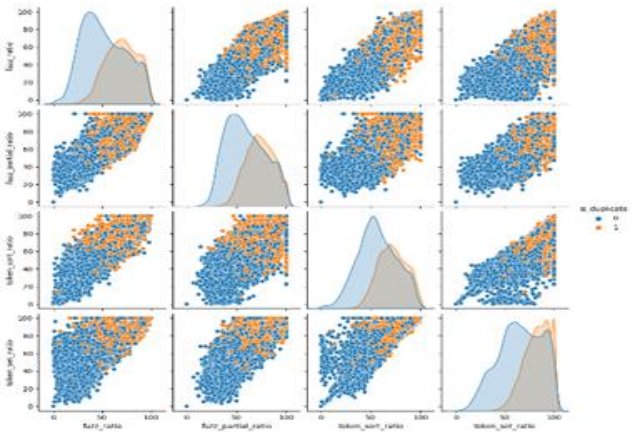
Fuzzy features involve calculating the similarity between two sentences based on measures like Lowenstein distance or cosine similarity.

Token features involve analyzing the individual words or tokens present in a sentence. This includes techniques like bag-of-words or TF-IDF (Term Frequency-Inverse Document Frequency) encoding.

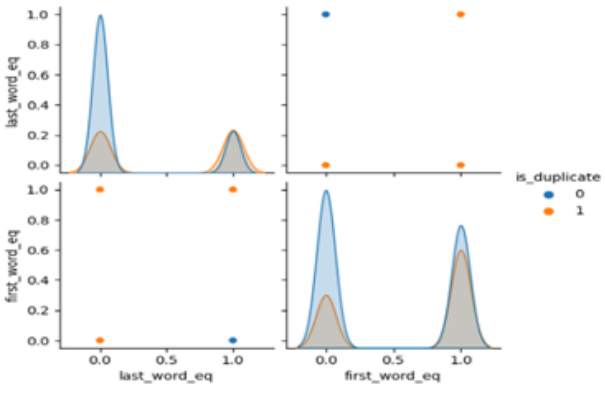
Considering the first and last word of a sentence can also provide valuable information. In some cases, questions that have identical first and last words are more likely to be duplicates. Therefore, examining these specific positions can serve as an additional clue in determining duplication.

On addressing the features to the model some of the graph to represent.





The length features and the fuzzy features are represented and had a great impact on the dataset as they provide the large differentiation of the questions duplicate.



1. **Random Forest**: The random forest model uses the formulas which are included in the process are
   1.SelectKBest

$$x^2 = \frac{\sum(Oij - Eij)^2}{Eij}$$

Random Forest:

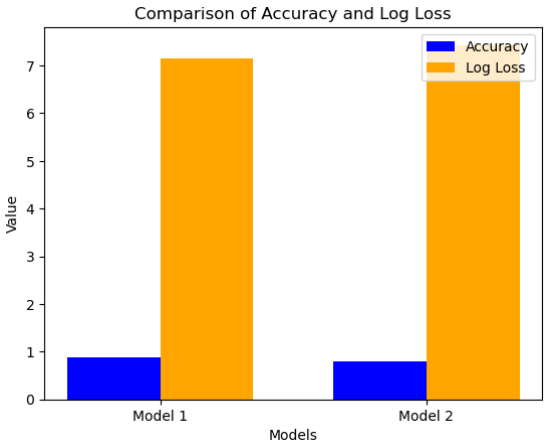
$$Gi = 1 - \sum_{i-1}^{n} (x_i)^2$$

2. **XGBoost:** The XGBClassifier uses the formula as:

$$l = \sum_{i-1}^{n} (Yi, \widehat{Yi^{(t-1)}} + ft(Xi))$$

Table -1: Accuracy of modules

SI NO	MODEL	ACCURACY(%)	CONFUSSION MATRIX	LOG LOSS
1.	Random Forest	89.8	array([[3268, 544], [747, 1441]], dtype=int64)	7.43
2.	XGBClassifier	79.2	array([[3228, 584], [660, 1528]], dtype=int64)	7.16



Here is the chart which represents the accuracy and logloss of the random Forest and XGBoost models where we can see that random forest give more accuracy.

```
XGBoost Classifier:
Accuracy: 0.7926666666666666
confussion matrix: [[3228  584]
 [ 660 1528]]
log loss: 7.161117466694687
```

Random Forest gives the more accuracy when the SelectKBest with chi area added to it and with minmax approach.

```
Random forest:
Accuracy: 0.89
confussion matrix: [[3268  544]
 [ 747 1441]]
log loss: 7.431665884471952
```

VI. CONCLUSIONS

The application of Random Forest and XGBoost models in analyzing duplicate or not questions has shown commendable performance, achieving an accuracy of 89% and accuracy of 78%. These ensemble learning algorithms are known for their ability to handle complex tasks and deliver robust performance. The high accuracy achieved indicates the models' success in identifying duplicate question pairs with a high level of correctness. With a precision of 78%, the models made accurate predictions when classifying pairs as duplicates. These results demonstrate the effectiveness of Random Forest and XGBoost models in accurately classifying duplicate question pairs, highlighting their potential for natural language processing and information retrieval tasks

VII. FUTURE ENHANCEMENTS

For future enhancements, improving duplicate question pair classification models can be achieved by incorporating advanced natural language processing techniques like word embeddings or contextualized representations (e.g., BERT) to capture richer semantic information. Employing deep learning architectures like CNNs or RNNs can enhance the model's ability to capture complex patterns. Utilizing ensemble methods with Random Forest and XGBoost can boost overall accuracy and precision. Additionally, considering domain-specific features and transfer learning approaches can further improve performance on niche datasets.

VIII. ACKNOWLEDGMENT

While working on synopsis and on completion of it, various people helped me guided me to see the way and helped me to complete it on time with 100per efforts of all of them.
I would like to thank our principal Dr. Aswath M U, Bangalore Institute of Technology, Bengaluru who gave me this wonderful opportunity to work on this synopsis. I would like to thank to our HOD and Guide Dr. T. Vijaya Kumar who gave me this wonderful opportunity to work on this synopsis Thanks to my dearest friends without whom I wouldn't be able to see different approaches to the topic that I've selected and giving me random information and their experiences, which helped me to see through various angles and problems also I would like to thank Dr. T. Vijaya Kumar for providing me the details of the project and also how to approach this problem in a way and all the details about software and hardware.

IX. REFERENCES

[1] Broder,A "On the resemblance and containment of documents. Proceedings of the Compression and Complexity of Sequences" 1997, SEQUENCES'97, Washington, DC, USA. IEEE Computer Society.

[2] Ameya Godbole, Aman Dalmia, Sunil Kumar Sahu "Siamese Neural Networks with Random Forest for detecting duplicate question"

[3] Quan Z. Sheng, Wei Emma Zhang, Jey Han Lau, Ermyas Abebe, and Wenjie Ruan.2018. "Duplicate Detection in Programming Question Answering Communities. ACM Trans. Internet Technol". 18, 3, Article 37 (April 2018),21 pages)

[4] Wenpeng Yin, Hinrich Schutze, Bing Xiang, and Bowen Zhou. "ABCNN: attention based convolutional network" abs/1512.05193.http://arxiv.org/abs/1512.05193.

[5] Tianqi Chen, Carlos Guestrin "XGBoost: A Scalable Tree Boosting System" August 2016

[6] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral "multi-perspective matching for natural language sentences" arXiv preprint arXiv:1702.03814, 2016

[7] Tim Rocktaschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, Phil Blunsom " Reasoning about entailment with neural attention" In ICLR 2016.

[8] Li zhang,Jing Jiang Liting Wing "Duplicate Question Detection with deep learning in Stack Overflow" 2020

[9] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. "Effective Approaches to Attention-based Neural Machine Translation". In EMNLP, 2015

[10] Jonas Mueller and Aditya Thyagarajan. "Siamese Recurrent Architectures for Learning Sentence Similarity". In AAAI, 2016.

[11] Oliver E. Durham Abric, Clarkand Shane McIntosh Can "Duplicate Questions on Stack Overflow Benefit the Software Development Community"

[12] LEO BREIMAN: "Random Forests Statistics Department", 2018

[13] Pritam Patil, Amol Dattu, Uday Patel, Prof Sujit Tilak and Renuka Khot "Quora Question Duplication Problem" May 2020

[14] Shuohang Wang and Jing Jiang. "Learning natural language inference with LSTM". In Proceedings of NAACL, 2016

[15] Gabor Angeli and Christopher D. Manning. Naturalli: "Natural logic inference for common sense reasoning". In EMNLP, 2014.

[16] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. "A large annotated corpus for learning natural language inference". In EMNLP, 2015

[17] Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, Wojciech Zaremba. "Addressing the Rare Word Problem in Neural Machine Translation". In ACL, 2015

[18] Kaggle: Your home for data science [online] available: https://www.kaggle.com/

[19] Quora question pairs — kaggle [online] available: https://www.kaggle.com/c/quoraquestion-pairs

[20] https://rb.gy/usj8sj. 2019

[21] https://rb.gy/7nzyoq. 2020