Duplicate Question Pair Detection with Machine Learning

Ms. Vishwaja M . Tambakhe

Department of Computer Science and Engineering , Government College of Engineering Amravati , Maharashtra , India.

Dr. Kishor P.Wagh

Department of Information and Technology, Government College of Engineering, Amravati, Maharashtra, India.

Abstract- **Duplicate or inconsistent records in databases can have a significant impact, which has led to the development of a variety of strategies for detecting such records in general databases. The most common issue found while using Q&A sites like Quora, Stack Overflow, Reddit, and others is question repetition. Answers become disjointed throughout one of a kind variation of the identical query because of the repetition of questions in these boards. This eventually leads to the loss of a rational searching, solution weariness, statistic segregation, and a scarcity of responses to the questioners. Machine Learning and Natural Language Processing can be used to detect duplicate inquiries. Tokenization, lemmatization, and the deletion of stop words are used to preprocess a dataset of over 400,000 question pairings obtained from Quora. The function extraction is performed on this pre-processed dataset. Machine learning techniques, in particular, are commonly utilised for locating duplicate records in large datasets, but only a few have been suggested. In this work we are using four classifiers for the classification using machine learning.**

Keywords – Machine Learning , Natural Language Processing , Question Duplication.

I. INTRODUCTION

Question-and-solution (Q&A) web sites together with Quora offer customers with a platform to invite 1 question that different customers at the web website online may also solution. However, a few of the questions being requested at any given time have already been requested via way of means of different customers, generally with a different phrasing or wording Ideally, the replica inquiries would be consolidated into a single canonical query, as this would provide the following benefits:

•If the query asker's question has previously been addressed at the web website online, it saves them time. Instead of waiting minutes or hours for a response, clients can have their answer right away.

• Repeated enquiries can irritate even the most devoted consumers, whose feeds get clogged with duplicate queries. Many customers who answer questions in a specific subject see light versions of the same query appearing repeatedly in their feed, which causes a terrible user experience for them.

• Customers and researchers pay more for Q&A data bases since there is a single canonical query and collections of replies, rather than the information being fragmented and spread throughout the web site online. This cuts down on the time it takes for customers to find the best responses and allows researchers to better understand the relationship between queries and answers.

• Having knowledge of many ways to phrase the same inquiry can help with search and discovery. The ability to search for full-text content is a valuable feature of Q&A sites, however its software is confined via way of means of wanting to question for near-genuine query phrasing. Having multiple illustration of the identical query can enhance this seek manner substantially for customers.

If the query askers indicated the same intent while creating the query, we say the inquiries are duplicates. That is, any legitimate answer to at least one question is also a legitimate answer to the other. For instance, “What is the shortest way to go from Los Angeles to New York?” “How do I go from Los Angeles to New York in the shortest amount of time?” and “How do I get from Los Angeles to New York in the shortest amount of time?” are said to be identical. It's worth noting that certain questions have intrinsic ambiguity based just on their texts, and we can't claim for certain that they all express the same purpose. “How do I make $100k USD?” and “How do I acquire 100 grands?”, for example, may be same if we assume that the “hundred grand” preferred is in US dollars, but this is not always the case true. Often, any human labeling manner will replicate this ambiguity, and introduce a few quantities of noise to the dataset .

II. RELATED WORK

W-shingling (Broder [1]) has been successfully utilised to quantify the similarity across textual content documents in traditional herbal language processing (NLP). However, because reproduction questions can be rephrased in a variety of ways, techniques that rely on phrase overlap fall short in this project, as we demonstrate in our tests . CNNs have shown substantial promise over traditional NLP methodologies when it comes to sentence classification and sentiment analysis (Wu [2]). Taking sentence inputs that have been shortened to the shortest possible length, the phrases of each sentence are turned into a matrix of pre-skilled phrase embeddings using word2vec (Mikolov et al. [3]) . The version has been shown to achieve exact results in a variety of sentence class tasks, including sentiment analysis. Bogdanova et al. [4] used StackExchange query data to test this method for reproducing query pair identification, and the results were very reliable on two very technical datasets (AskUbuntu forums) . Recent instructional interest in recreating 3 query pair came across has been noted since the release of Quora's initial public dataset. Wang et al. [5], just prior to the publication of this research, used bidirectional LSTMs to solve the problem of query pair identification, and then used today's results with hand-tuned cross-query capabilities in a system they dub "mutli-attitud matching”. These works laid the groundwork for attempting to apply an LSTM encoding to this project, which led to the development of a hybrid LSTM/CNN encoding. For many years, natural language sentence matching (NLSM) has been explored . Early strategies [Heilman and Smith, 2010; Wang and Ittycheriah, 2015] focused on building hand-craft functions to capture n-gram overlapping, phrase reordering, and syntactic alignments phenomena. [6], [7], and [8], respectively . This type of technique may work well for a specific assignment or dataset, but it is difficult to apply to other jobs. Many deep investigating methods for NLSM have been proposed as a result of the availability of large-scale annotated datasets [Bowman et al., 2015] [9]. The first type of framework is based on the Siamese architecture [Bromley et al., 1993] [10], in which sentences are encoded into sentence vectors using a few neural community encoders, and the relationship between sentences is then determined entirely based on the sentence vectors [Bowman et al., 2015; Yang et al., 2015; Tan et al., 2015] [11], [12], and [13]. This theory, on the other hand, ignores the fact that the lower stage interactive functions between sentences are critical. As a result, different neural community styles [Yin et al., 2015; Wang and Jiang, 2016; Wang et al., 2016] have been presented to suit phrases from various levels of granularity [14].

III.PROPOSED WORK

Question duplication is the most common issue on Q&A sites like Quora, Stack Overflow, and Reddit. Because of the repetition of questions, answers become dispersed among different variations of the same question, to overcome this NLP is used by using Machine learning. Natural Language Processing, or NLP, is a branch of Artificial Intelligence that allows robots to read, understand, and interpret human languages. To discover duplicate questions in order to reduce data redundancy , the process of finding word-based similarity between the two questions, then classify question pairs with similarity scores above a certain threshold as duplicates . Data Extraction, Data Pre-processing, and Feature Extraction are some of the components of the proposed system and Classification. The data is passed through each and every component respectively to maintain the structure of the data before it is trained by the Machine learning model. The data fed into the model can affect the efficiency if it is not passed through the above components.

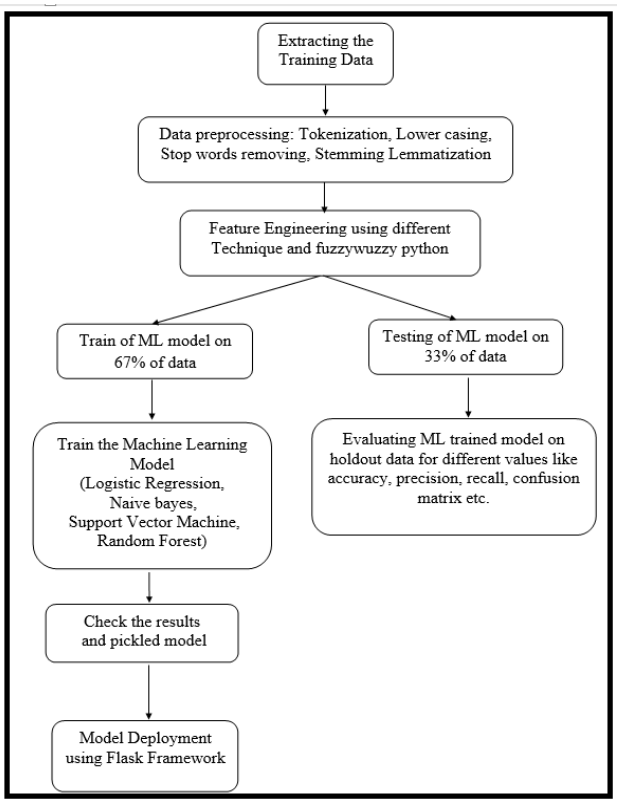

Figure 1: Flow of Proposed Methodology

The working of proposed methodology for detection of Duplicate Question following are the steps.

Step 1: Exploring the Dataset

On the Kaggle opposition internet page, we've get right of entry to the dataset and check statistics units. The dataset includes 404,290 query pairs (rows) and has 6 variables (columns). The columns are the row ID, query 1, query 2, query 1 ID, query 2 ID and the magnificence label that's zero for non-replica pairs and 1 for replica pairs. The query IDs (qids) uniquely pick out every query. The check set includes over one million query pairs (a lot of which might be purposefully laptop generated to save you dishonest withinside the contest). The check set includes three columns which might be the row ID, query 1 and query 2. I determined to alternatively paintings in simple terms at the educate set, on which I carried out a 67–33 dataset break up. For the ones unfamiliar, we break up the statistics into educate and check units and the check set isn't touched at the same time as the version is educated and at the same time as vectorizing our textual features. The dataset is what we educate our fashions on and the check set is how we compare the overall performance of our fashions. Improper utilization of the check set results in a hassle called statistics leakage. After our break up, our training and testing dataset have 67% non-replica pairs and 33% replica pairs. The distribution isn't best however honestly workable.

Dataset fields :-

- id - the unique identifier for a training set question pair
- qid1 and qid2 are the unique identifiers for each question (only available in train.csv)
- question1, question2 - each question's complete text
- is duplicate - the target variable, which is set to 1 if question1 and question2 have the same meaning, and 0 if they don't.

Step 2: Data Cleaning and Preprocessing

In the records cleansing stage, we attempt to cast off any awful observations from the records. These can be reproduction rows, i.e. rows that have the identical qids. Usually, there are a few observations which do now no longer make sense, as an instance if we had a column for "time taken", it cannot be negative. We might cast off that row. This is a completely area particular problem. In our case, I checked for reproduction rows and there had been none.
In the records preprocessing stage, we smooth up every individual row's records. This is an essential step withinside the in the data science and machine learning pipeline. The preprocessing steps had been achieved had been:
• Convert textual content to decrease case.
• Remove punctuation.
• Replace a few numerical values with strings (Eg: one million with 1m).
• Remove HTML tags.
• Replace a few characters with their string equivalents (Eg: $, % @ etc.).
• Decontract words ("don't" becomes "do not").
• Stop phrase removal.

Step 3: Feature Extraction

Feature Extraction is manner wherein we extract capabilities from our statistics. These capabilities are what we are able to be feeding into our device studying algorithms. Featurization is pretty probable the maximum crucial element of device studying. It is the so called "art" a part of statistics science. I actually have extracted the subsequent capabilities: Token capabilities are capabilities which might be extracted from studying the tokens in a query (a token in this situation is every phrase). The token capabilities I extracted are:
1. q1_len: Number of characters in query 1.
2. q2_len: Number of characters in query 2.
3. q1_words: Number of phrases in query 1.
4. q2_words: Number of phrases in query 2.
5. words_total: Sum of q1_words and q2_words.
6. words_common: Number of phrases which arise in query 1 and, repeated occurrences aren't counted.
7. words_shared: Fraction of words_common to words_total.
8. cwc_min: This is the ratio of the quantity of not unusualplace phrases to the period of the smaller query.
9. cwc_max: This is the ratio of the quantity of not unusualplace phrases to the period of the bigger query.
10. csc_min: This is the ratio of the quantity of not unusualplace forestall phrases to the smaller forestall phrase depend most of the questions.
11. csc_max: This is the ratio of the quantity of not unusualplace forestall phrases to the bigger forestall phrase depend most of the questions.
12. ctc_min: This is the ratio of the quantity of not unusualplace tokens to the smaller token depend most of the questions.
13. ctc_max: This is the ratio of the quantity of not unusualplace tokens to the bigger token depend most of the questions.
14. last_word_eq: 1 if the final phrase withinside the questions is same, zero otherwise.
15. first_word_eq: 1 if the primary phrase withinside the questions is same, zero otherwise.
16. num_common_adj: This is the quantity of not unusualplace adjectives in question1 and question2.
17. num_common_prn: This is the quantity of not unusualplace right nouns in question1 and question2.
18. num_common_n: This is the quantity of nouns (non-right) not unusualplace in question1 and question2.

FuzzyWuzzy is a Python library which has a few techniques to evaluate string equivalence. I used one-of-a-kind string assessment strategies from FuzzyWuzzy to extract fuzzy capabilities. The fuzzy capabilities are:

1. fuzz_ratio: fuzz_ratio rating from fuzzywuzzy.
2. fuzz_partial_ratio: fuzz_partial_ratio from fuzzywuzzy.
3. token_sort_ratio: token_sort_ratio from fuzzywuzzy.
3. token_set_ratio: token_set_ratio from fuzzywuzzy.

Step 4: Applying machine learning for classification

After cleaning, preprocessing and Featurization of dataset, the final step is to apply classifiers, here we are using random forest, support vector machine, naive bayes, logistic regression, and support vector machine .

i. Support Vector Machine

Support Vector Machine (SVM), a machine learning method extensively utilised by many academics to cope with static motions, was chosen as the classification approach. Its major goal is to locate multiple hyper-planes that can divide all data samples into groups and thus complete data categorization .

In a multi-dimensional space, hyperplanes can be thought of as decision boundaries that classify data points into their appropriate classes. Different classes can be assigned to data points on either side of the hyperplane.

Hyperplane is a plane that has been generalised :

- It's a line in two dimensions.
- It's a plane in three dimensions.
- It's known as a hyperplane in higher dimensions.

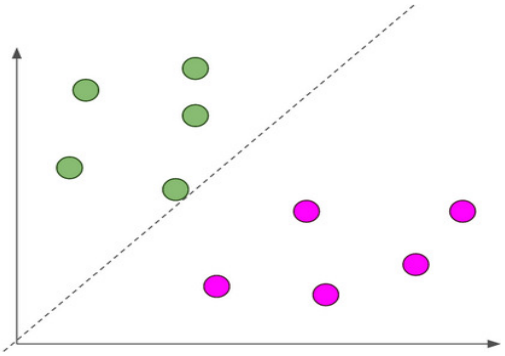

Figure 2: Representation of Hyperplane in SVM
(dotted line is hyperplane, separating blue and pink classes balls.)

ii. Naïve Bayes

Naïve Bayes set of rules is a supervised getting to know set of rules, that is primarily based totally on Bayes theorem and used for fixing class problems. Bayes' Theorem facilitates you study the opportunity of an occasion primarily based totally at the earlier information of any occasion that has correspondence to the previous occasion. Its makes use of are specially observed in opportunity principle and statistics. The time period naive is used withinside the experience that the capabilities given to the version aren't depending on every other. In easy terms, in case you alternate the cost of 1 function withinside the set of rules, it'll now no longer without delay affect or alternate the cost of the alternative capabilities. Consider for instance the opportunity that the rate of a residence is excessive may be calculated higher if we've a few earlier statistics just like the centers round it as compared to any other evaluation made without the information of the region of the residence.

$$P(A|B) = [P(B|A)P(A)]/[P(B)] \quad (1)$$

The equation above suggests the primary illustration of the Bayes' theorem wherein A and B are activities and: P(A|B): The conditional opportunity that occasion A occurs, for the reason that B has occurred. This is called because the posterior opportunity. P(A) and P(B): The opportunity of A and B with none correspondence with every other. P(B|A): The conditional opportunity of the prevalence of occasion B, for the reason that A has occurred.

iii. Logistic Regression

Logistic Regression is Supervised Machine Learning Algorithms in Machine Learning. It has an ability to predict whether an observation belongs to a certain class using an approach that is straightforward, easy-to-understand. Logistic Regression is a Binary Classifier. This means that the target vector may only take the form of one of two values. In the Logistic Regression Algorithm formula, we have a Linear Model, e.g., β0 + β1x, that is integrated into a Logistic Function (also known as a Sigmoid Function). The Binary Classifier formula that we have at the end is as follows:

$$P(y_i=1 \mid X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}} \quad (2)$$

Where:

The probability P(yi = 1 | X) of the ith observations target value, yi belonging to class 1.
B0 and β1 are the parameters that are to be learned.
e represents Euler's Number.

The Logistic Regression formula aims to limit or constrain the Linear and/or Sigmoid output between a value of 0 and 1. The main reason is for interpretability purposes, i.e., we can read the value as a simple Probability; Meaning that if the value is greater than 0.5 class one would be predicted, otherwise, class 0 is predicted.

iv.Random Forest

Random forest is a learning algorithm that is supervised. It creates a "forest" out of an ensemble of decision trees, which are commonly trained using the "bagging" method. The main notion of the bagging approach is that a mixture of learning models increases the total output. Random forest has the advantage of being able to solve classification and regression issues, which make up the majority of contemporary machine learning systems. In classification, random forest is sometimes referred to as the "building block" of machine learning. The hyperparameters of a random forest are quite similar to those of a decision tree or a boosting classifiers.

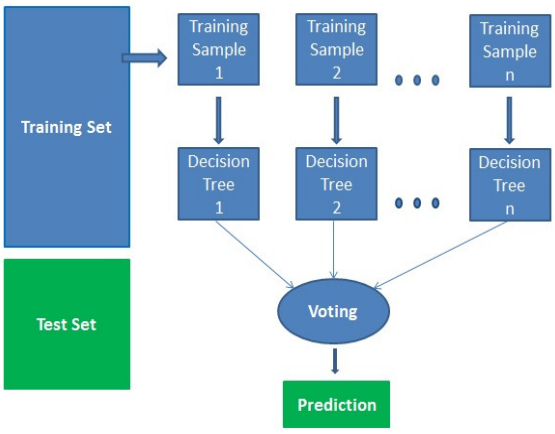

Figure 3: Working of Random Forest

IV. CONCLUSION

This study uses Machine Learning and Natural Language Processing to classify whether question pairings are duplicates or not in Q&A forums. The use of minimal cost architecture and the selection of highly dominating elements from the questions make it an effective template for detecting duplicate inquiries and subsequently finding high-quality answers.

REFERENCES

1) Broder, A. (1997) 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) Yoon Kim. (2014) Convolution neural networks for sentence classification. Proceedings of the 2015 Conference on Empirical Methods for Natural Language Processing, pages 1746-1751. Doha, Qatar.

3) Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. (2013). Efficient estimation of word representations in vector space. Proceedings of International Conference on Learning Representations, ICLR 2013, Scottsdale, AZ, USA.

4) Dagna Bogdanova, C´ıcero dos Santos, Luciano Barbosa, and Bianca Zadrozny. (2015). Detecting Semantically Equivalent Questions in Online User Forums. Proceedings of the 19th Conference on Computational Language Learning, pages 123–131, Beijing, China, July 30-31, 2015.

5) Zhiguo Wang, Wael Hamza, Radu Florian. (2017). Bilateral Multi-Perspective Matching for Natural Language Sentences. https://arxiv.org/abs/1702.03814.

6) [Bowman et al., 2015] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326, 2015.

7) [Bromley et al., 1993] 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. IJPRAI, 7(4):669– 688, 1993.

8) [Chen et al., 2016] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, and Hui Jiang. Enhancing and combining sequential and tree lstm for natural language inference. arXiv preprint arXiv:1609.06038, 2016.

9) [Cheng et al., 2016] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. arXiv preprint arXiv:1601.06733, 2016.

10) [He and Lin, 2016] Hua He and Jimmy Lin. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In NAACL, 2016.

11) [Heilman and Smith, 2010] Michael Heilman and Noah A Smith. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In NAACL, 2010.

12) [Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. ¨ Neural computation, 9(8):1735–1780, 1997.

13) [Kingma and Ba, 2014] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

14) [LeCun et al., 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 521(7553):436– 444, 2015

15) P.A. Jadhav, P. N. Chatur and K. P. Wagh, "Integrating performance of web search engine with Machine Learning approach," *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 2016, pp. 519-524, doi: 10.1109/AEEICB.2016.7538344.

16) P. P. Shelke and K. P. Wagh, "Review on Aspect based Sentiment Analysis on Social Data," *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2021, pp. 331-336, doi: 10.1109/INDIACom51348.2021.00057.

17) Ms. Vishwaja M. Tambakhe, Dr. Kishor P.Wagh, "Review on Exploring Similarity between Two Questions Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology

(IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 3, pp. 287-293, May-June 2021. Available at doi : https://doi.org/10.32628/CSEIT217360 Journal URL : https://ijsrcseit.com/CSEIT217360