Review on Question Pair Similarity Using Machine learning

Kshipra Deshmukh
Department of CSE
H.V.P.M.Collage of Engineering and Technology Amravati,India

Dr. A. B. Raut
Department of CSE
H.V.P.M.Collage of Engineering and Technology Amravati,India

*ABSTRACT:*Platforms for questionnaires like Quora and Stack Overflow that are built on features that lets users ask questions and allow them to react to questions frequently experience semantic matching or question duplication. We frequently have a propensity to construct one or two sentences that are distinctly influenced by the linguistic environment we live in, our accent, and our surroundings. These techniques may have a major negative impact on users who ask inquiries that are semantically similar but are not marked as duplicates. Although it's not a perfect solution, the pattern we've found could make the model better at predicting the duplication of questions across question pairings. Question duplication is the fundamental problem that Q&A communities like Quora, Stack Overflow, Reddit, etc. confront. Because questions are asked repeatedly in online forums, the solutions are spread out among many variations of the same question. As a result, there will finally be no reasonable search, answer fatigue, information segregation, and no responses to the questioners. Finding the duplicate inquiries can be done using machine learning and natural language processing.

Key-Words: - *Machine Learning; Question Pair Similarity; XGBoost;Logistic Regression; Random Forest.*

1.Introduction

The need for a software system to locate information was first mentioned in 1945 in Vannevar Bush's controversial essay "As We May Think." He offered study libraries with related explanations that function like hyperlinks. The first systematic search engine, Archie, developed by Alan Emtage in 1990, was followed by the World Wide Web A few examples of search engines are Wanderer, Aliweb, JumpStation, and WebCrawler, as well as Yahoo, Megallan, Excite, Infoseek, Inktomi, Northern Light, AltaVista, and Bing. Access to pertinent online pages didn't happen until Larry Page's Google search engine. The major objective of the study is to find the best machine learning approach to eliminate all the repetitive queries. A thorough study shows that many machine learning methods take a lengthy time to train on real-time information. The response time for using SQLite to preprocess the dataset is reduced by one-quarter when compared to preprocessing the same dataset manually, and this study will highlight the key considerations. The effectiveness of machine learning algorithms will then be assessed using response time and the error log loss function. Using the characteristics of questions 1 and 2, the feature extraction of the related question pairs in the Quora dataset is assessed. These characteristics include frequency, length, word count, common word count, overall word count, frequency sum, absolute frequency difference, etc.

A place to learn and share knowledge about anything is Quora. It is a platform that fosters connection, encourages the exchange of distinctive information and insights, and provides high-quality responses to a range of queries. Because each person has a unique perspective on the situation and differing viewpoints, this approach enables people all over the world to learn

a great deal from one another and improves our understanding of the world. Every day, millions of individuals visit Quora, which results in a wide range of queries from users and a wide range of replies from different writers for the same questions, confusing the user as to which response he should take into consideration as the proper one. Due to the fact that there are numerous possible responses to the identical questions, finding the proper answer takes a lot of time. Finding the best outcome or response is really challenging. Quora adheres to canonical questions because they provide a better user experience for both active seekers and writers, as well as longer-term benefits for both of these groups. In order to solve the problem of paring up the duplicate questions from Quora, NLP is employed in conjunction with machine learning. Then, identify the question pairings with similarity scores that are greater than a predetermined threshold as duplicates after determining the word-based similarity between the two questions. We must make sure that each distinct question appears on Quora only once in order to create a high-quality knowledge base. Readers should be able to locate a single canonical page with the question they're looking for, and writers shouldn't be required to provide the same response to numerous variations of the same topic. The following are our formal problem statements. Determine whether Quora questions are repeats of other questions that have already been posed.

- This can be helpful for providing prompt responses to questions that have already been addressed.
- Tasks that determine whether or not a pair of questions are duplicates.

A machine learning and natural language processing system is created to automatically recognise whether questions with the same purpose have been asked several times in order to stop duplicate questions from being on Quora. In order to train a model, a fundamental process flow must be followed, including data extraction, data pre-processing, and feature extraction.

2.Literature survey:

Because the problem of question duplication is common across several applications, research towards developing an accurate answer has been ongoing for some time. Previous studies identified question pairings that copied each other based on their semantics using Support Vector Machines (SVMs) [2] and other traditional machine learning techniques. However, since the introduction of machine learning techniques, a wide range of models have yielded some amazing results.

In a few sentimental analysis tasks, such as classification tasks and determining the true nature of the sentences by comparing them to the phrase's semantics, Convolution Neural Networks (CNN) [3] have shown promise. However, most of them are Xiv:1801.07288v in the CS .CL] 28 Jan 2018 [4]. Machine learning techniques that have been published for detecting sentence duplication use a Siamese neural network architecture. Prior to comparison using a distance metric[5], this architecture uses a neural network to gather features from the input phrases. You played a crucial role in several tasks. A proposed method combines the outputs from both smaller processing networks after splitting them into two smaller networks[6]. The model is lightweight and easy to train, but there is no particular relationship that guards against knowledge loss. To overcome the drawbacks of the Siamese neural network, a Compare-Aggregate model that analyses and notes the similarity of duplicate questions and is used to train duplicate question detection models was presented [7]. We gained a lot of knowledge from studies conducted at New York University and the Department of Computer Science at Stanford University [8]. Evaluation was done on a model

for extracting various features [9] that also considers the fuzzy idea and vector distances of the texts [10].

3.RELATED WORK

Multiple applications are experiencing the question duplication issue, and it is taking a long time to find an accurate answer. [1] Previous research has used typical machine learning methods like Support Vector Machines to identify question pairs that are duplicates based on the semantics of the duplicate question pairs. [2] But as Deep Learning, Artificial Intelligence, Neural Networks, and Natural Language Processing approaches have grown in popularity, this diverse variety of models has produced some striking outcomes. Convolution Neural Networks (CNN) [3] have demonstrated promising performance in a few sentimental analysis tasks, such as categorization and identifying the true character of the sentences by comparing them to the phrase's semantics.[4] However, Siamese neural networks have been exploited by deep learning techniques developed for validating sentence duplication. When comparing input words in neural network architecture, a distance metric is utilised. The input sentences are processed by utilising neural network to extract the features. Distance metric is supervised learning that determines how similar two data points are to one another by calculating their similarity [5].

In this proposed work, a Siamese neural network is used for processing. Neural networks have prominent rules that are applicable to a wide range of natural language processing applications. This siamese neural network is processed in such a way that it comprises two or more identical sub-networks, where identical is defined as using the same weights and setup settings to detect similarity. [6]. Siamese neural networks do have certain limitations. For instance, because the model's training was simple and quick, there is no assurance against information or data loss because there is no clear relationship between the processes. A Compare-Aggregate model was designed in order to circumvent the Siamese neural network; this model has excellent observational skills and can detect similarities between two texts. Duplicate questions won't train for the question detection models in the dataset that the Quora data science engineer publicly provided [7]. This suggested model uses that dataset. We were able to learn a lot from research conducted at the Department of Computer Science, Stanford University [8], and the University. The evaluation of a model for extracting various features, which also superimpose the idea of fuzzy and vector distances of the texts, was managed or administered in those places [9]. Prior to this, logical inference based on the Stanford Natural Language Inference Corpus has been the main focus of semantic matching of sentences. The subject of Rocktaschel's paper [10] was word-by-word concentration techniques employing LSTMs. The first challenge to feature a task based on question-question similarity was SemEval, which was focused on semantic similarity [11].

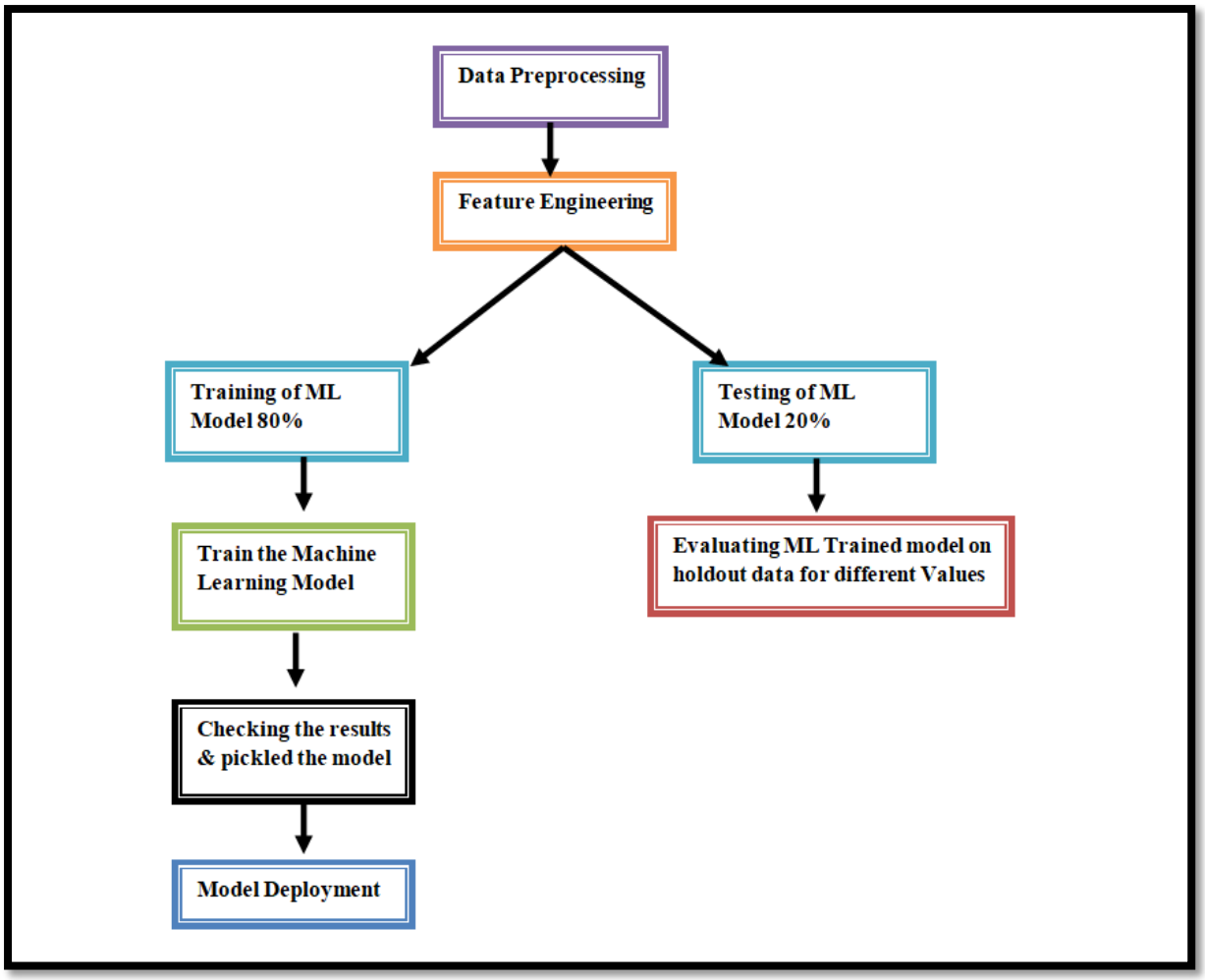
The larger task of semantic text similarity, which has been the focus of the SemEval challenges since 2012 [12], includes duplicate question detection. Word overlap and conventional machine learning methods, such as Support Vector Machine, were initially used in early attempts to identify sentence similarities [13]. In a variety of NLP problems, neural network techniques have been the state-of-the-art. A suggested Siamese neural network with two sub-networks connected at their outputs. [14]. We learned a lot from research in the Department of Computer Science at Stanford University and New York University. The notion of fuzzywuzzy and text vector distances were included in the evaluation of the model

used to extract different characteristics [15]. Following the public release of Quora's inaugural dataset, there has been an increase in interest in academic research for the detection of duplicate question pairs. And the author wang suggested a bilateral LSTMs to this duplication problem by combining the hand-tuned cross-question feature with the state-of-the-art result from the Archievung algorithm. This procedure is known as multi viewpoint matching. The goal of this work was to apply LSTM encoding, and as a result, a hybrid LSTM model was produced. [16]

There are two different deep learning frameworks for neural network models, both of which were suggested by NLSM. Siamese architecture served as the first framework, and matching aggregation served as the second. Siamese framework matching decisions were made exclusively based on two vector sentences, whereas attributes of two sentences were gathered by CNN or LSTM in matching aggregation frameworks. As a result, it needs more significance to be improved. [17]. The matching aggregation approach has various limitations. The first one is that it only examines the matching process phrase by phrase or word by word. The second disadvantage was that it only matches in one direction, P against Q, and ignored the other, reverse direction, Q against P. The BiMPM model overcame this flaw in matching aggregation by matching both P against Q and Q against P. The match was conducted using three layers: answer sentence selection, paraphrase identification, and natural language inference. In all of the jobs, our technique has achieved cutting-edge performance [18].

4.PROPOSED WORK

The main issue with Q&A websites like Quora, Stack Overflow, Reddit, etc. is question duplication. Due to question repetition, answers to the same question become dispersed throughout several versions. Machine learning is employed because it is a branch of artificial intelligence that enables machines to read, comprehend, and extrapolate meaning from languages used by people. The procedure of locating word-based similarity between the two questions, then classifying question pairs with similarity scores above a particular threshold as duplicates, is used to detect the duplicate questions for decreasing the redundancy in data. Data extraction, data pre-processing, feature extraction, and classification are some of the components that make up the suggested system. Before the machine learning model trains the data, the input is passed through each component in turn to retain its structure. If the data is not delivered via the aforementioned components, it may damage the efficiency of the model.



5.Conclusion

In order to classify whether or not question pairings are duplicates, machine learning methods are used in this research to solve the issue of question duplication in Q&A forums. It is an efficient typology to identify duplicate questions and eventually locate high-quality responses in Q&A forums thanks to the use of lowest cost architecture and the selection of very dominating aspects from the queries.

References:

[1] Martin Aabadi, Aashish Aagarwal, Paul Barhaam, Eugene Brvdo, Zhifng Chen, Craaig Citro, Greg S Corado,Andy Davis, Jeffrey Dean, Matthieu Devin, et al.2016. Tennsorflow:Largescale machiinelearnning on heterogeneous distrbute systems. arXiv preprintarXiv:1603.04467

[2] YEUNG, K. (2016, March 17). Quora has millions of daily visitors, up from 80 million in January. https://venturebeat.com/2016/03/17/quora-now-has-100-million-monthly-visitors-up-from-80-million-in-january

[3] Lili Jiang, S. C. (n.d.). Quora: https://engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning

[4] mccormickml. (2016, April 12). Retrieved from mccormickml: http://mccormickml.com/2016/04/12/googlespretrained-word2vec-model-in-python

[5] Machine Learning Mastery. (2017, June 15). https://machine learning mastery.com / prepare-textdata-machine-learning-scikit-learn.

[6] Brownlee, J. (2017, October 19). A Gentle Introduction to the Bag of Words Model.https://machinelearningmasterry.com/gentle-introduction-bag-words-model

[7] Rajaraman, A.; Ullman, J.D. (2011). "Data Mining". Minning of Masive Dataset(PDF). pp. 1–17.doi:10.1017/CBO981139058452.002. ISBN 978-1-139-05845-2

[8] Gilyadov, J. (2018, March 23). Word2VecExplained.Retrieved github.io: https://isrelg9.github.io/2018-03-23-Word2Vec-Explained.

[9] McComick, C. Google's trainedWord2Vec model in Python. Retrieved from mcrmickml.com:http://mccormickml.com/2016/04/12.

[10] Thakur, A. (2017, Feb 27). "Is That a Duplicate Quora Questions?" Retrieved from Linkedin: https://www.linkein.com/pulse/duplicate-quora-questiona bhishek - thakur.

[11] Tim Rocktahel, Edward Grefenstte, Karl Mortz Herman, Tomas, Phil Bluom. Reasoning about entailment with neural attention. In ICLR 2016

[12] E. Agirre, C. Banea, D. Cer, M. Diab, A. GonzalezAgirre, R. Mihalcea, G. Rigau, and J. Wiebe, “Semeval2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation,” in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 497–511.

[13] E. Agirre, A. Gonzalez-Agirre, D. Cer, and M. Diab, “Semeval-2012 task 6: A pilot on semantic textual similarity,” in Proceedings of 1st Joint Conference on Lexical and Computational Semantics, 2018, pp. 384–392.

[14] Andri Z Brder. 1997. On resemblane and containment of docuuments. In Compresion Complexiity of Sequences 1997. Proccedings. IEEE, pages 212–219.

[15] Kauntal, Ritevik Shrivast and Sarooj Kashiik. 2016. A paraphrase and semanticsimilarity detection systemfor user generated short text content on micro blogs. In COLING. pages 2880–2890.

[16] Broley, Jane, "Signature Verification Using A "Siamese" Time Delay Neural Network." IJPRAI 7.4 (1993): 669-688.

[17] Wang, Zhguo, Wael, and Radu. "Bilatreal MultPerspective Matchingfor Natural LanguageSentences." [arXivarXv:17020814 (2019)].

[18] Wng, Shuhang, and Jing Jang. "A ComparativeAggregate ModelforMatching TextSequeces." arXiv preprint arXiv:1611.01747 (2016).

[19] Addair, T. (2016, Feb 20). "DupliicateQuestionPairDetection". Retrieved from stanford.edu: https://web.stanford.edu/class/cs224n/reports/2759336.pdf

[20] Lei Guo, C. L. (2017, Jan 16). DupliicateQuoraQuestionsDetction. Retrieved fromsemanticscholar.org:https://pdfs.semanticscholar.org/4c19/2b8f45/b1he913ee7da32624cd75/59eccb0890.pdf