International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217360

Review on Exploring Similarity between Two Questions Using Machine Learning

Ms. Vishwaja M. Tambakhe*[1], Dr. Kishor P.Wagh[2]

*[1]Department of Computer Science and Engineering, Government College of Engineering, Amravati, Maharashtra, India

[2]Department of Information and Technology, Government College of Engineering, Amravati, Maharashtra, India

Article Info
Volume 7, Issue 3
Page Number: 287-293

Publication Issue :
May-June-2021

Article History
Accepted : 15 May 2021
Published : 22 May 2021

ABSTRACT

Question duplication is the main problem which is based on functionality of allowing users to ask questions. Questions are often answered, and the duplication problem is faced by question and answer sites such as Quora and Reddit, Stack-overflow, and others. Answers are segmented through various iterations of the same question due to question continuity. The aim is to detect the duplicate questions for reducing the redundancy in data. This is a worst experience of users, as the answers get segmented on various versions of the same question, it is bad for writers as well as seekers. Actually this problem also has been noticed on other platforms of Q&A. In this proposed work a simple neural architecture with natural language inference will be used. The approach gathers an attention to pound the problem into sub-problems that helps it to be solved separately, thus making it menially parallelizable. This work is just completely a new pattern, for the solution and it is also possible that it will not provide the complete solution to the problem but may help in increasing the efficiency of the model to predict the duplication's among several question pairs. Question duplication is the serious problem due to the segmentation of answers in various variants of the same question because ofduplication's in these discussion boards. Lastly, As a consequence, there is a lack of a rational search, solution indifference, knowledge separation, and an insufficiency of responses to the questioners. This could be avoided by employing Natural Language Processing as well as Machine Learning, which will help to improve the performance as well.

Keywords: Duplication of questions, Natural LanguageProcessing and Machine Learning.

I. INTRODUCTION

Quora is a site where you can learn about anything and share anything you've learned. It's a kind of forumwhich helps people to get connected or share the unique knowledge insight and gives the quality answer to the variety of questions. This process helps people to learn a lot from each other across the world, each person have its own perspective to the situation and have different opinions too, which helps to have a deeper understanding of the universe . Every daynumber of people visiting Quora, which results in varieties of questions from the seekers and various answers for the same questions from different writers, which results in confusion to the seeker that which answer is correct among all the answers and which answer should he consider being the correct one. The another major disadvantage is that, the process of searching the correct answer is very time consuming because of different answers to the same questions. It is very difficult to find the best result or answer. Quora has some values for their questions, it follows textual questions are useful because they provide a better experience for active seekers and authors, as well as providing more long-term value to all of these classes. In this we'll be dealing with the task of pairing up the duplicate questions from quora, to overcome this NLP is used by using machine learning. To find the word-based similarity between the two questions, then classify the pairs of question with similarity scores above a certain threshold as duplicates.It's important to ensure that each specific query appears only once on Quora in order to create a high-quality knowledge base. Authors will not have to write a same answers to different versions with same question, and viewers must be able to find the question they're looking for on a single canonical post.. More formally, the followings are our problem statements

Identify that question which already asked on Quora are duplicates of questions.

- It is helpful to get an instant answer to a question that's already been answered.
- The mission should be predict whether a pair of question are duplicates or not.

A machine learning and natural language processing framework is designed to automatically recognize when questions with the same aim response have been posed multiple times on Quora, preventing duplicate questions from appearing. For this a basic process flow has to befollowed such as data extraction, data pre-processing, feature extraction which will help to train a model.

II. RELATED WORK

A main issue of question similarity wasspreading on various applications and they finding the accuracy in the solution is taking extended time. [1] Previous work to identify duplicative ness among question pairs based on the similar of the duplicate question pairs are follows some of the SVM and other traditional machine learning algorithms. [2] But from the emergence of Deep Learning, NLP, AI , Neural networks and techniques, this wide range of models have shown some impressive results. Convolution Neural Networks(CNN) [3] have shown good results in few sentimental analysis tasks where they have identify the true nature of the sentences by weighing them against the semantics used in the phrase, and classification tasks.[4] However, Deep learning methods proposed for validating duplicative ness among the sentences have used a Siamese In neural network architecture a distance metric is used for the comparison of input sentences, the input sentence are carried out by extracting the features by using neural network. Distance metric is supervised learning which calculate the similarity in between data points, which helps to define the similarity between them [5].

Neural network has its prominent rules which play on wide range of natural language processing tasks, in

this proposed work, a Siamese neural network is used for processing. The processing of siamese neural network it has multiple identical forum. where identical means the same weight and parameters are taken for the configuration which helps to find the similarity [6]. There are some drawbacks in Siamese neural network, such as the process of training the model was light weight and straight forward and because of this, there is no guarantee of information or data loss because there is no specific inter relation between the processes. And to overcome this Siamese neural network, a Compare-Aggregate model was proposed, this proposed model have keen observation and note the similarity in between two sentences. This proposed model works on the dataset which is publically released by the data science engineer at Quora, where duplicate questions won't train for the models for detecting questions [7]. Research at Stanford University's Department of Computer Science [8] and the University provided us with a wealth of information. Where a model for extracting various features was controlled or administered, which superimposed the principle of fuzzy and vector distances of the texts as well[9]. Logic inference based on the Stanford Natural Language Inference Corpus has traditionally been the subject of semantic matching of sentences. Rocktaschel's paper [10] concentrated on LSTM-based word-by-word attention methods. The SemEval challenge [11], which focused on semantic similarity, was the first to include a task on question-question similarity.The task of detecting duplicate questions is part of the larger task of semantic text similarities, which has been the focus of the SemEval tests since 2012 [12]. Earlier attempts to detect sentence similarity relied on manually designed features such as word overlap, as well as standard machine learning algorithms such as SVM Classifier [13]. In a broad selection of NLP functions, neural network approaches have been the state-of-the-art. A Siamese neural network was suggested, which consists of two sub-networks that are linked at their outputs [14]. Research at Department of Computer Science, Stanford University, and New York University provided us with a wealth of knowledge. The concept of fuzzywuzzy and vector distances of the documents, as well as the concept of fuzzywuzzy, were evaluated as part of the model for extracting various features[15]. After the release of Quora's first dataset publically, the interest has begin and observed in academic for duplicate question pair detection. And the author wang proposed a bilateral LSTMs to this duplication problem by archievung state of art result and combine it with hand tuned cross question feature and this process has given a name as multi perspective matching. This work was attempting to apply LSTM encoding task and subsequently a hybrid LSTM model is created by encoding CNN [16]. In neural network model there are two types of deep learning framework which were proposed by NLSM. The first frame work was Siamese architecture and the other was Matching Aggregation. In Siamese framework matching decision was made solely base on two vector sentences and in matching aggregation framework features of two sentence were captured by CNN or LSTM, therefore it requires some more significance for improvement [17]. There are some limitation in matching aggregation approach, the first limitation is, that it explore the matching process word by word only or phrase by sentence, second was, it matches only in single direction i.e. P against Q but ignored the other i.e. reverse direction, Q against P. This drawback of matching aggregation was tackle by BiMPM model, where P against Q and Q against P, both approach were matched. The match was carried out on three layer and the three layers are identification, inference from natural language, and response sentence selection In all activities, this process has achieved state-of-the-art efficiency. [18].

III. APPROACH REGARDING PROPOSED METHODOLOGY

Question repetition is the most common problem that question and answer sites like Quora and Reddit, as well as Stack-overflow, face. Answers become segmented into various iterations of the same query due to question redundancy; to solve this, NLP is used in conjunction with machine learning. NLP is a branch of AI that allows machines to understand and deduce meaning from human language.NLP is a technique for allowing machines to understand and interpret human speech and text. Real-world applications such as sentiment analysis and stemming, parts-of-speech labeling, named entity recognition, and more are enabled by human-computer interaction. NLP is commonly used for machine translation and automatic question answering.

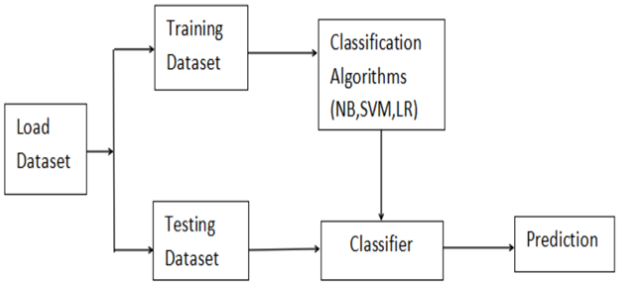


Figure 1 : Flowchart of Proposed Methodology

The above figure shows how to load a dataset before splitting it into training and testing datasets. Apply classification algorithms such as naive bayes, help vector machine, and logistic regression to the training dataset. Following the application of a classifier to a research dataset and a classification algorithm that predicts whether the questions are similar or not.

IV. MACHINE LEARNING ALGORITHMS

A. Naïve Bayes Classifier :

The Naive Bayes algorithm is a deterministic machine learning method for classification tasks. Naive bayes functions, like other supervised learning algorithms, are used to make a prediction on a target variable. The variance is that in naive bayes, features are assumed to be independent of one another and that there is no association between them. As it is a probabilistic model, the naive bayes algorithm can be coded up easily and the predictions made real quick .The algorithm is used to calculate the probability using the formula P(B|A):-

$$P(B|A) = \frac{P(A|B).P(B)}{P(A)} \quad (1)$$

Example :-

Fruits	**Like**
Strawberry	No
Orange	Yes
Watermelon	Yes
Watermelon	No
Strawberry	Yes
Orange	Yes
Strawberry	Yes

Frequency Table		
Fruits	Yes	No
Strawberry	2	1
Watermelon	2	0
Orange	1	1
Grand Total	5	2

Likelihood table			
Fruits	Yes	No	Probability
Watermelon	1	1	2/7 = 0.28
Strawberry	2	1	3/7 = 0.42
Orange	2	0	2/7 = 0.28

Therefore, The total probability of (Yes) = 5/7 = 0.71
And The Total probability of (No) = 2/7 = 0.28
Now, we will use the probability approach to solve it.
P(Yes | Strawberry) = P(Strawberry | Yes) * P(Yes) / P (Strawberry)
Now, P (Strawberry |Yes) = 2/5 = 0.4 , P(Strawberry) = 3/7 = 0.42 ,
P(Yes)= 5/7 = 0.7
Therefore, P (Yes | Strawberry) = 0.4 * 0.42 / 0.7 = 0.24, It is highest probability.

B. Logistic regression:

Logistic regression is a type of parametric classification model and is a supervised learning algorithm. It means that logistic regression models have a set number of parameters that are dependent on the number of input features and produce categorical predictions.The logistic regression uses a logistic function is also called a sigmoid function. The function outputs a value from 1 to 0 when an input „ ‟ is fed into it. The sigmoid function is as follows:

$$\emptyset(z) = \frac{1}{1+e^{-z}} \quad (2)$$

The sigmoid function outputs the probability of the occurrence of *z'*

Example :-

An explanation of how to construct a Logistic Regression model and use it to make predictions.

Parameters:

θ=1/12 ,b= -7

Equation:-

$$X = \frac{1}{12} * weight - 7$$

Consider the following scenario: we have two persons, one who weighs 140 kg and the other who weighs 70 kg. Let's see what happens when these numbers are entered into the model:

For person 1 ; 70 kg

$$x = \frac{1}{12} * 70 - 7 = -1.16$$

Probability of obese(x)

$$= \frac{1}{1+e^{-(-1.16)}} = \mathbf{0.76}$$

For Person 2 , 140 kg

$$x = \frac{1}{12} * 140 - 7 = 4.6$$

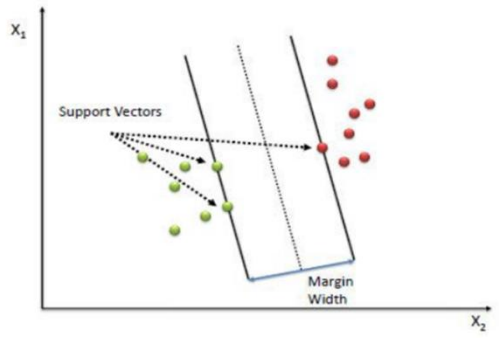
Probability of obese(x)

$$= \frac{1}{1+e^{-4.6}} = \mathbf{0.99}$$

As can be shown, the first person (70 kg) has a very low likelihood of becoming obese, while the second (140 kg) has a very high probability.

C. Support Vector Machine :

Support vector machine is a supervised learning model that requires data to be labeled and can be used on any form of data.
SVM is good at categorizing and predicting numbers.SVMs find a line (or hyperplane in order to maintain and increase than 2) among two classes of data points and therefore the distance on each side of that equation or hyperplane to another data point is the same on both sides .



The support vector machine algorithm's goal is to find a hyperplane in an N-dimensional space (N — the number of features) that distinguishes between data points. SVM can be used for regression as well as classification .The position of the vectors affects the position of the hyperplane, so SV points are very important in deciding the hyperplane. This hyperplane is also known as a margin maximizing hyperplane in technical terms.

V. CONCLUSION

This study uses natural language processing (NLP) to solve the problem of duplication in question and answer types by using machine learning to predict whether question pairs are identical or not. It selects highly superior features from questions and implements a low-cost architecture that allows it to identify duplicate questions and provide excellent answers to questions in Q&A.

VI. 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

Cite this Article

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