

**Literature Survey on 4th Year Project**

**Sudipta Halder**

**Smita Bandyopadhyay**

|  |  |
| --- | --- |
| **Team Number** | 1 |
| **Group Members** | Sudipta Halder (10400116059), Smita Bandyopadhyay (10400116078) |
| **Project Mentor** | Amit Kumar Das |
| **Project Title** | **Quora Question Pairs**  Can you identify question pairs that have the same intent? |

**Abstract:**

Quora is a question-and-answer site where questions are asked, answered, edited and organized by its community of users. Users can collaborate by editing questions and suggesting edits to answers that have been submitted by other users. This collaboration is displayed as a thread on a single question with a list of similar/related question so that users do not have to answer similar questions once again. Quora wanted to improve their similarity recognition system. So they released their dataset publicly so that a particular solution can be found out for increasing the already existing solution. The main aim of this work was to apply various Natural Language Processing (NLP) concepts for feature engineering from the given dataset and apply and compare some machine learning models such as K- Nearest neighbour, Decision Tree, Random Forest, Extra Trees, AdaBoost and XgBoost to predict the similarity.

**Introduction:**

Quora is a very popular website among internet users and a substantial amount of people from around the world visit Quora. It serves as a platform for individuals to interact with other individuals and ask and answer questions. Since the user base is quite high there are instances when there are multiple

questions which are based on the same topic and are redundant in nature. These redundant question decreases the efficiency and creates data that is repeating in the data servers. Since these questions have similar answers and users have to write similar content for each of these questions which is waste of time. User of the site have to view answers to the same redundant type of questions which results in wastage of time and reduces the efficiency. It would be beneficial if the redundant question can be reduced. This will help the users of the site in getting knowledge pertaining to a topic at one place rather than searching to same material over different questions which results in wastage of time and

resources. Machine Learning can be used for tackling this problem. We aim on devising some techniques that would help to judge the similarity between two questions in a more meaningful sense. Also we then aim to decide the similarity between a pair of question using various machine learning algorithm and compare the efficiency of different algorithm in tackling the problem. The dataset used is the globally released dataset by Quora. Site for Quora Question PairsDatabase source: <https://www.kaggle.com/c/quora-question-pairs/data>

**Related Works:**

The task of identifying duplicated questions can be viewed as an instance of the paraphrase identification problem, which is a well-studied NLP task that uses natural language sentence matching (NLSM) to determine whether two sentences are paraphrase or not [1].

With the renaissance of neural networks [3], two types of neural-based frameworks have been proposed for the task of paraphrase identification. The first framework is based on a “siamese” neural network consisting of two sub-networks joined at their outputs, where the sub-networks share the same weights at all levels and are responsible for extracting features from the input, and the output level computes the distance (cosine of the angle) between the two feature vectors generated by the sub-networks [2].

Although the siamese structure is lightweight and easy to train given that

the sub-networks share parameters, there is no interaction between the two sentences during the training process, which might cause information loss [1].

To cope with the limitations of the Siamese framework, a second framework named “compare-aggregate” is proposed [3], which captures the interaction between two sentences by performing a word-level matching and aggregating the matching results into a vector for the final classification. However, the “compare-aggregate” model fails to account for other types of granular matchings (e.g. phrase-by-sentence) and only performs matching in a single direction, which also neglects information in the sentence pairs [1].

**Discussion:**

In the current literature review, the machine learning models which are used mostly are logistic regression, logistic regression with hyperparameter tuning, linear svm, linear svm with hyperparameter tuning, XgBoost etc. and deep learning models based on “Siamese” structure and “compare-aggregate”.

We will try to apply more machine learning models like Extra Trees Classifier, AdaBoost and also some deep learning models like MaLSTM to improve the accuracy and precision and decrease the log loss.

**References:**

[1] Wang, Zhiguo,Wael Hamza, and Radu Florian. "Bilateral Multi-Perspective Matching for Natural Language Sentences." arXiv preprint arXiv:1702.03814 (2017).

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

[3] Wang, Shuohang, and Jing Jiang. "A Compare-Aggregate Model for Matching Text Sequences." arXiv preprint arXiv:1611.01747 (2016).

[4] <https://www.kaggle.com/anokas/data-analysis-xgboost-starter-0-35460-lb>

|  |  |
| --- | --- |
| **Mentor’s Endorsement**  **Signature and Date** |  |