Quora Question Duplication Problem

Uday Patel [1], Amol Dattu[2], Pritam Patil[3], Renuka Khot[4], Prof Sujit Tilak[5]

[1,2,3,4]Student at Pillai College of Engineering, Panvel 410206
[5]Prof. Sujit Tilak, Dept. of Computer Engineering, Pillai College of Engineering, Panvel, Maharashtra, India

---***---

Abstract - *The goal of our project is to present the results of systematic and comparative experimentation for automatic duplicate question detection, when different types of methods are applied to datasets of larger sizes.*

We will be applying different approaches to study the learning profiles of this task and also evaluate their merits. For example, the queries "What is the most populous state in the USA? " and "Which state in the United States has the most people?" should not exist separately on Quora because the intent behind both is identical. In this project, feature abstraction technique will be used, it is used to discriminate between the relevant and irrelevant parts of text. We will be testing logistic regression algorithm, Linear support vector machine algorithm and XGBoost algorithm and then we will evaluate these algorithms on the basis of their efficiency and accuracy. The main parameters are efficiency and accuracy. Keywords—

***Key Words*: NLP, Duplicate Questions, Vectorization, Tokenization, semantic analysis, machine learning, deep learning.**

1. INTRODUCTION

The fundamental of this project is to check whether the pair of questions are similar or not using the algorithms specified. Taking a dataset consisting of questions in paired format and pre-processing them for various operations performed by algorithms. The algorithms which would be used are Random forest, Logistic Regression, Linear Support Vector machine and gradient boost algorithm. The dataset will be in csv format (csv stands for comma separated values) which is CSV is a standard for storing tabular data in text format. Analyzing the outputs produced by algorithms. Graphs generated based on the statistics of the algorithms used as graphical representation is the most efficient representation to show the statistics of a dataset graphical representations using python only.

2. Literature Survey

In this there are some papers which were taken as reference for making this project. From those papers some conclusions were made and from which we made those decisions what features to take or what features to be used. Some papers were similar to the project to be made while some had the algorithms which were to be implemented, while some had the features which were to be implemented, the others had idea about pre-processing the data and how to clean the data before processing it through the algorithms.

The paper Can Duplicate Questions on Stack Overflow Benefit the Software Development Community? which is also a website similar to quora website, some information like fuzzywuzzy features, cosine similarity, vectorization with TFIDF, differentiating between root and duplicate questions (they recognise the duplicate questions and root questions from postID of the candidate's post if the postID is given then post is considered as root question else as duplicate question) were taken from that research paper. This paper also shows that whether the duplicate questions can be used as benefit instead of telling them as a disadvantage for the community. The answer is it can and can't because of its advantages and disadvantages. The advantage would be the duplicate information sometime can have some unique information

Example.

1. Mount Everest exists in Nepal and The mount Everest of height 8848m exists in Nepal. The meaning of the questions is same but they both have different information which might help the user to get to the answer without reading the actual answer and might come to some conclusion of their problems. But to detect this answer or like that it would have to read through the entire text. So this process would take time and complexity can increase which proves to be disadvantage for the Q&A websites which is a disadvantage.

2. Duplicate Question Detection in Stack Overflow: A Reproducibility Study this paper deals with the classification of the questions using SVM algorithm. SVM stands for Support Vector Machines. This algorithm functions mainly on creating a hyperplane between the two entities which are different but exists in a combined manner with other entities. This hyperplane divides the space into two parts. Here in this problem it will divide the analysis of the algorithms into two parts of similar and nonsimilar questions.

3. Exploring Deep Learning in Semantic Question Matching this paper deals with the methods of preprocessing the text like tokenisation, stemming, lowercasing, lemmatization, removing stopwords. The term vector distance which is a main a term which is used in machine learning and is helpful in recognising the string of text whether they are similar or not by calculating the vector distance between the pair of questions. Various vector distances: Cosine, Cityblock, Canberra, Euclidean and Minkowski distances are measured for all the question pairs. These distances are plotted against common words which gives the similar plot. These are some of the methods which we are going to use in deep learning in question matching or text matching. The overview of the TF-IDF method is also present which is also one of the method which will be used in this problem.

4.Learning Profiles in Duplicate Question Detection this paper gave the idea of using the SVM algorithm in text matching of the questions and checking whether they are duplicate or not. This also gives and overview of the Natural Language Processing and how to implement using SVM algorithm and xgboost algorithm, also some light has been shed on fuzzywuzzy features in the paper. Also we got the idea of performance degradation of the algorithms when algorithms are trained on variety of datasets and when tested on the narrow domain based dataset.

2.3 Summary of Related Work

The summary of methods used in literature is given in

Article	Publisher	Inference
Learning Profiles in Duplicate Question Detection	Chakaveh Saedi, Jo˜ao Rodrigues, .etal	SVM algorithm is used in DQD as well as in NLP because it is a good classification algorithm for text.
Exploring Deep Learning in Semantic Question Matching	Ashwin Dhakal, Arpan Poudel, Sagar, .etal	Data extraction, dataset preprocessing, fuzywuzzy parameters, Word Mover Distance and vector distance.
Duplicate Question Detection in Stack Overflow	Rodrigo F. G. Silva, .etal	Stack overflow is also a similar website to Quora. This website also uses SVM to classify the questions.
Can Duplicate Questions on Stack Overflow Benefit the Software Development Community?	Durham Abric, Oliver E. Clark,.etal	The preprocessing of text like removing of html tags, differentiating tokens.

Table1: Summary of literature survey

3. Proposed Work

To achieve this fundamental like to check whether the pair of questions are similar or not using the algorithms we would divide this problem into three parts:

Taking a dataset consisting of questions in paired format and pre-processing them for various operations performed by algorithms.

The algorithms which would be used are Logistic Regression, Linear Support Vector machine and gradient boost algorithm.

The dataset will be in csv format (csv stands for comma separated values) which is CSV is a standard for storing tabular data in text format. Analyzing the outputs produced by algorithms. To see whether which algorithm and/or its feature gives best accuracy and output in terms of algorithmic loss. Graphs generated based on the statistics of the algorithms used as graphical representation is the most efficient representation to show the statistics of a dataset graphical representations using python only.

3.1 System Architecture

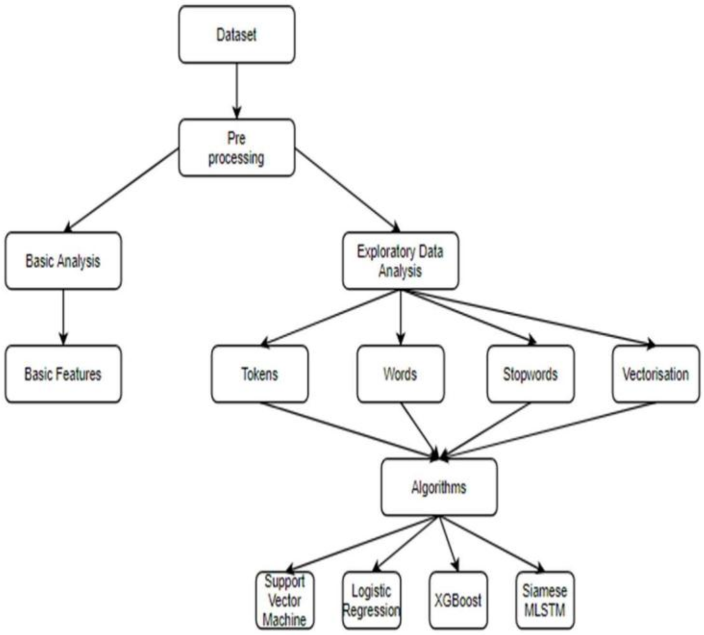


Fig 3.1.1 Proposed System Architecture

3.2 Implementation Details and results

Implementation of Our project are as follows: -

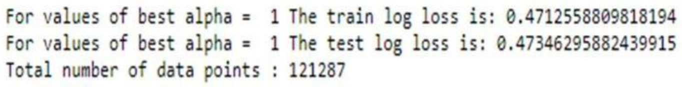
Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.

$$h\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

The Hypothesis of logistic regression

Result using logistic regression:

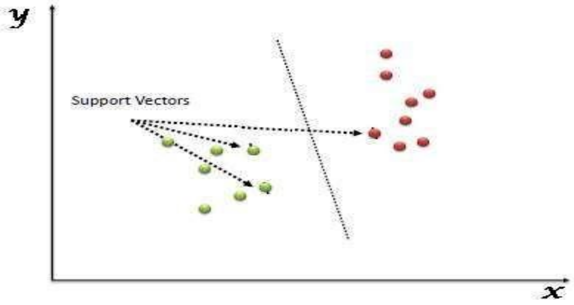


```
For values of best alpha =  1 The train log loss is: 0.4712558809818194
For values of best alpha =  1 The test log loss is: 0.47346295882439915
Total number of data points : 121287
```

Support Vector Machine

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in ndimensional space (where n is number of features you have) with the value of each feature being the value of a

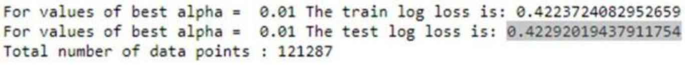
particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot).



3.2.1 Sample SVM graph [3]

Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line). We got accustomed to the process of segregating the two classes with a hyper-plane.

Result using SVM:



```
For values of best alpha =  0.01 The train log loss is: 0.4223724082952659
For values of best alpha =  0.01 The test log loss is: 0.42292019437911754
Total number of data points : 121287
```

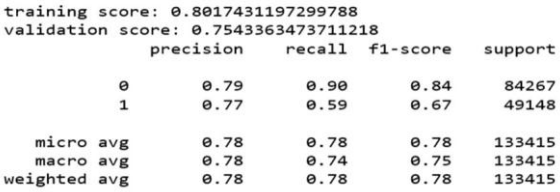
XGBoost:

Its name stands for eXtreme Gradient Boosting, it was developed by Tianqi Chen and now is part of a wider collection of open-source libraries developed by the Distributed Machine Learning Community (DMLC). XGBoost is a scalable and accurate implementation of gradient boosting machines and it has proven to push the limits of computing power for boosted trees algorithms as it was built and developed for the sole purpose of model performance and computational speed.

The implementation of XGBoost offers several advanced features for model tuning, computing environments and algorithm enhancement. The algorithm was developed to efficiently reduce computing time and allocate an optimal usage of memory resources. Important features of implementation include handling of missing values (Sparse Aware), Block Structure to support parallelization in tree construction and the ability to fit and boost on new data added to a trained model (Continued Training).

XGBoost is an ensemble learning method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models. The models that form the ensemble, also known as base learners, could be either from the same learning algorithm or different learning algorithms. Bagging and boosting are two widely used ensemble learners. Though these two techniques can be used with several statistical models, the most predominant usage has been with decision trees.

Results using XGBoost:

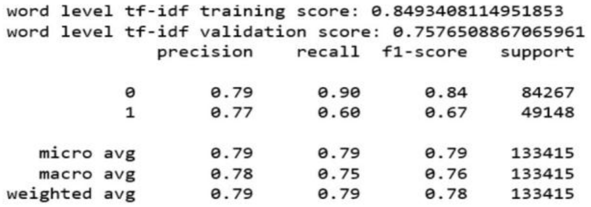


```
training score: 0.8017431197299788
validation score: 0.7543363473711218
             precision    recall  f1-score   support

          0       0.79      0.90      0.84     84267
          1       0.77      0.59      0.67     49148

  micro avg       0.78      0.78      0.78    133415
  macro avg       0.78      0.74      0.75    133415
weighted avg      0.78      0.78      0.78    133415
```

Result using XGBoost word level tfidf



```
word level tf-idf training score: 0.8493408114951853
word level tf-idf validation score: 0.7576508867065961
              precision    recall  f1-score   support

           0       0.79      0.90      0.84     84267
           1       0.77      0.60      0.67     49148

   micro avg       0.79      0.79      0.79    133415
   macro avg       0.78      0.75      0.76    133415
weighted avg       0.79      0.79      0.78    133415
```

Using Siamese Manhattan LSTM

MaLSTM (“Ma” for Manhattan distance), its architecture is depicted (diagram excludes the sentence preprocessing part).

Notice that since this is a Siamese network, it is easier to train because it shares weights on both sides.

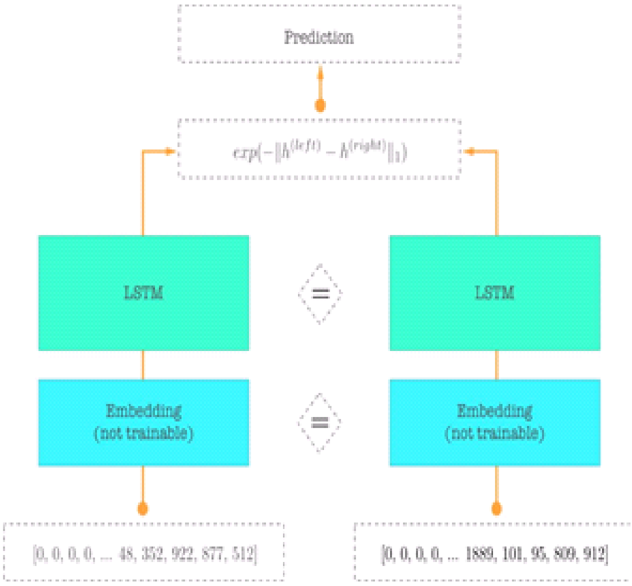
MaLSTM’s architecture — Similar colour means the weights are shared between the same-coloured elements

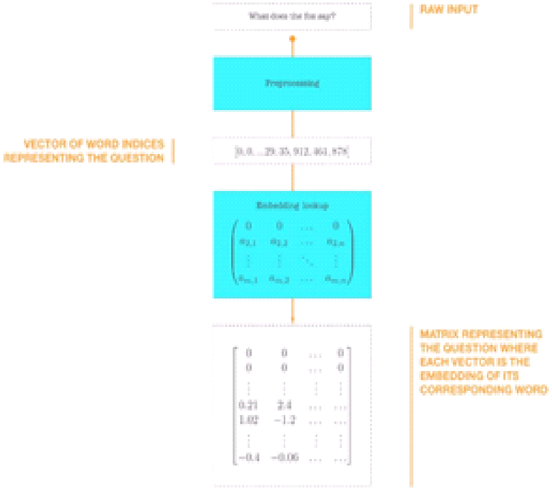
Siamese networks are networks that have two or more identical sub-networks in them.

Siamese networks seem to perform well on similarity tasks and have been used for tasks like sentence semantic similarity, recognizing forged signatures and many more.

In MaLSTM the identical sub-network is all the way from the embedding up to the last LSTM hidden state. Word embedding is a modern way to represent words in deep learning models. Essentially, it’s a method to give words semantic meaning in a vector representation. Inputs to the network are zero-padded sequences of word indices. These inputs are vectors of fixed length, where the first zeros are being ignored and the nonzeros are indices that uniquely identify words.

Those vectors are then fed into the embedding layer. This layer looks up the corresponding embedding for each word and encapsulates all them into a matrix. This matrix represents the given text as a series of embeddings. Google’s word2vec embedding, same as in the original paper. The process is depicted in figure.




What does the fox say?
RAW INPUT
Preprocessing
VECTOR OF WORD INDICES REPRESENTING THE QUESTION
[0, 0, ... 29, 35, 912, 461, 878]
Embedding lookup
MATRIX REPRESENTING THE QUESTION WHERE EACH VECTOR IS THE EMBEDDING OF ITS CORRESPONDING WORD

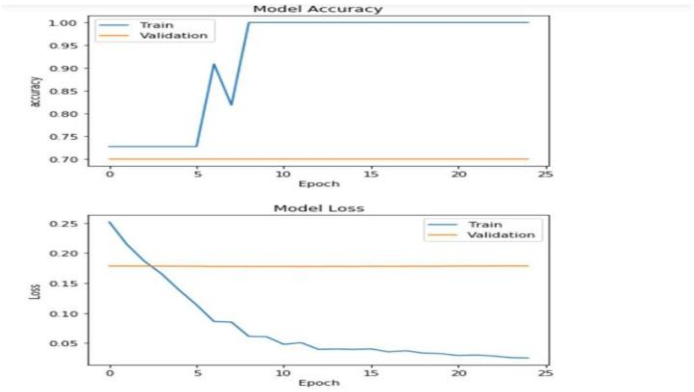
We have two embedded matrices that represent a candidate of two similar questions. Then we feed them into the LSTM (practically, there is only one) and the final state of the LSTM for each question is a 50-dimensional vector. It is trained to capture the semantic meaning of the question. This is represented by letter h. By now we have the two vectors that hold the semantic meaning of each question.

We put them through the defined similarity function (below)

$$exp(-\|h^{(left)} - h^{(right)}\|_1)$$

MaLSTM similarity function

and since we have an exponent of a negative the output (the prediction in our case) will be between 0 and 1.



MaLSTM is doing OK, getting an 82.5% accuracy rate on the validation data.

3.3.1 Sample Dataset Used:

An experiment is conducted in order to identify the input/output behavior of the system. Identify inputs. Specify the sample inputs that would be used in the experiments. The sample dataset used in the experiment are identified and given in Table below:

ID	QI D1	QID 2	Question1	Question2	Is_duplicate?
0	1	2	What are steps to invest in share market in india?	What is the step by step guide to invest in share market?	0
1	3	4	Is Kohinoor diamond real?	Will Kohinoor diamond will ever be recovered?	0

Table 3.2.2: Sample Dataset Used for Experiment

Conclusion

From the above executed result, we can conclude that the XGBoost algorithm performs and gives the best accuracy of 0.8. The major focus was on the semantic analysis of the sentences. The MALSTM gave an accuracy of 82% which is almost similar to XGBoost. The linear regression took much time to execute and SVM gave poor accuracy score as compared to XGBoost and MALSTM.

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. Sandeep M Joshi who gave me this wonderful opportunity to work on this synopsis.

I would like to thank our H.O.D Dr. Sharvari Govikar who gave me this wonderful opportunity to work on this synopsis.

I would like to thank our guide Prof. Mimi M Cherian whose devotion towards the topic and the kind of knowledge she has on this is exceptional and I'm very much lucky to have her as guide.

Thanks to my dearest friends Vaishnavee Deshmukh, Rahul P. 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 Ms. Harshada Chaudhari 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.

REFERENCES

[1]. Chakaveh Saedi, Jo˜ao Rodrigues, Jo˜ao Silva, Ant´onio Branco, Vladislav Maraev Learning Profiles in Duplicate Question Detection:, 2017

[2]. Ashwin Dhakal, Arpan Poudel, Sagar Pandey, Sagar aire, Hari Prasad Baral ,Exploring Deep Learning in Semantic Question Matching:, 2018

3]. Rodrigo F. G. Silva, Klérisson Paixão,Duplicate Question Detection in Stack Overflow, A Reproducibility Study: Marcelo de Almeida Maia, 2018

[4]. Durham Abric, Oliver E. Clark, Matthew Caminiti, Keheliya Gallaba, and Shane McIntosh Can Duplicate Questions on Stack Overflow Benefit the Software Development Community? :, 2019

[5]. LEO BREIMAN:Random Forests Statistics Department, 2018

[6]. Prince Mahmud,Md. Sohel Rana, Kamrul Hasan Talukder ,An Efficient Hybrid Exact String Matching Algorithm to Minimize the Number of Attempts and Character Comparisons, 2018

[7]. Chakaveh Saedi, Jo˜ao Rodrigues, Jo˜ao Silva, Ant´onio Branco, Vladislav Maraev IEEE International Conference on IRI Learning Profiles in Duplicate Question Detection University of Lisbon.2017

[8]. Ashwin Dhakal 1 Arpan Poudel 2 Sagar Pandey 3 Sagar Gaire 4 Hari Prasad Baral Exploring Deep Learning in Semantic Question Matching ,2017

Biographies:

Author 1: Uday Patel. Computer Engineering student at Pillai College of Engineering Panvel

Author 2: Amol Dattu. Computer Engineering student at Pillai College of Engineering Panvel

Author 3: Pritam Patil. Computer Engineering student at Pillai College of Engineering Panvel