

# Information Retrieval

Quora Question Pair Similarity using LSTM Model Semester Project Report

# Course Instructor: Sir Rafi Section: E

Group Members: Ahmed Raza 19K-0134

# Mubeen Siddiqui 19K-1276

## Abstraction:

Quora is famous platform to get the questions answered. The questions and answers are primarily posted by users. However, with the platform being ever growing, there exists a problem of different users posting questions of similar intent, hence producing duplicate questions. Successfully detecting duplicate questions would help in guiding the users to the effective answers posted already, saving time and thus improving user experience. In this research, we use the Quora Question Pairs dataset available at Kaggle. First, three types of word embeddings, namely Google news vector embedding, FastText crawl embedding with 300 dimensions, and FastText crawl sub words embedding with 300 dimensions are implemented individually to vectorize all the questions and train the model. The final features used for prediction are blend of these three types of word embeddings. Then, Siamese MaLSTM (‘‘Ma’’ for Manhattan distance) Neural Network model is applied for prediction of duplicate questions in the dataset to classify questions in the binary (1 for duplicate and 0 for not duplicate).

## OBJECTIVES:

Objectives of the project include to determine

-To Classify similar questions together

-User gets the answer of previously asked questions instantly

-Experts don't have to answer same questions again

## INTRODUCTION:

Quora is a social media website where questions are posted by users and answered by experts who provide quality insights. Other users can cooperate by editing questions and suggesting more accurate answers to the submitted questions. According to statistics provided by the Director of Product Management at Quora on 17 September 2018, Quora receives 300 million unique visitors every month, which raises the problem of different users asking similar questions with same intent but in different words. Multiple questions with similar wording can cause readers to spend more time to find the best answer, and make writers answer multiple versions of the same question. Therefore, Quora has an important principle for having a single question thread for logically different questions. For example, questions like ‘‘How can I be a good photographer?’’ and ‘‘What should I do to be a great

photographer?’’ are identical because both have the same meaning and should be answered only once. Some questions, like ‘‘How old are you?’’ and ‘‘What is your age?’’ do not have same wording. However, the context remains the same. Therefore, such questions are also considered duplicate. It can be an overhead to have different pages for such questions. Thus, identifying the duplicate questions at Quora and merging them makes knowledge sharing more efficient and effective in many ways. This way, the seekers can get answers to all the questions on a single thread and writers do not need to write the same answer on different locations for the same question.

## METHODOLOGY:

1. **Word Embeddings**

The LSTM model only understands input in vector format, in order to achieve that, The input text is passed to embedding layer which converts the text into vector format which is then understood by the model, We used The Embeddings of 300 dimensions and maximum sequence length of 20, We used GoogleNewsVector, Fast Text crawl, and Fast Text crawl subwords

## GoogleNewsVector

Google provides pre-trained word embedding based on news corpus. This word embedding contains 3 million English words with 300 - *dimensions*, providing 3 billion wordvectors

## FastText

FastText is an efficient word representation learning library provided by the Facebook research team. It contains 2 million common crawl words with 300 - *dimensions*, providing 600 billion word-vectors. It is different from Google word embedding because it provides the n-gram characterlevel representation of words

1. **FastText Subwords**

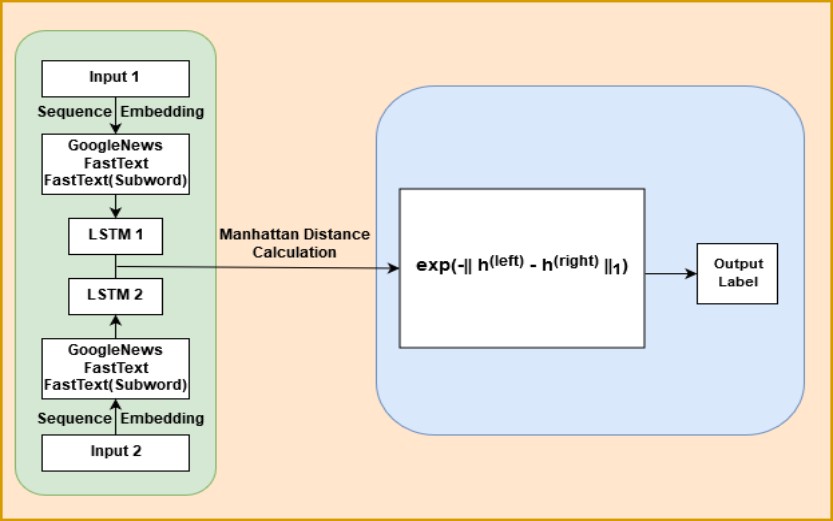
FastText Subword contains 2-million-word vectors trained with subword information on Common Crawl (600*B* tokens). Subword embedding provides us more details by converting each word into its sub words. If we want to get the subwords of word ‘where’ with *n* D 3 the resulting subwords will be, ‘whe’, ‘her’, and ‘ere’. At the end, it provides the dictionary of union of these subwords

1. **Siamese Network**

In a Siamese Network two or more inputs are processed at a time, and their result is merged, they are used to compare the similarity between two input vectors, both input vector here contains same parameters and weights. Siamese Network is a binary classification model, which after learning the pattern tells if the input belongs to the same class or not and thus means duplicate Advantage of Siamese includes more robust to class imbalance and learning from semantic similarity.

## LSTM

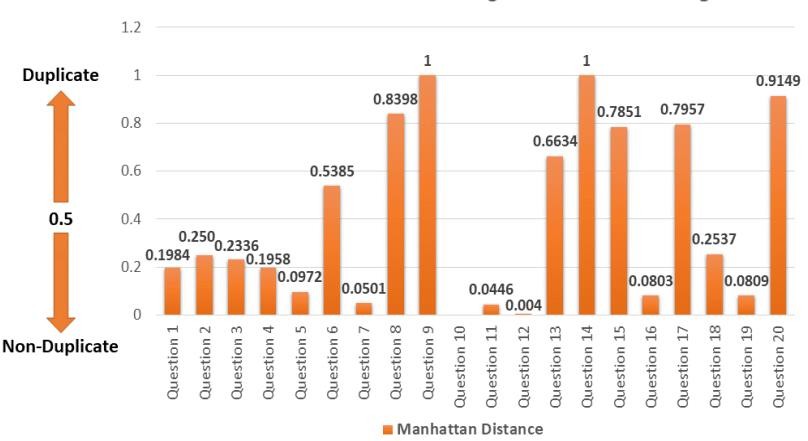
Long Short-term memory (LSTM) is an Artificial Neural network, they are well suited to classifying, processing, making predictions by using multiple inside layer, it consists or 4 components output gate, cell memory block, input gate, forget gate, input to LSTM is in the form of input vector, the hidden gate is sequentially updated between the gates.the update steps mainly rely on cell memory block. These 4 gates combine to make the predictions in LSTM model.



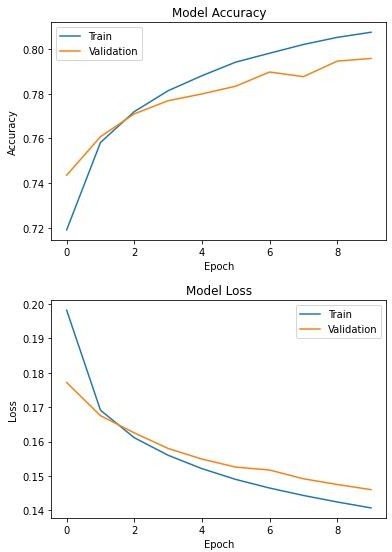
## RESULTS:

In the final set of experiments, the Siamese LSTM model is first trained on each word embedding (GoogleNewsVector, FastText and FastText subword) individually and later we use the blend of these trained models’ prediction for final prediction. The models are trained on 303K number of samples and are tested on 100k instances The training is performed on Google COLAB using GPU The training takes

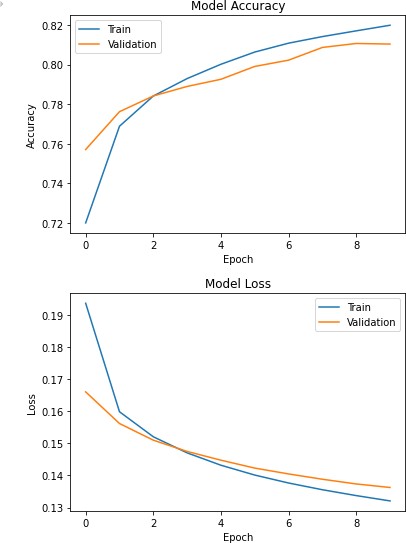
3.5 hours to run epochs on ‘Quora Question Pair Dataset’ on each embedding and to show the classification results. By training Siamese MaLSTM with Google News Vector, FastText Crawl and FastText Crawl Subword, the model achieved 81*:*77%, 82*:*77% and 82*:*57% accuracy, respectively. It can be observed that upon combining the predicted results of all three of these approaches with 33% of Google News Vector, 33% of FastText Crawl and 34% of FastText Crawl subwords, the obtained accuracy is 91*:*14% which is much higher than other state of the arts models



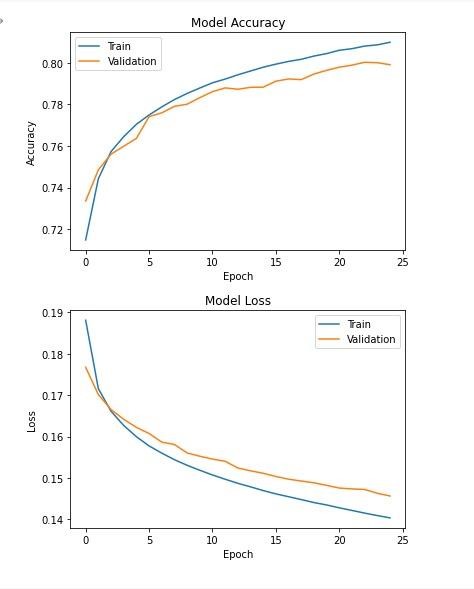
## Google News Vector Accuracy



**Fast Text Accuracy**



**Fast Text Subwords Accuracy**



**CONCLUSION:**

This work proposed a model that identifies duplicate question pairs by combining the three word embedding (i.e., Google News Vector, FastText Crawl, and FastText Crawl Subword) feature extraction techniques which results in a much better accuracy as compared to these embeddings individually.

Furthermore, this work proposed a novel Siamese MaLSTM model which accounts the Manhattan distance to determine the semantic similarity among the questions with 95% accuracy which is way better then state-of-the-art works. Upon closely looking at the manhattan values, in blend of different word embedding predictions, the manhattan score classifies the question pairs in more accurate way than any other embedding. The duplicate question score is very close to 1 while the non-duplicate pair values are much closer to zero. This determines the correctness and exactness of our proposed technique.

**REFERENCES:**

1. S. AbdelRahman and C. Blake, ‘‘A rule-based human interpretation system for semantic textual similarity task,’’ in *Proc. 1st Joint Conf. Lexical Comput. Semantics, 6th Int. Workshop Semantic Eval.*, vol. 2, 2012,

pp. 536–542

1. Z. Wang, W. Hamza, and R. Florian, ‘‘Bilateral multi-perspective matching for natural language sentences,’’ *CoRR*, vol. abs/1702.03814, pp. 1–7, Jul. 2017.
2. D. Tang, B. Qin, and T. Liu, ‘‘Document modeling with gated recurrent neural network for sentiment classification,’’ in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, Sep. 2015,

pp. 1422–1432

1. M. Tsubaki, K. Duh, M. Shimbo, and Y. Matsumoto, ‘‘Non-linear similarity learning for compositionality,’’ in *Proc. 13th AAAI Conf. Artif. Intell.*,

2016, pp. 2828–2834

1. B. Rychalska, K. Pakulska, K. Chodorowska, W. Walczak, and P. Andruszkiewicz, ‘‘Samsung Poland NLP team at SemEval-2016 task 1: Necessity for diversity; combining recursive autoencoders, WordNet and ensemble methods to measure semantic similarity,’’ in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, San Diego, CA, USA, 2016, pp. 602–608.
2. R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, ‘‘Skip-thought vectors,’’ *CoRR*, vol. abs/1506.06726, pp. 1–11, Jun. 2015
3. Y. Kim, ‘‘Convolutional neural networks for sentence classification,’’ in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha,

Qatar, 2014, pp. 1746–1751

1. R. Johnson and T. Zhang, ‘‘Supervised and semi-supervised text categorization using LSTM for region embeddings,’’ in *Proc. ICML*, 2016, pp. 1–9.
2. C. Zhou, C. Sun, Z. Liu, and F. Lau, ‘‘A C-LSTM neural network for text classification,’’ Nov. 2015, *arXiv:1511.08630*. [Online]. Available: https://arxiv.org/abs/1511.08630
3. P. Sravanthi and B. Srinivasu, ‘‘Semantic similarity between sentences,’’ *Int. Res. J. Eng. Technol.*, vol. 4, no. 1, pp. 156–161, 2017.