Quora Pairs Assignment

Nikolaos Athanasopoulos, Zoltan Kunos, Lily Voge

April 2023

1 Introduction

To begin with, we have to underline the fact that we split the initial dataset in three subdatasets:

- Train (90%)
- Validation (5%)
- Test (5%)

It is also important to mention the fact that we created a function called preprocess text, that cleans the text by doing the following actions:

- Convert all text to lowercase so that the model does not treat uppercase and lowercase letters differently.
- Used the Natural Language Toolkit (nltk) library to tokenize the text. Tokenization is the process of splitting the text into individual words.
- Created a set of stopwords, which are common words that are often removed from text because they do not carry much meaning (e.g., "the", "and", "it").
- Applied a lemmatizer to each token in the token list. In general, lemmatizer is used to reduce words to their base form (e.g., "running" to "run").

We have created two models as per the assignment. In this document, we will discuss these two models and weigh out their (dis)advantages. Just to briefly mention that we used the following two models:

- Perceptron
- Logistic Regression

1.1 Simple Solution

In this section, we answer the following questions: What problems/limitations do you think the model has? What type of errors do you get? What type of features can you build to improve the basic naive solution?

Our first model was based on a perceptron model. A perceptron is a type of neural network model that can be used for binary classification tasks, where the output is either 0 or 1 which is the case for this task. The output is either the same question (1) or two different questions (0). The perceptron model is a simple and fast algorithm that can be used for linearly separable problems, but may not work well for more complex problems that require non-linear decision boundaries. As our problem is quite complex our accuracy will be lower compared to other models, but it will be fast to train. Some of the limitations of the simple solution could include the lack of complex features that capture the semantic meaning of the questions, the lack of fine-tuning the hyperparameters of the model, and the use of only basic pre-processing techniques.

In terms of errors, we can see that the model is better at predicting negative instances (questions that are not duplicate) compared to positive instances (questions that are duplicate). This can be seen from the relatively lower values of precision and recall for the positive class compared to the negative class.

We decided to implement a more complex model (Logistic Regression) to improve our performance.

1.2 Improved Version

For our improved version we used a Logistic Regression model. Logistic Regression is a commonly used supervised learning algorithm for binary classification problems, where the goal is to predict a binary output (e.g., true/false, yes/no). In the case of the Quora challenge, the task is to predict whether a pair of questions are duplicate or not. Therefore, it is a binary classification problem.

Logistic Regression models are particularly useful for binary classification tasks because they output a probability estimate between 0 and 1, representing the likelihood that a given instance belongs to the positive class (in this case, duplicate questions). This probability can then be thresholded to make a binary prediction.

Another advantage of Logistic Regression is that it is a relatively simple and interpretable model. The coefficients learned by the model can provide insight into which features are important for the classification task, which can be useful for understanding the problem domain. Last but not least, Logistic Regression can handle a large number of features, which is important in natural language processing tasks where the feature space can be very high-dimensional.

Based on the results, we can see that the Logistic Regression model has achieved better performance compared to the Perceptron model on the task of solving the Quora challenge.

However, there is still room for improvement in terms of improving the model's ability to predict positive instances (questions that are duplicates).

Overall, the improved Logistic Regression model provides a good baseline for future experimentation and optimization on this task.

Other models that we could use if we had more time in order to improve performance even more might be Random Forest, Gradient Boosting etc.