Question Or Not?

Question detection using artificial intelligence

Anfal Alatawi 1617134

Najwa Noorwali 1507697

Bayader Alsahafi 1606346

DEPARTMENT OF COMPUTER SCIENCE

KING ABDULAZIZ UNIVERSITY

Contents

1 Introduction

In machine learning, a new field is emerging with much potential, due to the advancements in computational power and the abundance of data, that field is the field of **Natural Language Processing (NLP)**.

NLP is crucial in order to automate tasks usually taken out by humans, as the higher goal of automation is to eventually replace entire humans by machines (essentially models). In order to replace humans without affecting their performance, the tasks being replaced have to be taken out exactly in the same manner as before replacement. That is, users should be able to communicate with the models the same way they communicated with humans.

One of the branches of NLP is **text classification**, taking a piece of text and categorizing it into predefined categories. Building text classification models is usually done using the approach of supervised machine learning, where a large corpus of labelled instances is used as training data, to train a model to predict the label on unseen instances.

1.1 Importance of the Problem

This project explores the application of text classification in the business sector, specifically, text classification in automated customer services. Many costumer service providers use chat-bots, that is due to the similarity and redundancy of the inquiries received. In this project, we explore the use of classifying customer inquiries into either sentences or questions, either to aid chat-bots, or human customer service representatives.

This classification is useful in the context of automated filtering of questions, or can go further with the use of automated replies to the detected questions, which can either aid the performance of chat-bots by flagging possible questions, or aid in performing initial filtering of inquiries, so that the chat-bots or humans receives questions only.

Moreover, the detection of questions in customer feedback can be also useful in measuring the performance of a customer service provider, participating in a company's KPIs. For example, it can be used to measure the percentage of questions answered, the average response time of customer service (to clients with questions), and much more statistical information, which is all based on first automatically detecting the existence of a question in customer feedback.

2 Proposed Model

The model built in this project takes a sentence and classifies it into either a questions, or not a question, as shown in Figure 1.

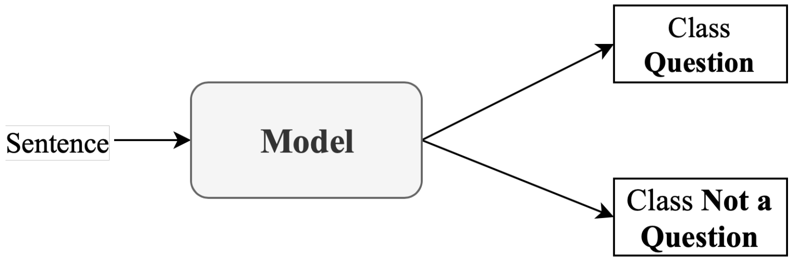

Figure 1: Overview of the machine learning model

3 Related Work

When reviewing the literature to find similar problems to the this project's, we have found that researches with the same idea were limited to detecting questions in audio files. Those researches may depend on features that are extracted from the sound it self rather than the textual script extracted from the audio file. Here, we only consider researches that look into text NLP related features.

In (Boakye, Favre, and Hakkani-Tür 2009) work, the authors worked on detecting English questions in meetings. The authors examined several features, which can be divided into three groups: features related to words and syntax, features related to the turn-taking nature of conversational speech and acoustic features related to pitch/intonation. For the Lexico-syntactic features, they used uni-grams and bi-grams to identify initial Wh-words, the second-person pronoun “you”, and word order. They also used part of speech tags and parse trees to get the syntax representation. For the classification, they used adaptive boosting, which produces a classification rule by combining a set of weak classifiers. In their work, the authors used one-level decision trees (stumps). They trained the classifiers for 1000 rounds and used f1 score as an optimizer. As for their results, they achieved an F1 score of 67.57% by only considering a subset of all the Lexico-syntactic features. This result outperformed the two other groups efficiency when considering each individually.

In (Blanchard et al. 2016), the authors investigate automatic detection of key teacher instructional strategies from recordings of teachers’ speeches. Their corpus contains transcripts of audio recordings of 37 class sessions, taught by 11 teachers. In their work, they developed 37 domain general features rather than word specific features, like n-grams or parse trees; because they sampled different teachers and classes, and the topics covered vary significantly; for that reason, a content heavy approach would likely overfit to specific topics. They considered the Naïve Bayes classifier using the WEKA machine learning toolbox. It was chosen based on preliminary experiments with several other standard classifiers (e.g., logistic regression, support vector machine, k-nearest neighbor, decision tree, random forest). Furthermore, they used a leave-one-teacher-out cross-validation technique to validate the model, in which the model is built on data from 10 teachers (training set) and validated on the held-out teacher (testing set). In their experiments, the authors trained the model on three feature types, NLP, timing, acoustic and a combination of the three. They also trained the model on predicting five instructional segments: (Question & Answer, Procedures and Directions, Supervised Seatwork, Small Group Work, and Lecture). By considering all the features, the model achieved 55% F1 score in detecting questions. It also has been found that the 10 most usefull features in detecting questions are all NLP features, and included features such as the number of occurrences of certain question words (e.g., “why”, “what”) and the number of proper nouns.

4 Methodology

In this project, we first collected the data, pre-processed it, extracted its features, then fed it into SVM model. Finally we reported the results of this experiment.

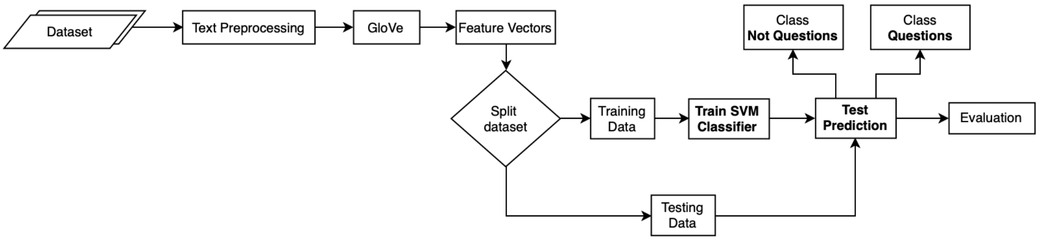

Figure 2: Methodology of developing question classifier

5 Data

The dataset in this project is unique, and has been developed it for the purposes of this project. It is constructed from the following resources:

1. **Quora:** The first source of the dataset is **Quora**. Quora is a question-and-answer website where questions are asked, answered, and edited by users. We obtained 8351 questions from this dataset. For the source of this dataset, go to the next URL: `https://github.com/vishaljain3991/cnn_topic_classification`.

2. **OPUS:** The second source of data used in this project is the Open Parallel Corpus collection (OPUS). OPUS is an open source databse of sentences and their translations. OPUS includes data from various resources. The collected dataset in this project contained over 10 million English sentences initially, from the Wikipedia website.

5.1 Data Preprocessing

The data-set went through several steps in order to clean it and to transform it into an understandable format, to finally obtain a balanced data-set. However, since Quora data-set was already clean, these steps were done only on the OPUS data-set.

- **Removal of short and long sentences:** Sentences with less than 3 words or more than 50 words were removed.
- **Removal of questions:** Since OPUS data-set contained both questions and sentences, all questions were removed (To decide whether a sentence is a question or not, the sentence was checked if the last character was a question mark '?' or not).
- **Conversion to lower case.**
- **Removal of punctuation.**

6 Model Design

6.1 Feature Extraction

In order to represent text in natural language to a model, the text has to go through a step that models the text in values the machine learning model is able to understand: vectors. The mapping from textual data to real valued vectors is called **feature extraction**.

There are many methods that facilitate representing instances of text as vectors. One method is the **Bag Of Words (BOW)** method, this algorithm makes a list of unique words in the text data-set called vocabulary. Then represents each sentence as a vector with each word represented as 1 for present and 0 for absent from the vocabulary. A major downside of this method is that the length of the vector will equal the number of unique words in the data-set, which is infeasible. Another disadvantage of BOW is that it discards word order thereby ignoring the context and in turn meaning of words in the sentence.

Word Embeddings is another method of feature extraction from text. It is a representation of text where words that have the same meaning have a similar representation. In other words, it represents words in a coordinate system where related words, based on a data-set of relationships, are placed closer together.

In this project, Word Embeddings were chosen to perform feature extraction of the instances. The reason behind that is that word embeddings conserve the meaning of the words, rather than a frequency vectorization of words, such as BOW.

The Global Vectors for Word Representation, or GloVe, algorithm is an implementation of Word Embeddings. GloVe constructs an explicit word-context matrix using statistics across the whole text data-set that it has been trained on. The result is a model that maps text to word vectors.

In this project, a pre-trained GloVe model was used. The model chosen was the "GloVe 300-Dimensional Word Vectors Trained on Wikipedia and Gigaword 5 Data" (Pennington, Socher, and Manning 2014). This choice is justified after inspecting a number of different models with different dimensions, such as the GloVe model trained on Twitter data with a dimension of

25. The aforementioned model did not fit out problem, because the domain of the text used in this project is different from the social networking (Twitter) domain. Also, the Wikipedia GloVe model worked best on this project's data, since part of the data was extrected from wikipedia.

6.2 Model Selection & Hyper parameters

In this project, we chose a Support Vector Classifier to train. This choice is justified by the good performance of SVCs on classification problems. The hyper parameters chosen for the model were the default parameters.

7 Results

We evaluated the performance of our model based on 4 key criteria: accuracy, precision, recall, and F1, which were calculated from the two classes.

Table 1: Results of SVM Model

Measure	**SVM**
Training Accuracy	0.897
Testing Accuracy	0.804
Precision	0.87
Recall	0.76
F1	0.82

Based on the result, which is shown in Table 1, the scores show that the model preforms well.

8 Conclusion

In this report, we presented a project that classifies sentences to a question, or not. We constructed our data-set from two resources, Quora and Wikipedia websites, which is comprised of 16,698 sentences. The data-set contains exactly 8351 questions and 8347 sentences, in order to make sure it is balanced and does not affect the model performance. In order to classify the data-set, we performed feature extraction on our text data using a GloVe pre-trained model. Then, we used SVM classifier model on the extracted features, which produced high performance scores.

For future work, there are many deep learning models that we would like to explore such as convolutional neural networks, since they tend to produce better results on text models. And recurrent neural networks, since they consider the sequence of inputs and that might affect results of text models positively. As well as the addition of different feature extraction mechanisms. Moreover, we want to improve our data-set by different techniques such as collecting more data and removing stop words that don't convey any notion of meaning, to improve both the efficiency and accuracy.

References

Blanchard, Nathaniel et al. (Sept. 2016). “Identifying Teacher Questions Using Automatic Speech Recognition in Classrooms”. In: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Los Angeles: Association for Computational Linguistics, pp. 191–201. DOI: 10.18653/v1/W16-3623. URL: https://www.aclweb.org/anthology/W16-3623.

Boakye, Kofi, Benoît Favre, and Dilek Z. Hakkani-Tür (2009). “Any questions? Automatic question detection in meetings”. In: *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 485–489.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.