Deriving from Pre-trained Word Embedding

Akanksha Mishra, Sukomal Pal
Indian Institute of Technology (BHU), Varanasi - 221005
India
{akanksham.rs.cse17, spal}@itbhu.ac.in

ABSTRACT: *Question answering and classification is analysed from different angles by researchers. Organising and classifying questions in the query formulation is addressed using different techniques where the word embedding is one. IN this paper we present our work related to using pre trained GloVe word embedding. Besides, we processed the results with different other embedding.*

Keywords: Quora Insincere Question, Bidirectional Long Short Term Memory, GloVe Embedding

Received: 3 April 2020, Revised 6 June 2020, Accepted 20 June 2020

DOI: 10.6025/pms/2020/9/2/52-56

1. Introduction

In today's era, all the major websites on the internet are facing the challenge of providing appropriate content on their space. They are relying on the user of the website for marking unsafe content so that they can remove those contents and hence make the content of their website safe. Quora is one of the widely used websites on the internet with the user base of 300 million monthly users1. We can ask questions on any topic that affect us or the world; moreover, if we want to have an opinion of experts about any real-world incidents or how others would have tackled any particular situation. We can get interesting answers about all the stuff that we care about.

2. Task Definition

The people are frequently using web forums like StackOverflow, Quora, and many more for getting answers to their information-seeking questions. However, with the vast user base, some people tend to ask questions with objectionable content. Sometimes, some questions are posted to target some specific group or spread hate speech. It is challenging for the human moderators to filter out insincere questions manually due to the vast number of questions. The task is about the identification of non-information seeking questions with the varying characteristics into one of the six categories:-

[1] https://expandedramblings.com/index.php/quora-statistics/

- Rhetorical questions
- Sexual Content
- Hate Speech
- Hypothetical
- Other
- Not an Insincere question

3. Data and Resources

Task organizers provided participants with the question id and label assigned to each question for 900 training instances. Due to dataset sharing constraints of kaggle, they could not provide the question text; hence the participants were supposed to extract the question text from the competition organized on the kaggle[2] website using question id. We extracted the question text from the data section of the competition on kaggle and formed the training set. Each instance of the training set consists of the question id, question text, and label assigned to it. A sample of the training set is given below in table 1.

Figure 1 shows the distribution of training instances among different categories. It can be seen from figure 1 that there is an imbalance of training instances among different categories.

qid	qtext	label
03c5993d2c4898c57e49	Is it OK to be white?	1
05981e7a85209fe81046	Why have 50% of women in Finland been raped?	2
0599b2f6bf4ce21d0dea	Why Mohajirs hate Punjabi, Sindhi, Baloch, Pathan and pretty much everyone?	3
05e945cc9bf993ae9a5d	Can I get my desired gazetted officer government job by the help of black magic?	4
0485c68c75ca7ce02272	Can bleach cure autism?	5
034dadb82db0211e2ca7	How do I sign up for Quora account?	0

Table 1. Sample of training set

4. System Description

This section discusses the implementation of our approach as shown in Algorithm 1. Firstly, we perform preprocessing on the question text, followed by feature extraction using pre-trained embedding and training on the bidirectional Long Short Term Memory model.

Data Preprocessing: We perform preprocessing by removing punctuation, ‘#’,‘@’ and ‘https’ symbols. We keep the stop words and hashtag words to get a better understanding of the context during training. Also, we removed numerals as with different contexts they play different roles in question texts; hence, we feel it is better to remove them. We also lowercased all the texts of the question.

Feature Representation: We use pre-trained GloVe [3] word embedding to represent words in the form of vectors. Different versions of glove pre-trained embedding exist; however, we use embedding trained of dimension 300 on common crawl

[2] https://www.kaggle.com/c/quora-insincere-questions-classification

using 840B tokens and 2.2M vocabulary[3]. We generated random embedding of dimension 300 for out of vocabulary words.

Model Description: We determine the maximum length of the sentence from question texts of all training instances. We perform padding of the question text in each training instance whose sentence length is less than the maximum length of the question texts. We use bidirectional Long Short Term Memory [1, 4] layer followed by dropout layer to avoid overfitting. We added a fully connected dense layer at the end.

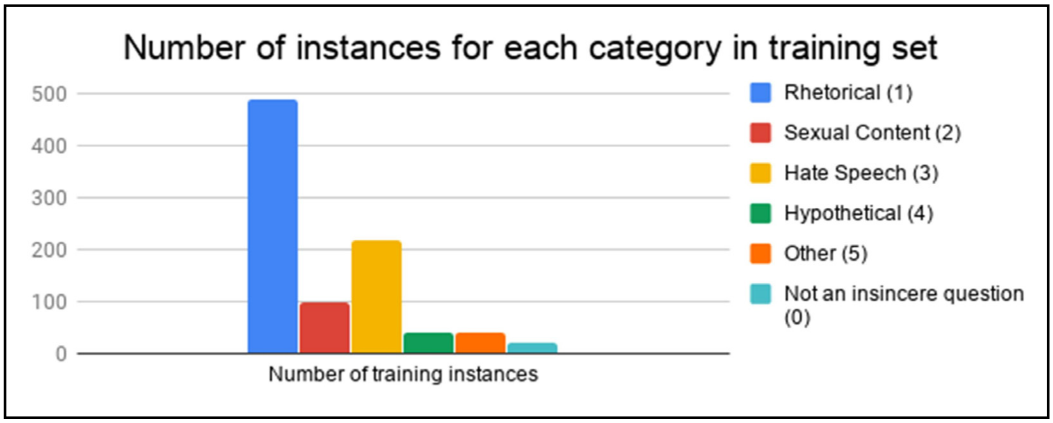

Figure 1. Number of instances in each category in training set

```
Algorithm 1: System Description
Data: Training and Test instances
Result: Predict true labels of all test instances
1 Generate a list of tokens of each instance of the train and test set
2 Remove punctuation, '#', '@' and 'https' symbols from the generated tokens
3 Remove digits and convert all tokens to lowercase from the generated tokens
4 Obtain a maximum length of an instance from the train and test data
5 Pad all instances whose length is less than the maximum length obtained in Step 5
6 Generate a list of vocabulary of the dataset
7 Extract embedding of dimension 300 using GloVe pre-trained embedding for each word of the vocabulary
8 If any word of the vocabulary is not present in the GloVe pre-trained embedding, then generate random embedding of dimension 300
9 Represent all labels using one-hot encoding
10 Build a sequential model consists of Bidirectional LSTM, dropout and dense layer
```

5. Results

In this section, we will discuss the experimental settings, results obtained with the model, and further analysis of the results.

[3] https://nlp.stanford.edu/projects/glove/

Experimental Settings: We use Keras[4] neural network library for training our model which uses Tensorflow as backend. The model is trained for ten epochs with a batch size of 32. We use a validation split of 0.3 to analyze the overfitting using validation loss that may occur during training. Table 2 list out the values for parameters and hyperparameters used for training the model.

Parameters / Hyper parameters	**Values**
BiLSTM Activation Function	tanh
Recurrent Dropout	0.2
Dropout	0.3
Dense Activation Function	softmax
Optimizer	adam
Loss	Categorical Cross Entropy

Table 2. Parameters and Hyper parameters

Results: We obtained an accuracy of 64.35% on test set which consists of 101 instances. The accuracy was calculated and shared by the task organizers.

Analysis: The task organizers shared the true labels of the test instances hence we used different other word embeddings for further analysis. The accuracy obtained using different embeddings with Bidirectional LSTM model is listed in the table 3. Model M2 was submitted for the evaluation. We trained vocabulary of the train set using word2vec [2] continuous bag of words architecture in model M1; however, we used pre-trained embedding paragram [5] in model M3. In the case of GloVe, it is observed that there are 60 out of vocabulary words; however, only 40 words of the vocabulary were not present in paragram embedding. All three embeddings represent each word of dimension 300.

Model	**Embedding**	**#OOV words**	**Model**	**Accuracy**
M1	Word2Vec	-	BiLSTM	65.34%
M2	GloVe	60	BiLSTM	64.35%
M3	Paragram	40	BiLSTM	63.36%

Table 3. Accuracy with different word embeddings

6. Conclusion

We used a bidirectional Long short term memory model for classification of insincere questions on Quora. The system can be used for the automatic classification of insincere questions. We obtained an accuracy of 64.35% on the test set. We can incorporate linguistic features to improve the system.

References

[1] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*. 9 (8) 1735–1780 (November). https://doi.org/10.1162/neco.1997.9.8.1735, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[4] https://keras.io

[2] Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space.

[3] Pennington, J., Socher, R., Manning, C. D. (2014). Glove: Global vectors for word representation. *In*: Empirical Methods in Natural Language Processing (EMNLP). p. 1532–1543, http://www.aclweb.org/anthology/D14-1162

[4] Schuster, M., Paliwal, K. (1997). Bidirectional recurrent neural networks. *Trans. Sig. Proc*. 45 (11) 2673–2681 (November). https://doi.org/10.1109/78.650093, http://dx.doi.org/10.1109/78.650093

[5] Wieting, J., Bansal, M., Gimpel, K., Livescu, K. (2015). From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics* 3, 345–358.