Classification of Insincere Questions

Sicheng Li (sl4814) | Judge Madan (tm3004) | Di Mu (dm3686) | Yanyun Chen (yc4014)

Abstract

We analyze the effectiveness of various embeddings and machine learning models on the classification of insincere questions on Kaggle's "Insincere Questions Classification" dataset[1].

Introduction

Quora is a question and answer website where people go to find information. What differs it from other search engines is that it focuses on high-quality questions and answers. It is not surprising that Quora considers insincere questions (non-neutral tone, inflammatory, false statements, or contains sexual content) as potential threats to its image as a safe place to share knowledge and potentially reduces engagement on the platform. In this project, our goal is to determine if a Quora question is an insincere question or not. This leads us to the problem we are trying to solve:

How do we best classify questions as sincere or insincere?

We decided to explore the use of the following models:

1. A logistic regression model with a bag-of-words embedding.
2. A Long short-term memory (LSTM) model, based on Recurrent Neural Networks (RNNs).
3. A transformer, using the Word2Vec embedding.
4. The Bidirectional Encoder Representations from Transformers (BERT) model.

The logistic regression model acts as a simple baseline model, with a bag of words embedding. In a bag-of-words embedding, we take the frequency of each word in the corpus and encode it into a vector. In this representation, the relative location of words are lost.

The LSTM model is an RNN, which allows the connection of previous words to the current word, forming contextual information. LSTMs are capable of learning long-term dependencies. Due to this, we chose to embed text by assigning each word in the corpus an integer, and converting sentences to a 100-length vector of integers, padded with 0s. The vector is then passed into an embedding layer.

The simple transformer model here is implemented as a transformer layer after the two embedding blocks: One for token

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

embedding using Word2Vec, and the other for index/position embedding. It then feeds to several dense layers.

The BERT model is a state of the art language model, which is focused on the use of bidirectional training of the transformer attention model. Compared to previous transformer approaches, which process text in one direction (e.g. left to right), the bidirectional model can pick up on context more clearly.

Data Analysis

The data only contains 6.2% insincere samples, making it an extremely imbalanced dataset. Normally speaking, F1 score, a measurement that considers both recall and precision, is an appropriate measure under such circumstances. However, considering the large scale and imbalance of our data, we have decided to take a balanced subset of 50,000 positive samples and 50,000 negative samples for performance reasons. Thus accuracy, precision and recall can all be used as metrics on this subset.

The data set provided by Quora contains 1.3 million rows and 3 features, namely:

- qid: unique question identifier
- question_text: Quora question text
- target: a question labeled "insincere" has a value of 1, otherwise 0

There are no missing values and no duplicated rows.

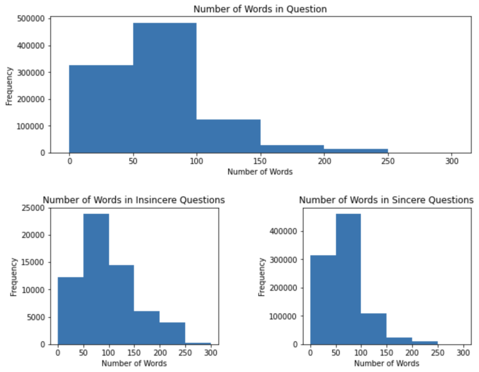

Fig I. Length of sentences for sincere/insincere text.

On initial data exploration, we found that sentences are generally longer for insincere questions with the average length being 98.06, compared to sincere questions with an average length of 68.87. Moreover, sentences with censored words have a higher proportion of (around 60%) insincere words. These insights could be later used as new features for our models. We also found that the distribution of labels for text with censorship (which suggests the use of negative words), shows a higher proportion of insincere speech at 57.9% compared to 6.2% baseline.

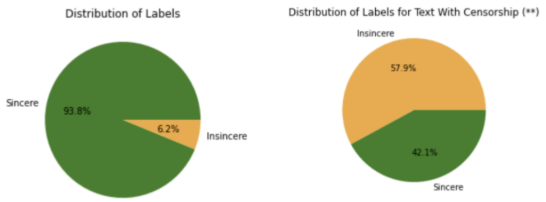

Fig II. Distribution of labels for text with censorship.

Initial data exploration also suggested that insincere questions usually contain profanity, sexual language, or other language that incites violence or racism[2]. On the other hand, sincere questions aim to

[2] https://docs.google.com/presentation/d/1mEpdI2GyeglVNfP6Hd3r8vy3swv8JVJYH2Zko6mE1x8/edit#slide=id.g100a519429e_1_135

seek a helpful answer, and are based on reality.

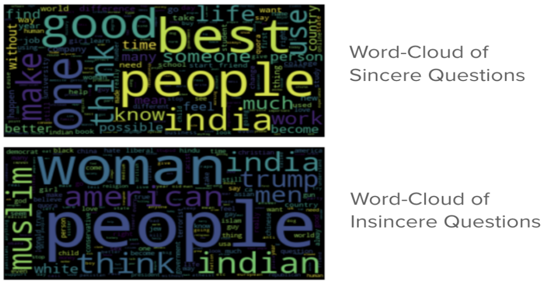

Fig III. Word cloud of data.

Word clouds created from the two classes of questions also revealed some insights: insincere questions contain divisive language about particular groups of people (“muslim”, “woman”, “american”, “men”), while sincere questions contain more neutral words.

Methodologies & Metrics

For each model, we follow a sequence of steps for evaluation, including:

1. Data cleaning/text preprocessing
2. Finding a suitable text representation or embedding
3. Training the model with an 60/20/20 train/validation/test split.
4. If applicable (for models which are trained in epochs), using early stopping to identify the highest validation accuracy.
5. Measure the accuracy of the model on the test dataset.

For data preprocessing, we split our sentence into a list of words, while removing common words and punctuation as they are redundant. Each word is reduced into its base form to extract the base meaning of each sentence according to the text processing pipeline demonstrated below.

Fig IV. Data Processing Pipeline

Input Data:
“ How did Quebec nationalists see their province as a nation in the 1960s? ”

Strip: Extra whitespace has no effect on meaning.
“How did Quebec nationalists see their province as a nation in the 1960s?”

Lowercase: Words should be stored in a common format.
“how did quebec nationalists see their province as a nation in the 1960s?”

Tokenize: Sentences need to be split into words for vectorization.
[“how”, “did”, “quebec”, “nationalists”, “see”, “their”, “province”, “as”, “a”, “nation”, “in”, “the”, “1960s?”]

Remove Punctuation: Helps model consider only key features.
[“how”, “did”, “quebec”, “nationalists”, “see”, “their”, “province”, “as”, “a”, “nation”, “in”, “the”, “1960s”]

Remove Stop Words: Helps model consider only key features.
[“quebec”, “nationalists”, “see”, “province”, “nation”, “1960s”]

Lemmatize: Combine words with the same meaning.
[“quebec”, “nationalist”, “see”, “province”, “nation”, “1960s”]

Results

Following the methodologies described, we achieve the following results. Due to the balanced subset used, accuracy is sufficient to compare the models. As aforementioned, models were chosen based on the highest validation accuracy, which was then evaluated on the test set.

Model	Accuracy (%)
Logistic Regression[3]	87%
LSTM	86%
Transformer	85%
BERT-LSTM	83%

[3] No improvement was found after adding indicator of censored words as a feature.

Evaluation

No significant difference in accuracy is found with each of the techniques, with accuracy between 83-87%, with simpler representations performing better.

While the other models use embeddings that provide semantic context, multidimensional logistic regression with the bag of words embedding performed the best, at 87% accuracy. This suggests that word placement, semantics and context within a sentence is not integral in deducing whether a question is sincere or not. Rather, the frequency of words used within a sentence provides adequate information to complete the task, as noted by similar accuracy across the board.

This aligns with our exploratory data analysis, as we noted that insincere questions contained more divisive language, while sincere questions contain more neutral language. This could mean that grammar and context could have minimal effect on the target class, rather only the vocabulary used. While we can not solely attribute the performance due to these reasons, we must consider the possible ramifications on accuracy of the harder ability to tune hyperparameters for larger and more complex models.

Due to the high computational complexity of fine-tuning BERT on the large dataset, we used the pre-trained bert-base-uncased model as embedding. This might lead to an accuracy drop on the BERT-LSTM without fine-tuning the model. Besides, BERT is known to perform well on more complicated tasks like summarization and dialog systems. In this task, where the boundary between sincere and insincere is usually decided by a certain word, static embeddings such as BoW, TF-IDF and Word2Vec can outperform BERT.