Quora Insincere Questions Classification

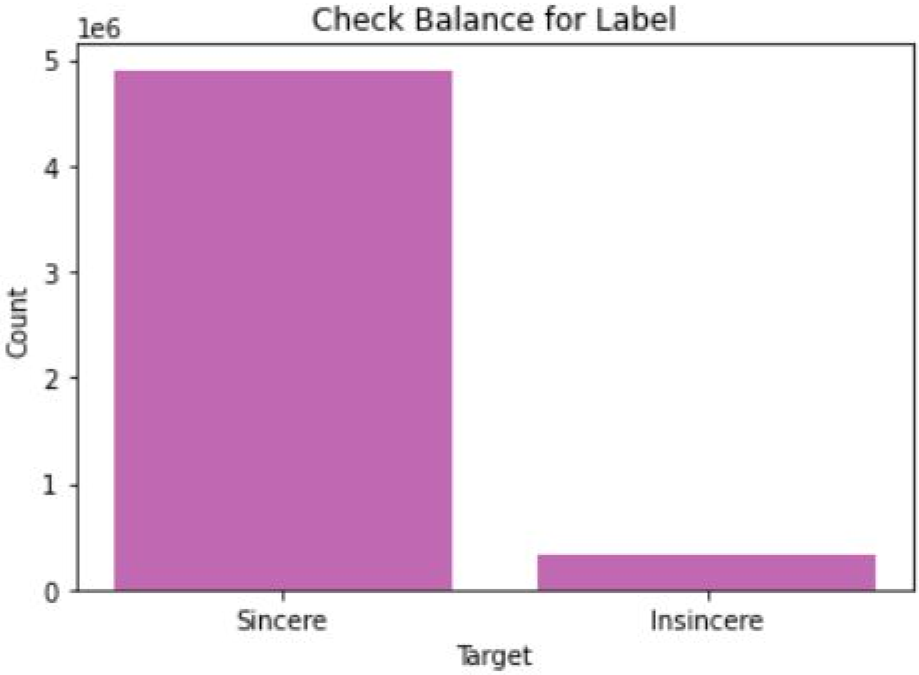
ANLY-590 Audrey Yuan

Introduction

For many websites today, a major problem is to deal with a diversity of content and material. Quora, one of the main websites, serves as a great platform for its users to share their knowledge and experiences, and learn from each other. Users can ask questions and Quora and read about others' unique insights and quality answers. A key challenge is to set apart insincere questions. Sincere questions are intended to make an inquiry or seek help about a specific subject or situation. Questions should be reported as insincere when their primary goal is to make a claim or assert an opinion, rather than looking for helpful answers for a specific inquiry. In this project, we will do sentiment analysis by using some neural network models to address this challenge.

Dataset

We use the Quora Insincere Questions Classification dataset for this project. The dataset contains 5,224,488 rows. Each row contains a question asked on quora and its label where 0 represents sincere and 1 represents insincere. The dataset is highly imbalanced with 4,901,248 sincere questions, and 323, 240 insincere questions.

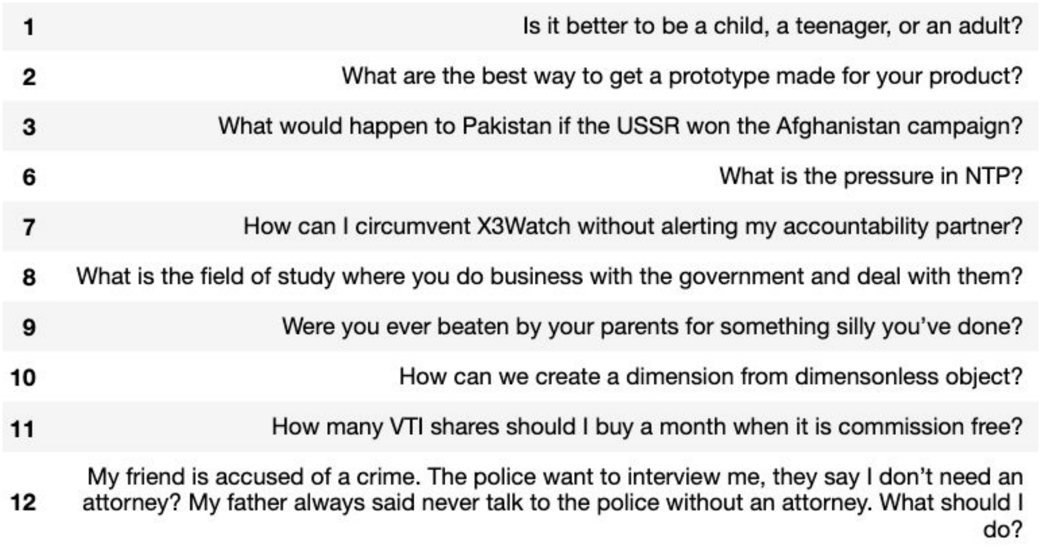


plot 1

To handle the data imbalance, we downsample the dataset to 323,240 sincere questions and 323,240 insincere questions to get rid of potential issues caused by data imbalance.

Exploratory Data Analysis

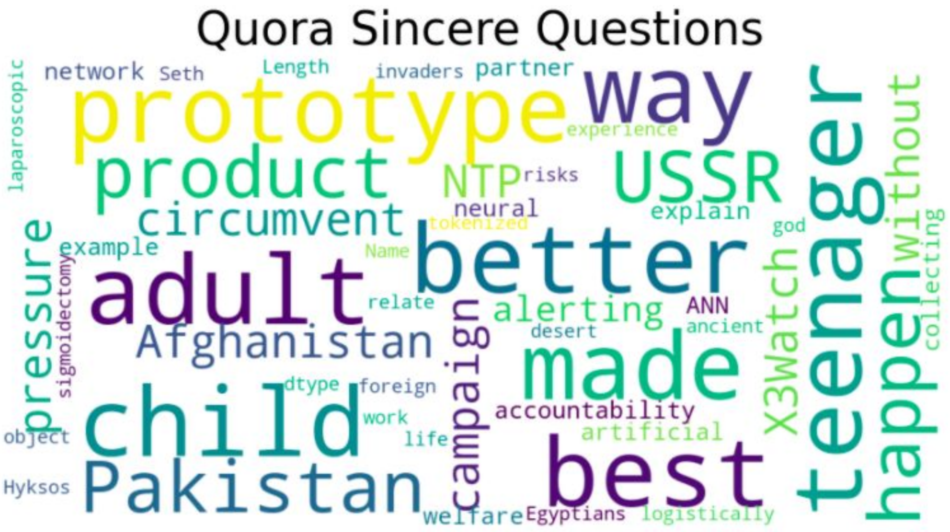
With sincere and insincere questions being defined earlier, we are going to visualize the dataset to better understand what each type of questions looks like. To start off, let’s take a look at some sincere questions and the word cloud of all sincere questions.



1	Is it better to be a child, a teenager, or an adult?
2	What are the best way to get a prototype made for your product?
3	What would happen to Pakistan if the USSR won the Afghanistan campaign?
6	What is the pressure in NTP?
7	How can I circumvent X3Watch without alerting my accountability partner?
8	What is the field of study where you do business with the government and deal with them?
9	Were you ever beaten by your parents for something silly you’ve done?
10	How can we create a dimension from dimensonless object?
11	How many VTI shares should I buy a month when it is commission free?
12	My friend is accused of a crime. The police want to interview me, they say I don’t need an attorney? My father always said never talk to the police without an attorney. What should I do?

plot 2

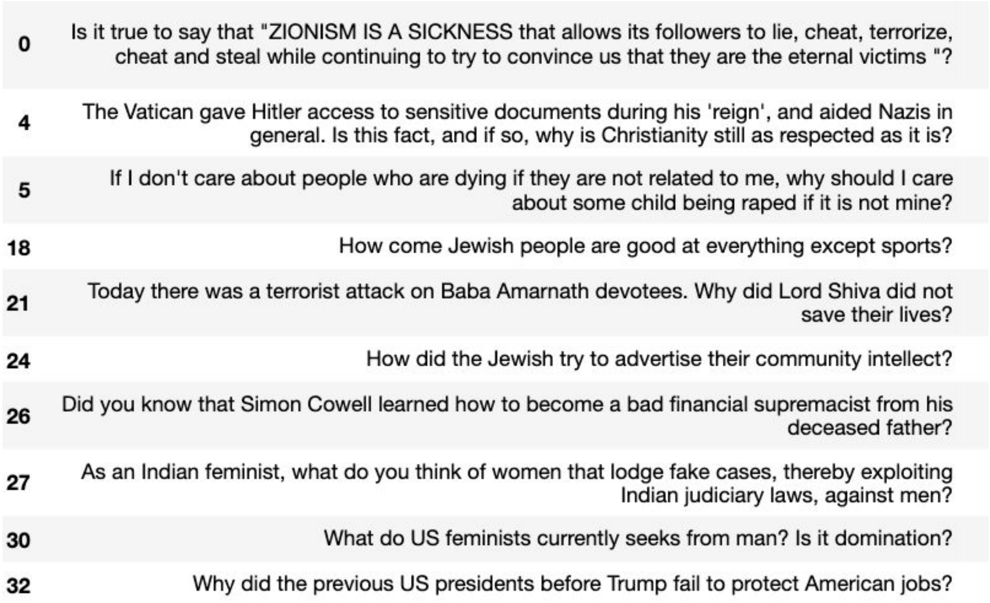
We can see that some typical sincere questions are meaningful and they target a specific area or topic. It seems like the users who asked these questions attempt to receive helpful information and replies. We can also take a look at the word cloud of all the sincere questions.



plot 3

We notice that the most of the commonly used words in sincere questions are either neutral, positive, or very specific for example, better, best, campaign, accountability.

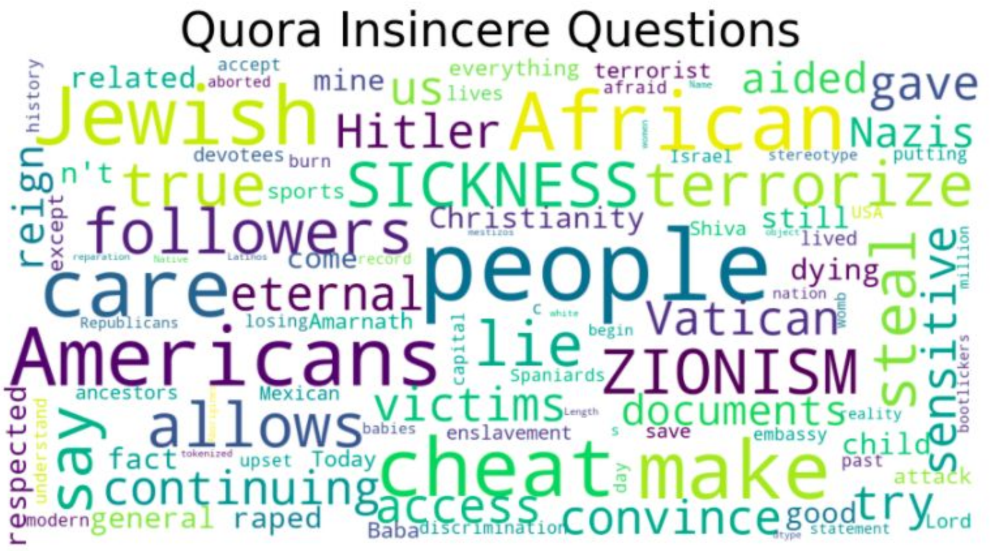
We do the same for insincere questions, and here are some of the typical ones.



0	Is it true to say that "ZIONISM IS A SICKNESS that allows its followers to lie, cheat, terrorize, cheat and steal while continuing to try to convince us that they are the eternal victims "?
4	The Vatican gave Hitler access to sensitive documents during his 'reign', and aided Nazis in general. Is this fact, and if so, why is Christianity still as respected as it is?
5	If I don't care about people who are dying if they are not related to me, why should I care about some child being raped if it is not mine?
18	How come Jewish people are good at everything except sports?
21	Today there was a terrorist attack on Baba Amarnath devotees. Why did Lord Shiva did not save their lives?
24	How did the Jewish try to advertise their community intellect?
26	Did you know that Simon Cowell learned how to become a bad financial supremacist from his deceased father?
27	As an Indian feminist, what do you think of women that lodge fake cases, thereby exploiting Indian judiciary laws, against men?
30	What do US feminists currently seeks from man? Is it domination?
32	Why did the previous US presidents before Trump fail to protect American jobs?

plot 4

A word cloud is also provided to better understand the common characteristics shared between the insincere questions.



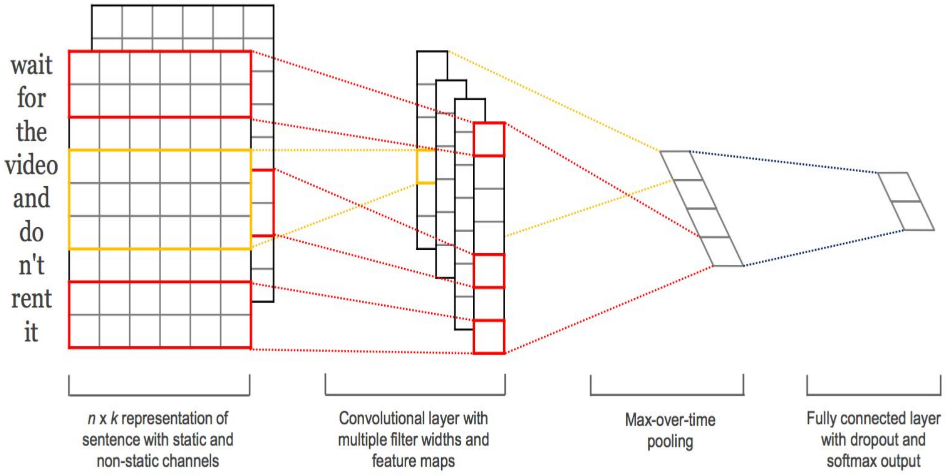
plot 5

Compared to sincere questions, insincere questions contain more aggressive, negative, and sensitive words for instance, sickness, cheat, terrorize, steal, raped, lie, and etc.

Models

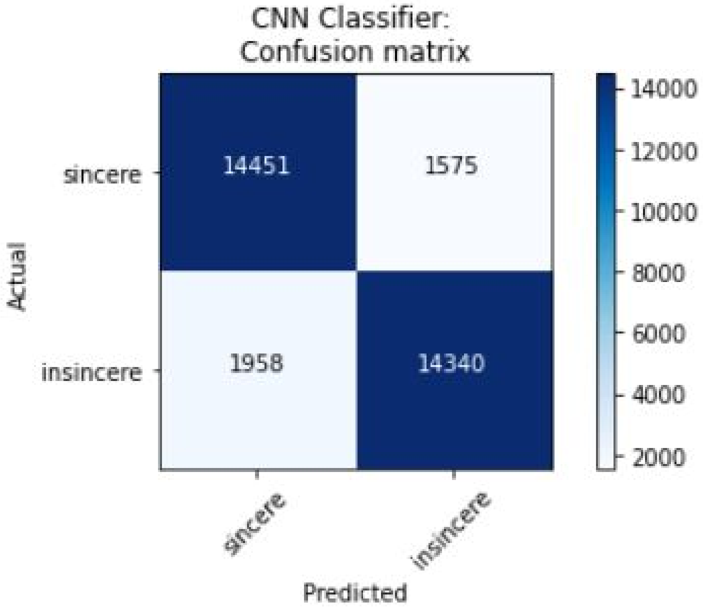
We used both convolutional neural network and recurrent neural network to train the classification model. For both model, we will need an embedding layer, which could either be a pre-trained word embedding or trained on specific corpus. In this project, we trained our own word embedding because pre-trained word embedding models are trained on general sources which may not work well on our specific Quora dataset.

For CNN model, we use 1D convolutional layers. The sizes of 1D convolutional layers are 3, 4, and 5. We have 100 convolutional layers for each size. We used dropout layer to prevent overfitting. The convolutional layers work like sliding windows to scan the sentences, multiple consecutive words at a time. It is also worth mentioning that since the lengths of the sentences are different, we will have to do zero padding for the shorter sentences and at the end, make all sentences have equal lengths.


wait
for
the
video
and
do
n't
rent
it
n x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

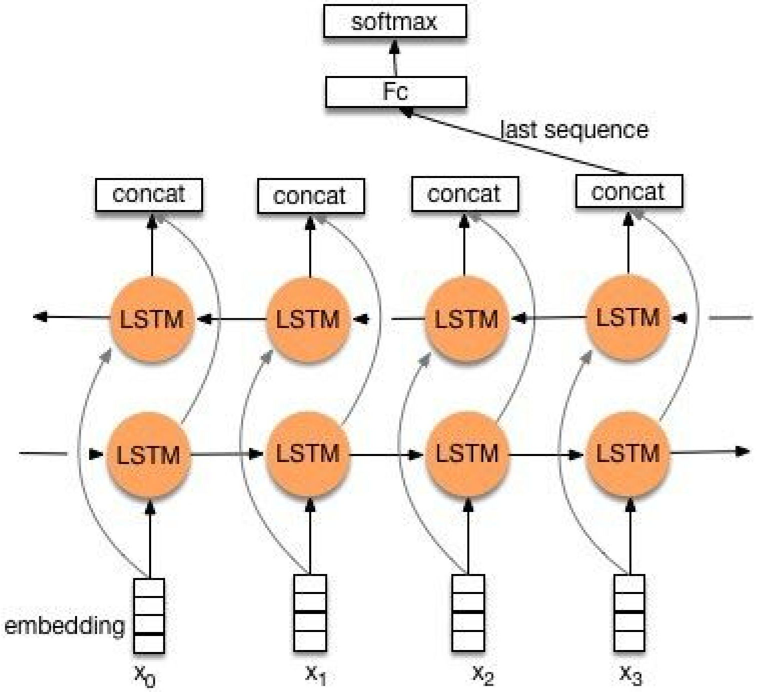
plot 5

The confusion matrix for the CNN model is as follows:



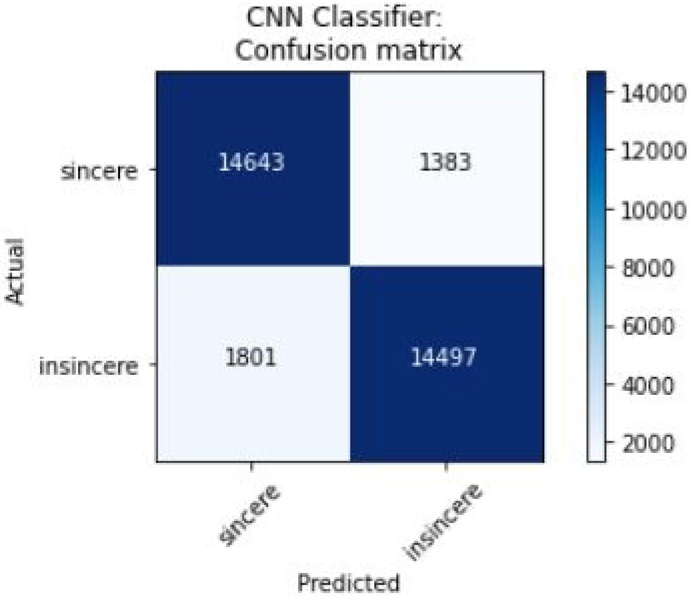
plot 6

For RNN model, we use bidirectional LSTM, which could read the sentences in both directions - from start to end and from end to start. For hidden layers, we set the dimension as 256. We also used dropout layer to prevent overfitting. Since the network is bidirectional, at the end, we concatenate the hidden layers in two directions together to get the final result.



plot 7

The confusion matrix for the CNN model is as follows:



plot 8

Results

	CNN	LSTM
Accuracy	89.23	90.07
F1	89.22	90.06

We evaluate the CNN and LSTM using the accuracy and macro f1 score. Overall, both neural networks work well. The LSTM is slightly better than the CNN, by approximately 1 percent.

Conclusion

Both CNN and LSTM are very powerful in text classification tasks. The LSTM is a little bit better than CNN because it reads through the whole sentence in both directions, which means that the prediction is based on all the previous contexts. On the other hand, the CNN only uses a sliding window to scan through the sentences by one pass. And it only checks the context in certain windows. This is the reason why LSTM is better than CNN in doing text classification tasks.