EE258 - Neural Networks

Project 2 - Quora Insincere Questions Classification

EE258_F18_PN on Kaggle

Name: Parvathi Chandrasekhar
ID : 012474995
Kaggle ID: https://www.kaggle.com/parvathich

Project Partner Name: Neeshitha Sudhakar
ID: 012459005

1. DATA

1.1. Dataset Description

- The Kaggle competition we picked is the Quora Insincere Questions Classification. The dataset provided here is a collection of questions from Quora with labels identifying them as sincere or insincere. Questions that are worded strongly for shock value, have disparaging terms and disingenuous motives are labeled as insincere.

- The training data contains 1306122 samples with 3 columns, including a question id, the question text and the label which classifies it into an insincere or sincere question.

- The test data contains 56370 samples with 2 columns.

```
Train shape :  (1306122, 3)
Test shape :  (56370, 2)
```

- Sample questions:

	qid	question_text	target
0	00002165364db923c7e6	How did Quebec nationalists see their province...	0
1	000032939017120e6e44	Do you have an adopted dog, how would you enco...	0
2	0000412ca6e4628ce2cf	Why does velocity affect time? Does velocity a...	0
3	000042bf85aa498cd78e	How did Otto von Guericke used the Magdeburg h...	0
4	0000455dfa3e01eae3af	Can I convert montra helicon D to a mountain b...	0

1.2. Dataset Distribution

- Before working on developing the model, initially the first step taken was to inspect the dataset. The problem here is a classification text related problem.

- Analyzing the ratio of sincere to insincere questions, it is seen that the number of sincere questions [0's] are higher than 1's. With this realization, there was an idea to not set accuracy as a performance metric in any model that would be run as the dataset is unevenly distributed.

- F-1 score is a good metric to use as it combines both precision and recall which are useful metric when dealing with a text-based dataset.

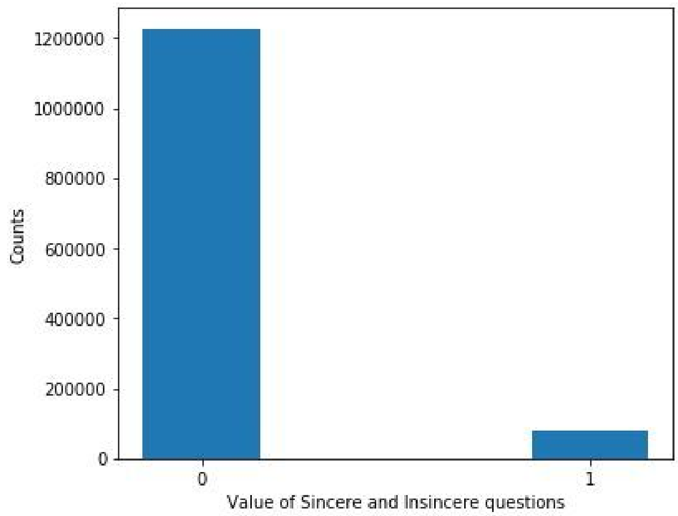


Fig.1. Value counts of insincere vs sincere questions

Some other features we decided to look at were:

1. Number of words/question

	question_text	word_count
0	How did Quebec nationalists see their province...	13
1	Do you have an adopted dog, how would you enco...	16
2	Why does velocity affect time? Does velocity a...	10
3	How did Otto von Guericke used the Magdeburg h...	9
4	Can I convert montra helicon D to a mountain b...	15

2. Number of characters/question

	question_text	char_count
0	How did Quebec nationalists see their province...	72
1	Do you have an adopted dog, how would you enco...	81
2	Why does velocity affect time? Does velocity a...	67
3	How did Otto von Guericke used the Magdeburg h...	57
4	Can I convert montra helicon D to a mountain b...	77

3. Average word length/question

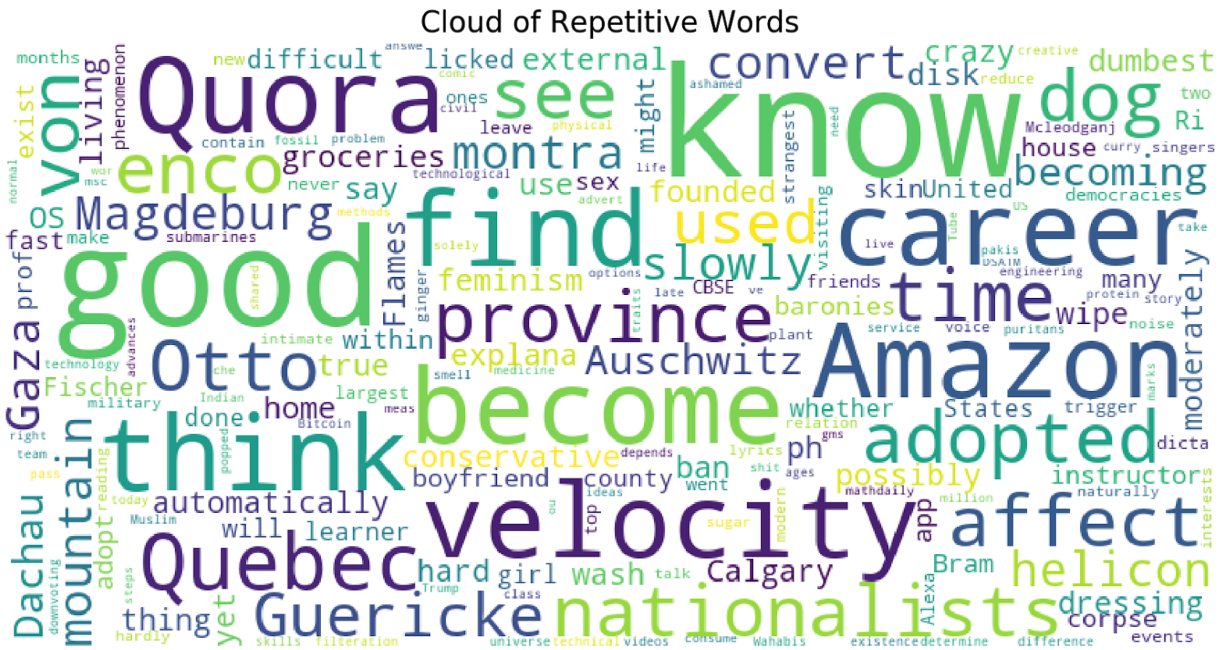
	question_text	avg_word
0	How did Quebec nationalists see their province...	4.615385
1	Do you have an adopted dog, how would you enco...	4.125000
2	Why does velocity affect time? Does velocity a...	5.800000
3	How did Otto von Guericke used the Magdeburg h...	5.444444
4	Can I convert montra helicon D to a mountain b...	4.200000

4. Stop words/question

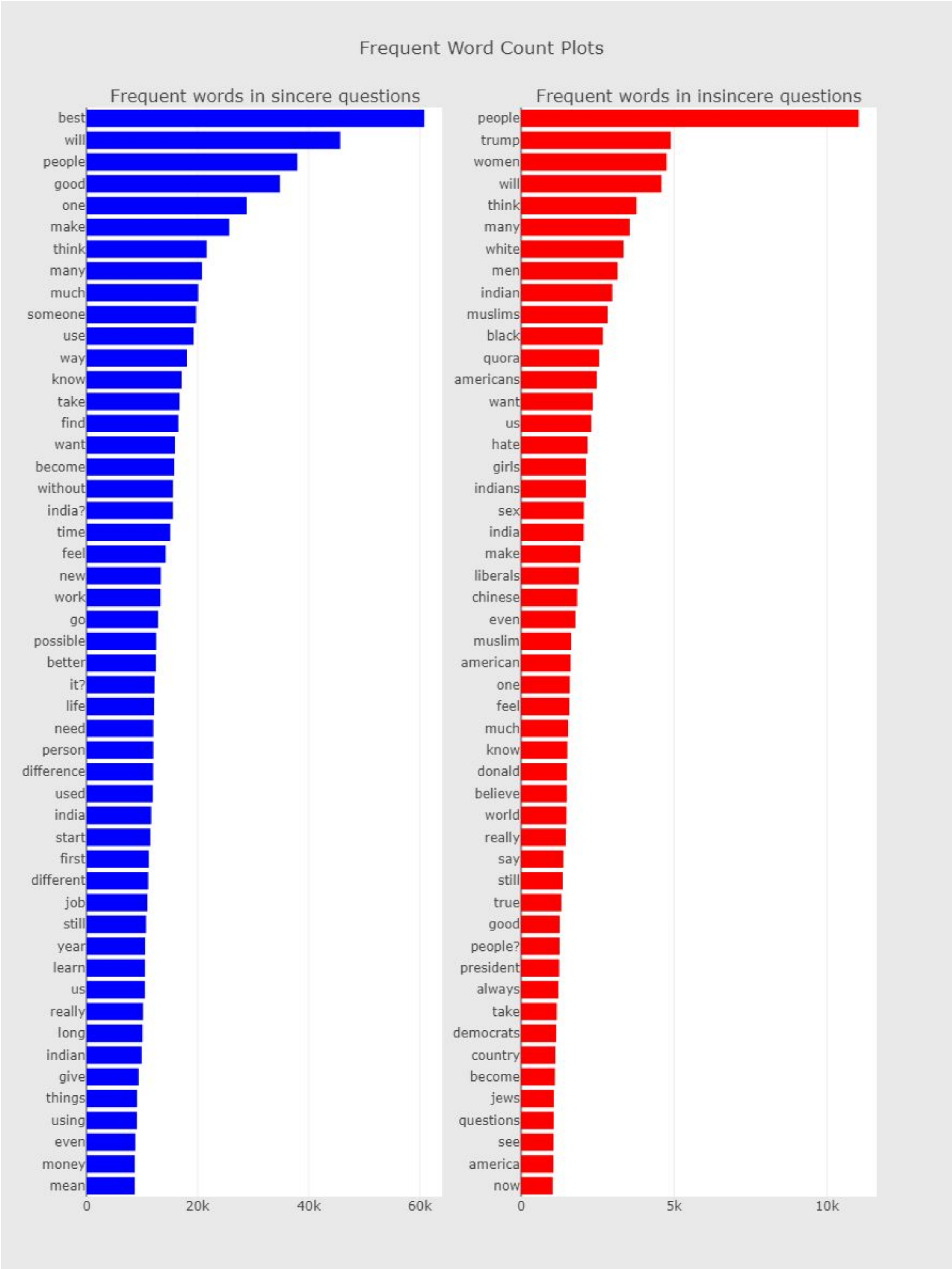
	question_text	stopwords
0	How did Quebec nationalists see their province...	6
1	Do you have an adopted dog, how would you enco...	8
2	Why does velocity affect time? Does velocity a...	1
3	How did Otto von Guericke used the Magdeburg h...	2
4	Can I convert montra helicon D to a mountain b...	5

5. Word Cloud

- To check the frequency of words in the dataset.

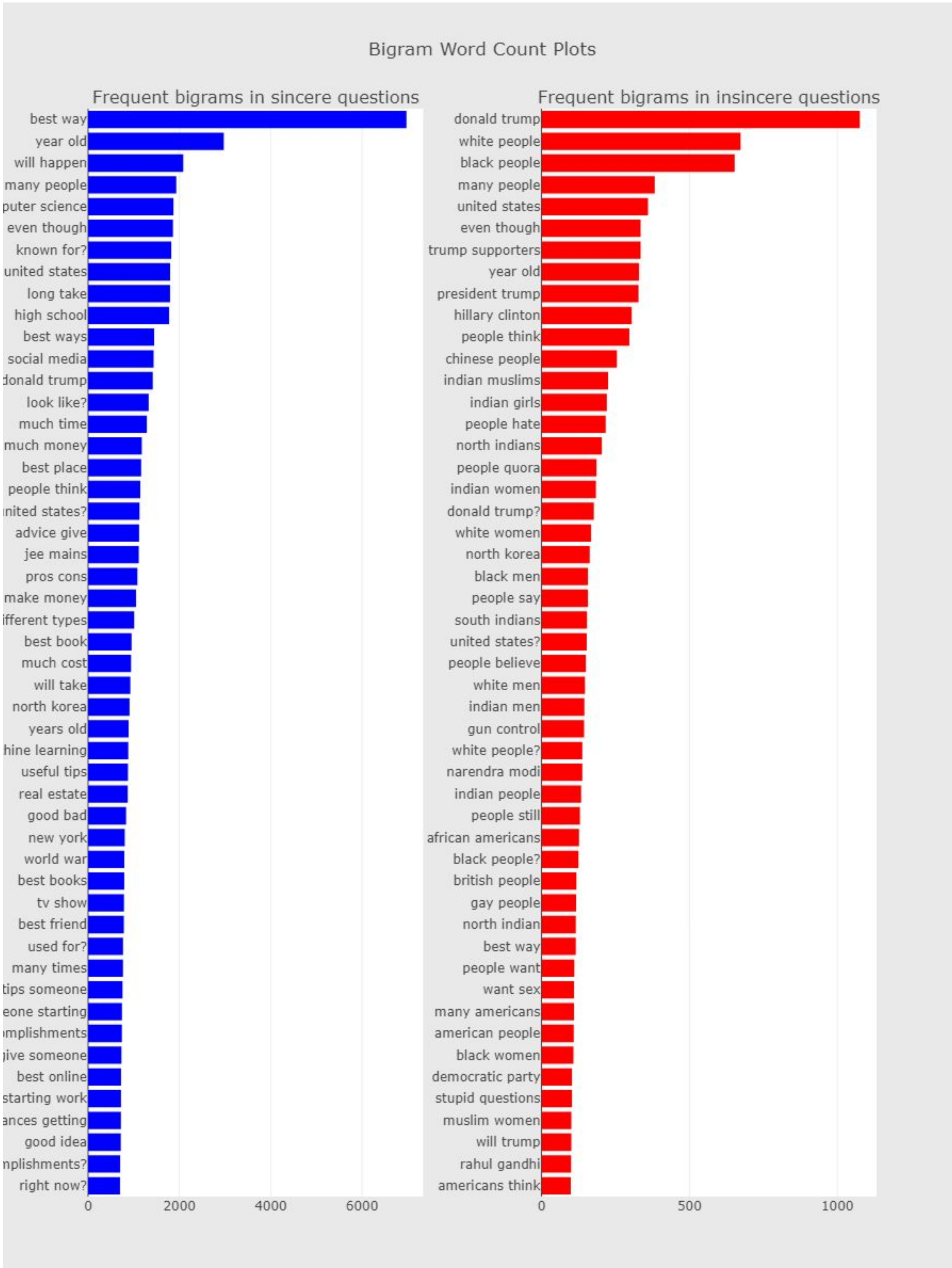


6. Frequency of words in insincere and sincere questions

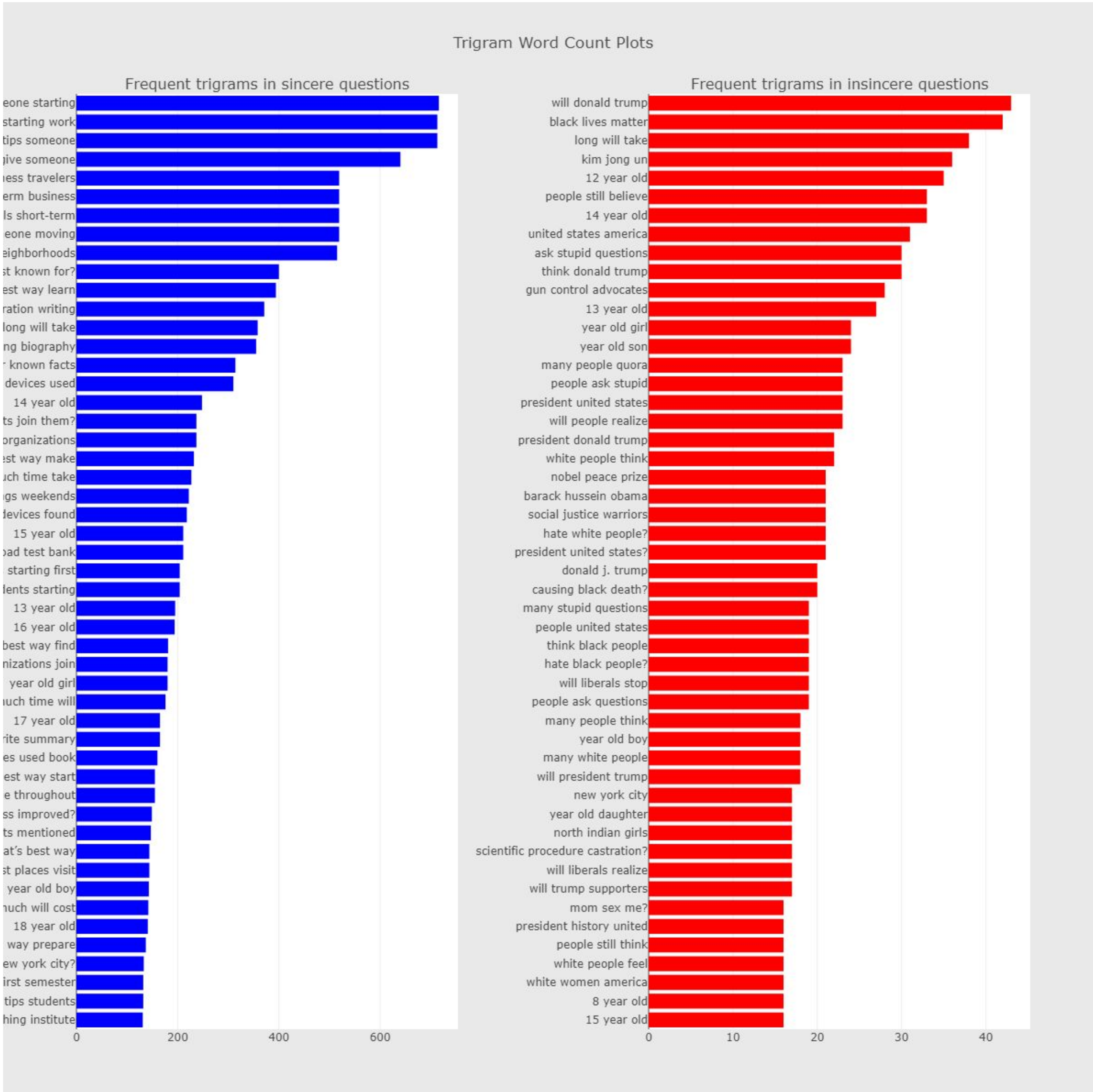


Frequent Word Count Plots
Frequent words in sincere questions
best
will
people
good
one
make
think
many
much
someone
use
way
know
take
find
want
become
without
india?
time
feel
new
work
go
possible
better
it?
life
need
person
difference
used
india
start
first
different
job
still
year
learn
us
really
long
indian
give
things
using
even
money
mean
0
20k
40k
60k
Frequent words in insincere questions
people
trump
women
will
think
many
white
men
indian
muslims
black
quora
americans
want
us
hate
girls
indians
sex
india
make
liberals
chinese
even
muslim
american
one
feel
much
know
donald
believe
world
really
say
still
true
good
people?
president
always
take
democrats
country
become
jews
questions
see
america
now
0
5k
10k

7. Bigram Frequency Plot



8. Trigram Word Plot



1.2. **Dataset Pre-processing**

1. Removal of Punctuation

```
'How did Quebec nationalists see their province as a nation in the 1960s'
```

2. Removal of Stop Words

```
'How Quebec nationalists see province nation 1960s'
```

2. **METHODOLOGY**

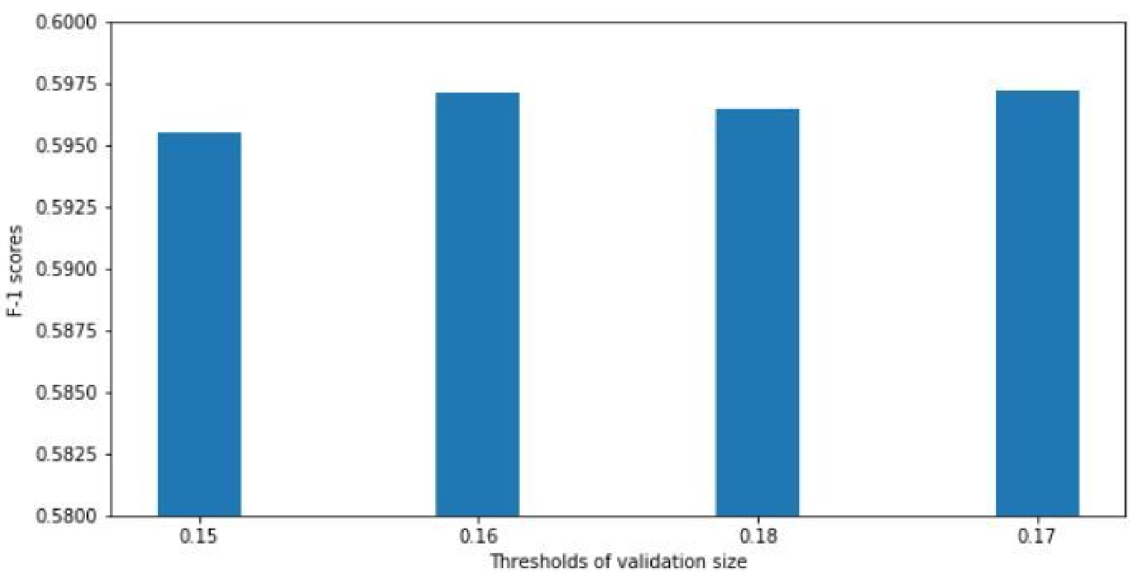
2.1. Baseline Model

- The model used for our baseline is a logistic regression model [activation function from 0 to 1, and good for classification problems]. The reason logistic regression was picked is because we wanted to experiment with the baseline model quite a bit before using a sequential model such as RNN which is known for working best with text based data.

- K-fold cross validation [with shuffling] was also done to split the training dataset into 5 folds, one fold is then used as validation while the other 4 folds are used as training.

- Logistic regression works by predicting the probability of the class that the input may belong to. The learning algorithm used here is a stochastic average gradient for faster convergence.

The resulting F-1 score from using this model:



- It’s observed that the best F-1 score is at a validation threshold of 0.17, and the score being 0.5972.

2.2. LSTM model

- After pre-processing, we use the modified dataset as inputs to an LSTM model.

- The LSTM model used here uses pre-trained embeddings given with the dataset. We opted for this method as it saves time. Embeddings are geometrical encodings of vectors from their co-occurrence or frequency of words. We used the GloVe model, which is a “count-based” model. GloVe counts the co-occurrence of words appearing in a large text based data by constructing a co-occurrence matrix. It constructs a matrix of words [rows] vs how many times it appears in a context in the text data [columns].

- We implemented a Bidirectional LSTM model with attention layers, a dropout layer [this is the regularization technique we use], and dense layers with the output dense layer containing one output and a sigmoid activation function.

3. MODEL IMPROVEMENTS

- Our biggest improvement from the baseline model is using a Recurrent Neural Network such as Bidirectional LSTM.

- It provided a much better F-1 score by exploiting the sequential nature of text based data.

- The attention, dropout and dense layers implement the LSTM model here as compared to a simple logistic regression model used in the baseline.

- The F-1 scores obtained here was by changing the epoch parameter to give better results. F-1 score was opted as our performance metric due to the uneven distribution of dataset [number of sincere vs insincere questions].

- The graph of varying epochs and F-1 score:

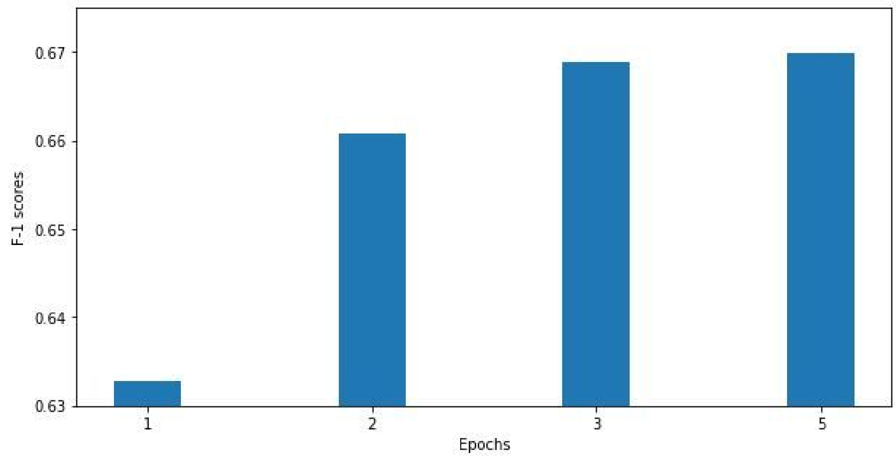


Fig.2. F-1 scores vs Epochs

- The graph of varying validation accuracy:

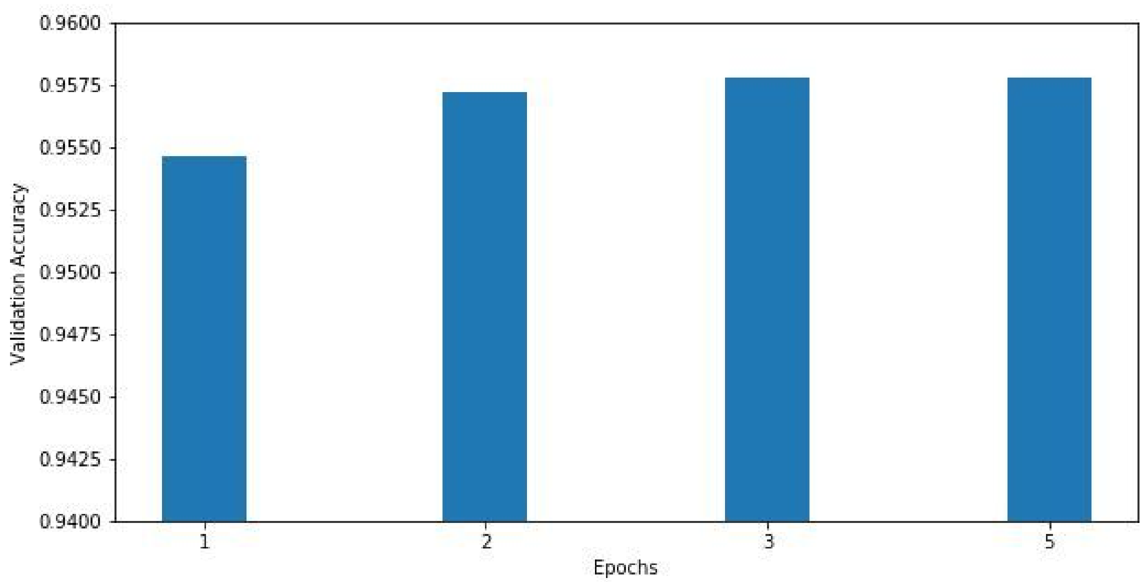


Fig.3. Validation Accuracy vs Epochs

4. RESULTS and FUTURE WORK

- The best F-1 score was from 5 epochs and was 0.6739.

- Better results can be obtained by adding more layers to the LSTM as well as running for more epochs.

- Better results could also have been obtained by tweaking activation functions of output layer as well as hidden layers, but we did not have time to do that.

4.1 References

- https://www.kaggle.com/mihaskalic/lstm-is-all-you-need-well-maybe-embeddings-also
- https://www.kaggle.com/nikhilroxtomar/embeddings-cnn-lstm-models-lb-0-683
- https://www.kaggle.com/demery/character-level-tfidf-logistic-regression
- https://www.kaggle.com/sudalairajkumar/a-look-at-different-embeddings