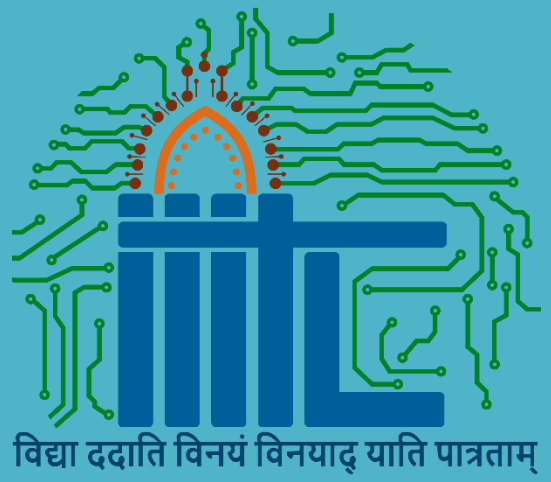


Quora insincere questions classification

By- Surbhi Suman

MCS20003

What is Quora?

- Quora is a platform that empowers people to learn from each other.
- On Quora, people can ask questions and connect with others who contribute unique insights and quality answers.
- A key challenge is to weed out insincere questions i.e. those founded upon false premises, or that intend to make a statement rather than look for helpful answers.

Problem statement-

- Detect inappropriate content and improve online conversation

Solution approach-

- Taking it as classification problem
- Find out the definition of what an “insincere ” question actually is.
- Point out keywords i.e. feature engineering
- Using several models
- Train different models
- Take their mean as final output

Models-

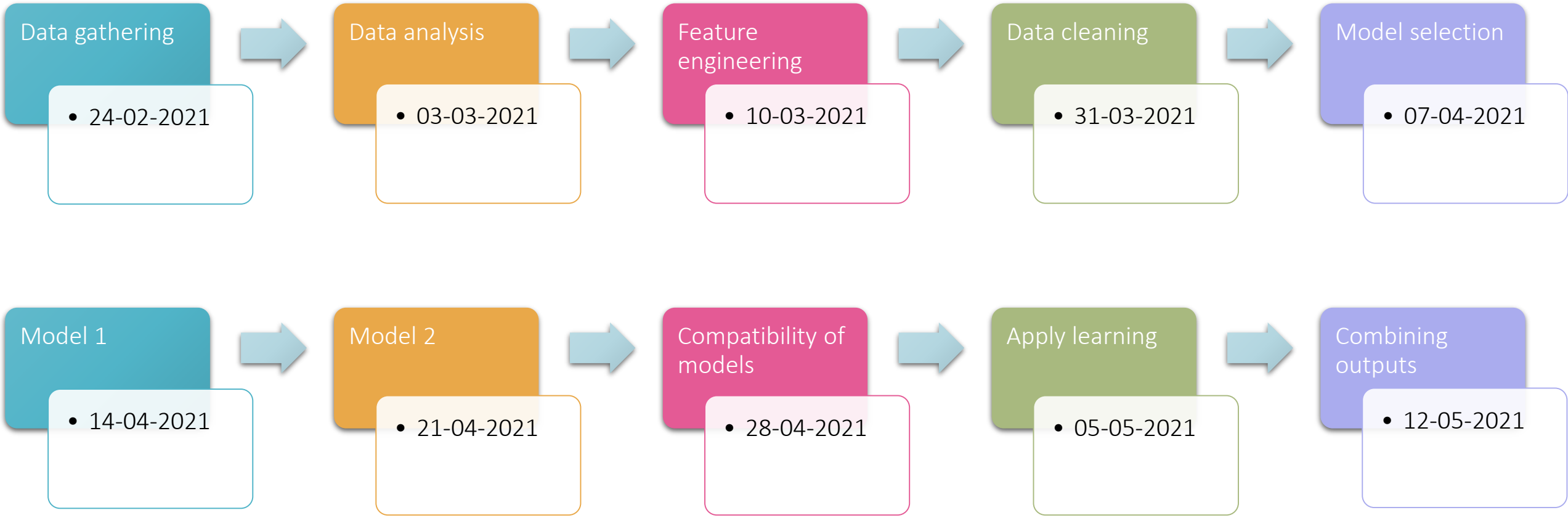
- LSTM- **Long short-term memory** is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video).
- L2 regularization- **Regularization** is a process of introducing additional information in order to prevent overfitting. A regression model which uses L2 is called ***Ridge Regression***.
- NB-SVM- Support vector machine with Naïve Baeyes
- Bidirectional LSTM- The NLP CNN is usually made up of 3 or more 1D convolutional and pooling layers unlike traditional CNNs. This helps reduce the dimensionality of the text and acts as a summary of sorts which is then fed to a series of dense layers.

Metric

- As it is a classification based problem so F1 score is taken as metric for evaluation.
- F1-score is given by harmonic mean of precision and recall

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})}$$

Timeline


Data gathering
• 24-02-2021
Data analysis
• 03-03-2021
Feature engineering
• 10-03-2021
Data cleaning
• 31-03-2021
Model selection
• 07-04-2021
Model 1
• 14-04-2021
Model 2
• 21-04-2021
Compatibility of models
• 28-04-2021
Apply learning
• 05-05-2021
Combining outputs
• 12-05-2021

Data gathering

Data for this problem statement was available in kaggle.

There are 4 kind of files provided by kaggle-

1. The training set- train.csv
2. The testing set- test.csv
3. Embeddings- as external data isn’t allowed so these embedding were provided- Google news, Glove, Paragram, Wiki-news
4. Sample output format

qid	question_text
1306122 unique values	1306122 unique values
00002165364db923c7e6	How did Quebec nationalists see their province as a nation in the 1960s?
000032939017120e6e44	Do you have an adopted dog, how would you encourage people to adopt and not shop?
0000412ca6e4628ce2cf	Why does velocity affect time? Does velocity affect space geometry?
000042bf85aa498cd78e	How did Otto von Guericke used the Magdeburg hemispheres?

Training Data

qid	question_text
375806 unique values	375806 unique values
0000163e3ea7c7a74cd7	Why do so many women become so rude and arrogant when they get just a little bit of wealth and power...
00002bd4fb5d505b9161	When should I apply for RV college of engineering and BMS college of engineering? Should I wait for ...
00007756b4a147d2b0b3	What is it really like to be a nurse practitioner?
000086e4b7e1c7146103	Who are entrepreneurs?
0000c4c3fbe8785a3090	Is education really making good people nowadays?
000101884c19f3515c1a	How do you train a pigeon to send messages?

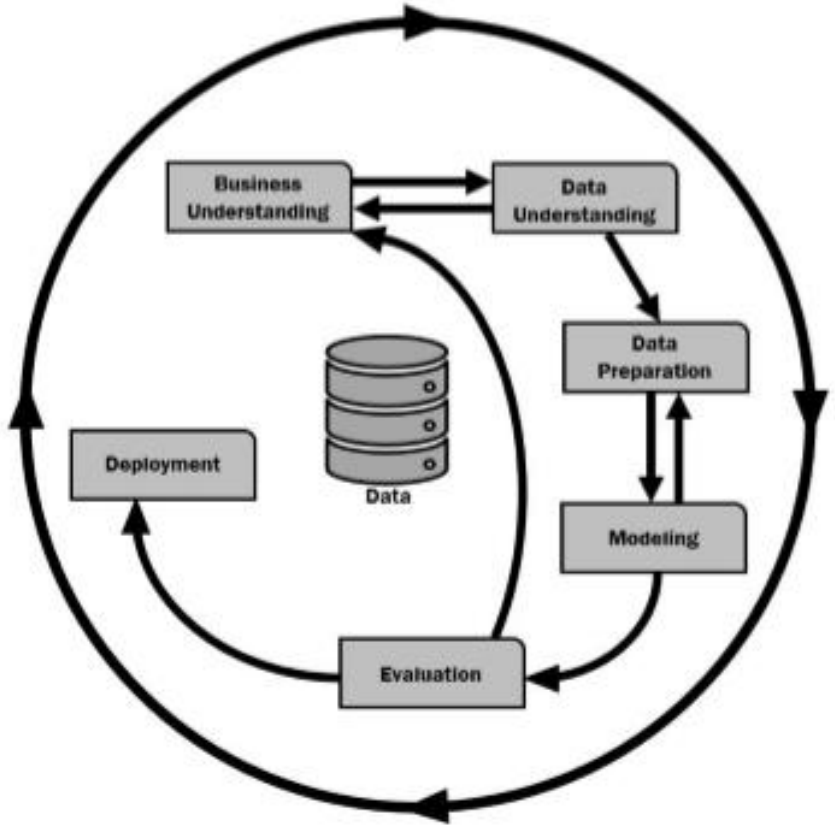
Test Data

Data fields:

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

Literature survey

- By- National college of Ireland
- Data pre-processing- word tokenization, removal of stop words, finding root word
- Text vectorization- TF-IDF and DTM
- Models – Logistic regression, Naïve Bayes, combination of previous ones, SVM, Decision tree, Random Forest.



❖ CRISP-DM Methodology

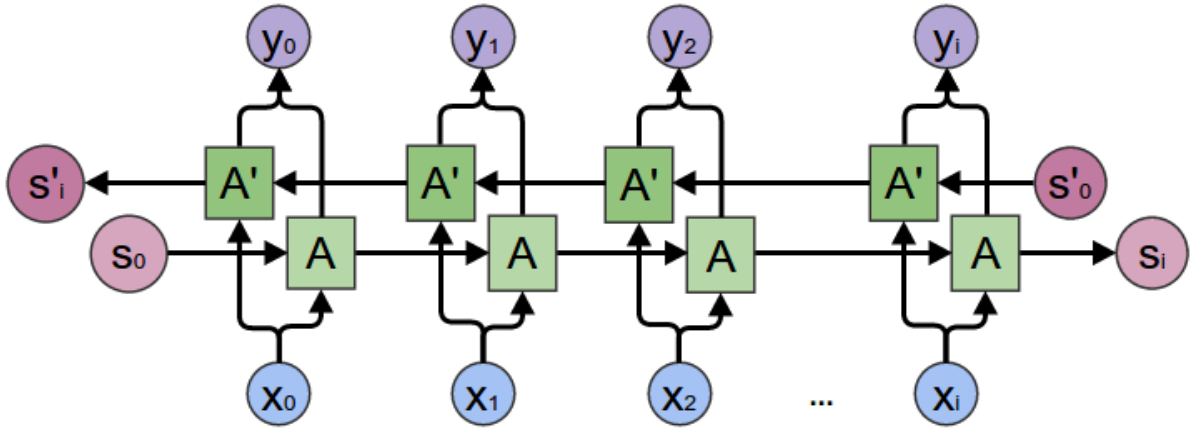
Parameter	Pre-Processing	Accuracy	Precision	Recall	F1 Score
Naïve Bayes	Count Vectorizer + Lemmatization	0.95	0.63	0.62	0.63
Logistic Regression	Count Vectorizer	0.95	0.63	0.62	0.63
Naïve Bayes + Logistic Regression	TF-IDF	0.95	0.64	0.62	0.63
SVM	TF-IDF + Lemmatization	0.93	0.45	0.42	0.43
Decision Tree	Count Vectorizer + Lancaster Stemmer	0.93	0.49	0.43	0.46
Random Forest	Count Vectorizer + Lemmatization	0.94	0.63	0.30	0.41

Author- NC state ECE

Key-points-

- Data pre-processing- Removal of URLs, Removal of punctuations, Lemmatization
- Vector Embedding- GloVe projects the words into vector space such that similar words are closer to each other
- Model- Bidirectional LSTM

Result- F1-score-0.89



Author- **Ronak Vijay**

Key-points-

- Feature Engineering(before cleaning)-Number of words, Number of capital_letters, Number of special characters, Number of unique words, Number of numerics, Number of characters, Number of stopwords
- Data pre-processing- by lemmatization and TF-IDF
- Models-
  i. *Bidirectional RNN(LSTM/GRU)*
  ii. *Bidirectional GRU with Attention Layer*
  iii. *Bidirectional LSTM with Convolution Layers*

Result- Optimal F1-0.6974 at threshold 0.3636

Author- Deepshi Mediratta, Nikhil Oswal (University of Ottawa)

Key-points-

- Data pre-processing - Vectorization, TF-IDF
- Resampling- cluster centroids, Random under sampler
- Models- SVM, Naïve bays,GRU, LSTM

Result- Best result by GRU using Glove Embedding with Accuracy = 89.49, F1 score = 0.72

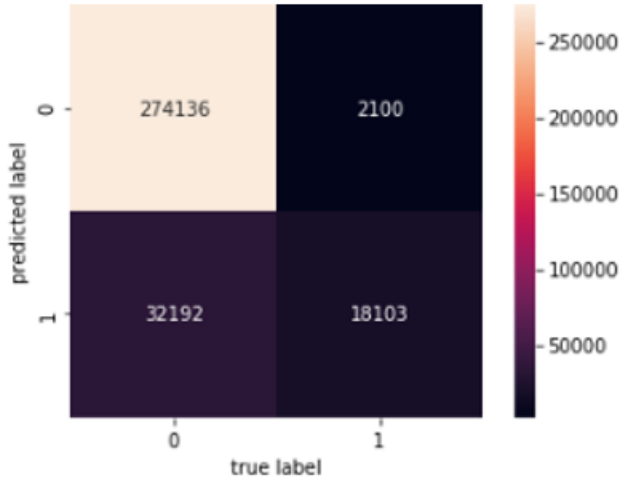
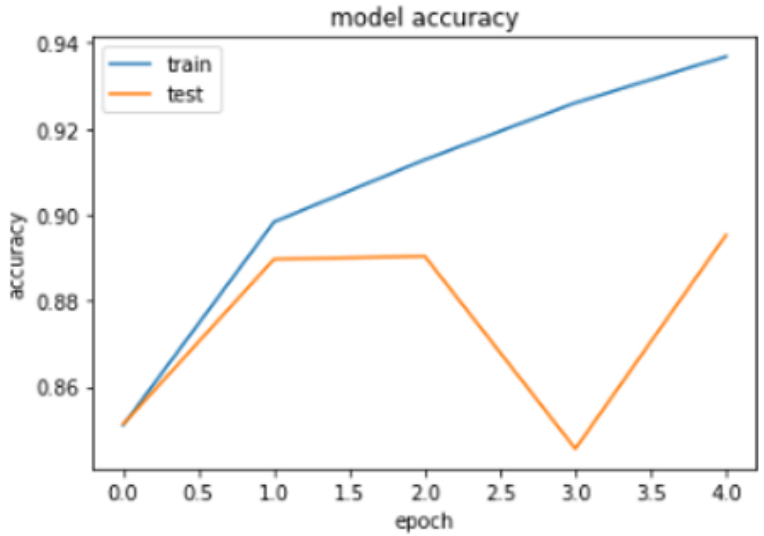
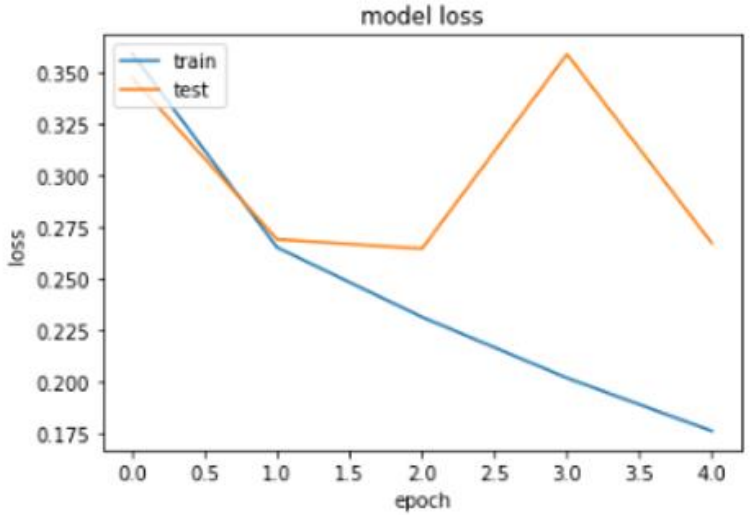


Fig. 12. Confusion Matrix – CuDNNGRU – Undersampling - Using Glove Embedding





•By- Hao Mao (haomao@), Rekha Kumar (rekha123@), Jerry Chen (jchen98@) {stanford.edu}

•Data pre-processing- punctuation removal, tokenization, stemming

•Models- Naïve Bayes, Logistic regression, Averaged perceptron, RNN, BERT

➢ Averaged Perceptron -The perceptron is a classic learning algorithm for finding a linear decision boundary for classification problems. The issue with the vanilla perceptron is that it counts later data points more than it counts the earlier data points. Averaged Perceptron is a modification of the Perceptron algorithm where we maintain a running sum of the averaged weight vector.

$$\hat{y} = \text{sign}\left(\sum_{k=1}^{K} c^{(k)}\left(\boldsymbol{w}^{(k)} \cdot \hat{\boldsymbol{x}} + b^{(k)}\right)\right)$$

➢BERT (Bidirectional Encoder Representations from Transformers) .BERT applies the bidirectional training of Transformer, a popular attention model, to language modeling.

•Result- 0.63 F1-score with Averaged perceptron

Data analysis

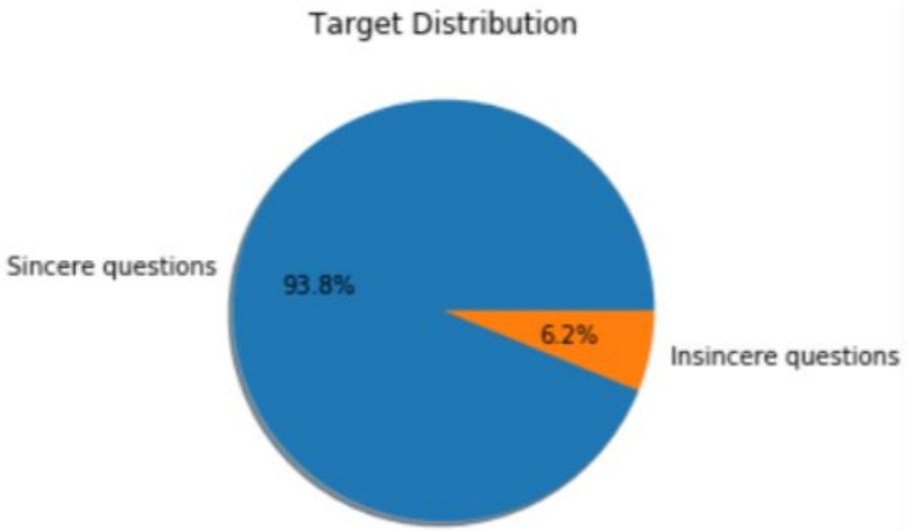
- Size of testing data=1306122 rows and 3 columns
- Size of training data=375806 rows and 3 columns
- There are no null value. Nearly pre-cleaned data.
- 2 columns are object type and the third column is int type

```
trdata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1306122 entries, 0 to 1306121
Data columns (total 3 columns):
 #   Column         Non-Null Count    Dtype
---  ------         --------------    -----
 0   qid            1306122 non-null  object
 1   question_text  1306122 non-null  object
 2   target         1306122 non-null  int64
dtypes: int64(1), object(2)
memory usage: 29.9+ MB
```

Percentage calculation-

- sincere data= 93.81
- Insincere data=6.18
- Data is highly unstable




```
In [6]: sincere=len(trdata.question_text[trdata["target"]==0])/len(trdata["question_text"])*100
        print(sincere)

        93.81298224821265
```