

Quora - A social platform where people can ask questions and connect with others who contribute unique insights and quality answers.

The Problem - Detecting/handling toxic and divisive content.

Aim - Develop models that identify and flag insincere questions. This project is a way to combat the problem and uphold Quora's status as a place for sharing and growing the world's knowledge.

#### The Data



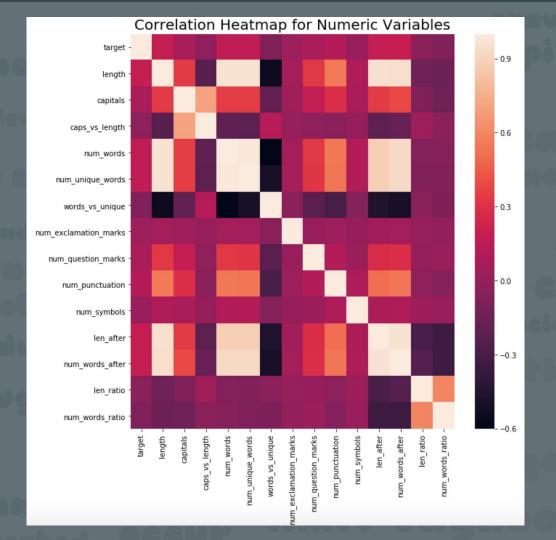
#### ~1.3 Million Questions

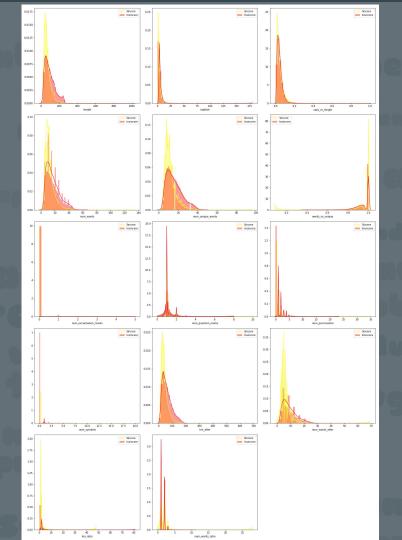


**Sincere Questions** 

**Insincere Questions** 

Understanding numerical features of question text





Distribution of numeric variables in Sincere vs Insincere questions

#### Word Cloud of Sincere Questions



# Understanding textual features of question text

#### **Word Cloud of Insincere Questions**

turnglyeword distance of the second of the s

Top 20 1,2,3 and 5 words phrases in sincere questions year old high school computer science united state good book long take good place donald trump many people iee main social medium even though make money good friend north korea good thing major accomplishment ty show much time 100000 60000 80000 4000 Top 20 1,2,3 and 5 words phrases in insincere questions useful tip someone start work tip someone start good hotel shortterm business traveler someone start work useful tip someone useful tip student start first advice give someone tip student start first semester give someone move donald trump knowledge experience affect way read white people black people experience affect way read book year old movie greenlit backstory movie make president trump good way learn tip student start first year trump supporter much time take united state take consideration write hillary clinton year old girl consideration write biography many people good way make north indian lesser know fact literary device use improve stay competitive strategically meet people think high school student help improve stay competitive strategically indian girl new york city asset help improve stay competitive black man people asset help improve stay chinese people someone white man much white woman man african american want 400 600 black life matter jew deny cause black death ask stupid question freedom shoot people walk street year old girl people walk street bomb civilian united state america kim iong un female female specie testicle come year old boy human female female specie testicle president donald trump testicle replace testicle another animal think donald trump yearold girl bore tit hole gun control advocate swedish woman think ethnic swede remove testicle replace testicle another hate black people castrate remove testicle replace testicle make america great happen castrate remove testicle replace people still believe many people ask stupid question donald j trump people say bully jealous victim hate white people social justice warrior president united state ejaculate inside human female female islam religion peace victim well finally kill victim people ask stupid throw victim well finally kill

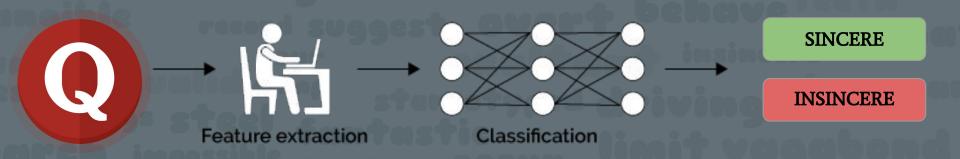
1. Insincere questions have more words than sincere questions on average.

Statistical Inferences about demographic features of the data

2. Insincere question have more number of punctuations than sincere ones.

#### Machine Learning

- → Logistic Regression
- → Naive Bayes
  Classifier
- → KNN Classifier
- → XG Boost Classifier



### Deep Learning Deep Neural Network and LSTM

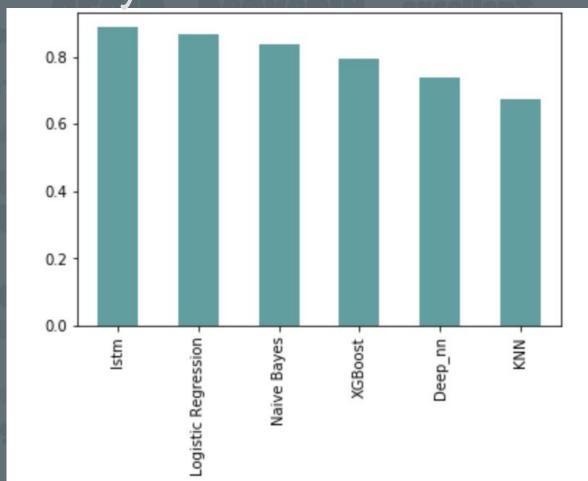
Output	Shape	Param #
(None,	512)	256512
(None,	512)	0
(None,	512)	262656
(None,	512)	0
(None,	512)	262656
(None,	512)	0
(None,	2)	1026
(None,	2)	0
	(None, (None, (None, (None, (None, (None, (None,	Output Shape  (None, 512)  (None, 52)

Total params: 782,850 Trainable params: 782,850 Non-trainable params: 0

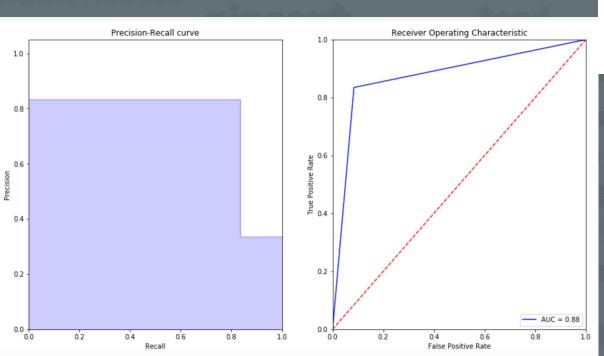
Layer (type)	Output	Shape	Param #
embedding_1 (Embedding)	(None,	1, 128)	21564032
spatial_dropout1d_1 (Spatial	(None,	1, 128)	0
lstm_1 (LSTM)	(None,	64)	49408
dense_1 (Dense)	(None,	1)	65
Total params: 21,613,505			

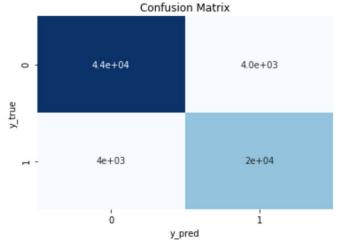
Total params: 21,613,505 Trainable params: 21,613,505 Non-trainable params: 0

#### **Accuracy Scores for Different Models**



## Confusion matrix for LSTM neural network

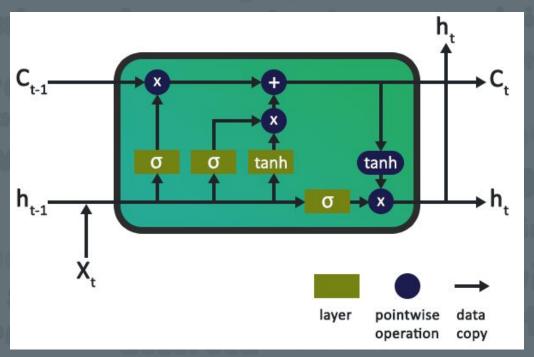


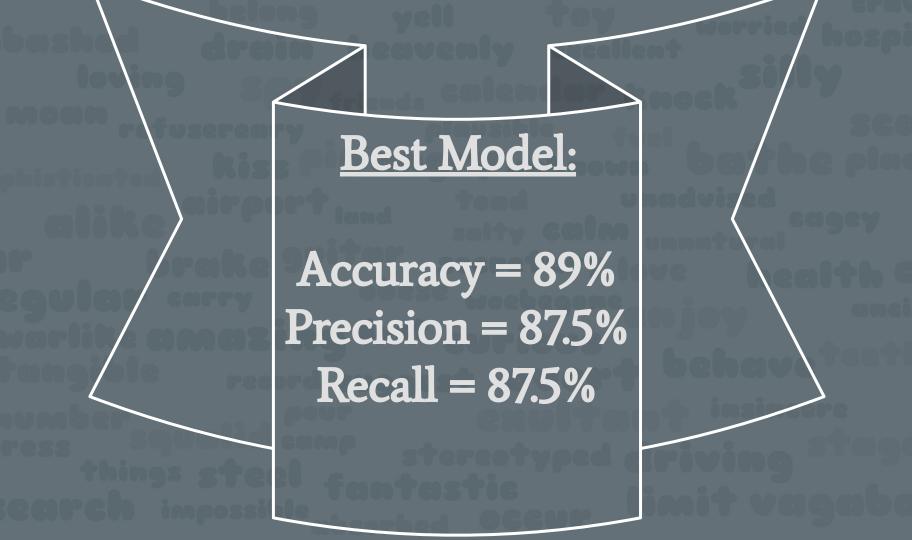


## Precision-Recall Curve and ROC Curve

#### Conclusion

#### LSTM results in best accuracy.

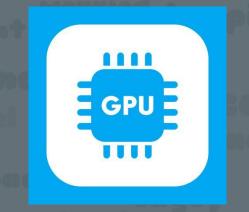


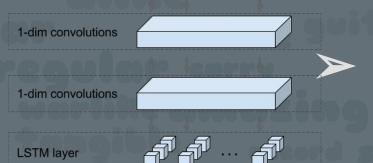


#### **Future Directions**

Train models using GPU

More epochs







input data

1-dim convolutions

Softmax layer

2

C classes

Combine different models to increase accuracy

- Bidirectional LSTM
- Convolutional Neural Network