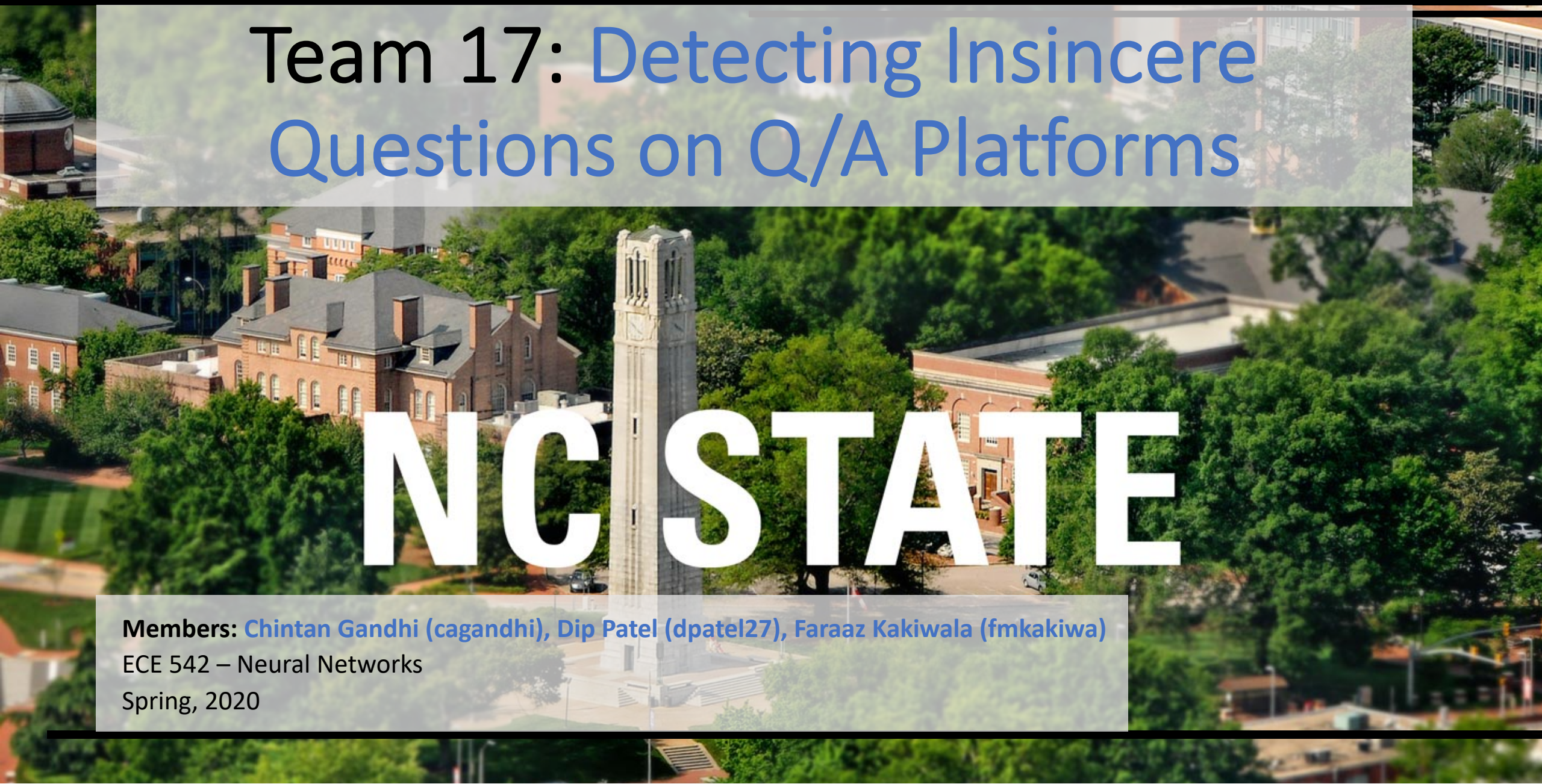
Team 17: Detecting Insincere Questions on Q/A Platforms



Members: Chintan Gandhi (cagandhi), Dip Patel (dpatel27), Faraaz Kakiwala (fmkakiwa)

ECE 542 – Neural Networks

Spring, 2020

Introduction

- Handling toxic questions is a burning problem for Q/A platforms.
- The questions that don't seek information but rather make statements, spread false information, founded upon false premises or attack a group of people are deemed insincere.
- Such questions need to be weeded out so that users can feel safe while sharing knowledge.
- The aim of our study is that given a question, we need to classify whether it is insincere or not (binary classification problem)

Challenges Faced

- The primary difficulty in detecting insincere questions is that the language used varies by region and people.
- Thus, the solution is not to merely build a list of stopwords.
- For example, "*Why do black people have black faces?*" v/s "*How can I remove black heads from my face?*".
- Both these questions employ similar words but their usage is completely different with the former being an insincere question.

Data Source

- Dataset provided on Kaggle challenge: *Quora Insincere Questions Classification*[1]
- Training data consists of 3 columns: Question ID, Question Text and the target (label) of the question (0-sincere / 1-insincere).
- Extreme imbalance in dataset as 1.2M sincere questions versus 80k insincere questions. Only 6% of training data are sincere questions.

0004a41beea5f02d85ef	What are some good songs for a long journey?	0
0004a7fcb2bf73076489	If blacks support school choice and mandatory sentencing for criminals why don't they vote Republican?	1
000500e5d543e112707e	What should be added to thrice the rational number -8/9 to get 4/7?	0

Data Pre-processing

- Removal of URLs
- Removal of punctuations
- Lemmatization

- **Original Text:** Why is Bannon trying to help Swedish democratic party (extreme right wing party)? https://www.dn.se/nyheter/varlden/bannon-weve-studied-the-sweden-democrats-for-a-while/
- **Removing URL:** Why is Bannon trying to help Swedish democratic party (extreme right wing party)?
- **Removing punctuations:** Why is Bannon trying to help Swedish democratic party extreme right wing party
- **Lemmatization:** Why be Bannon try to help Swedish democratic party extreme right wing party

Vector Embeddings

- GloVe[2] is an unsupervised algorithm to obtain vector representations for words.
- Projects the words into vector space such that similar words are closer to each other.
- We use Wikipedia 2014 + Gigaword 5 pre-trained model with 6B tokens.

vec(King) - vec(Man) + vec(Woman) = vec(Queen)

Bidirectional LSTM

- BiLSTM[3] is a special variant of RNN, which runs the input in 2 ways: one from past to future and other from future to past
- Combining the hidden states, we are able to preserve information about the past and the future at any point of time

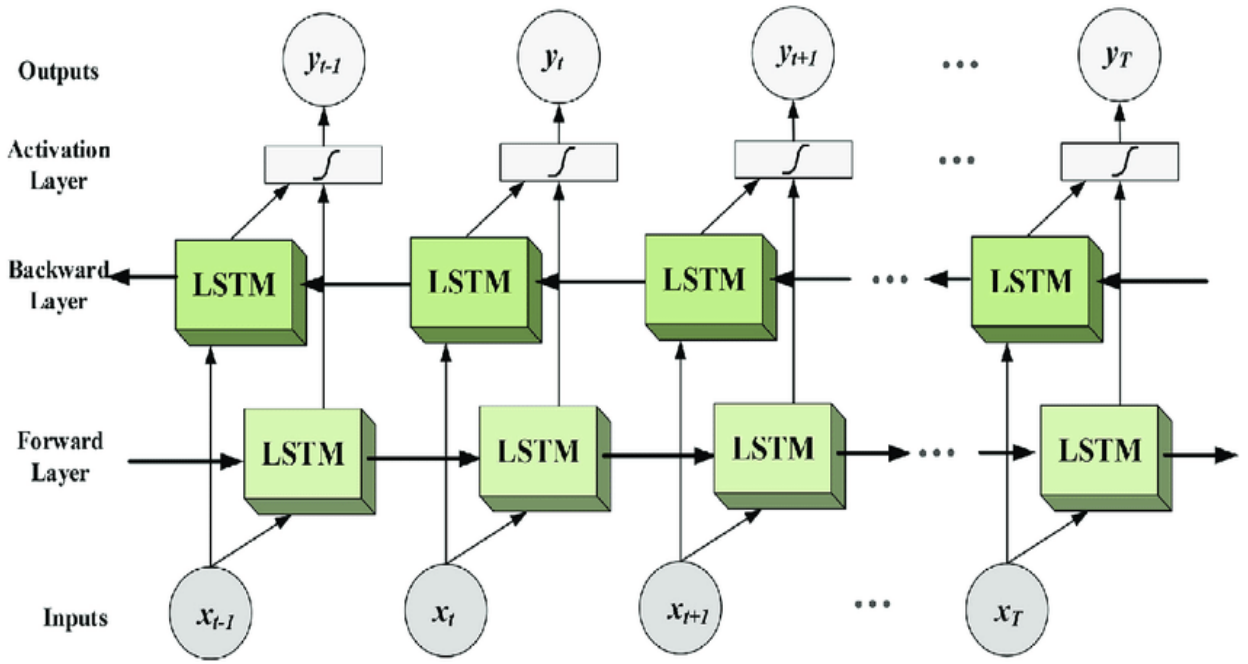
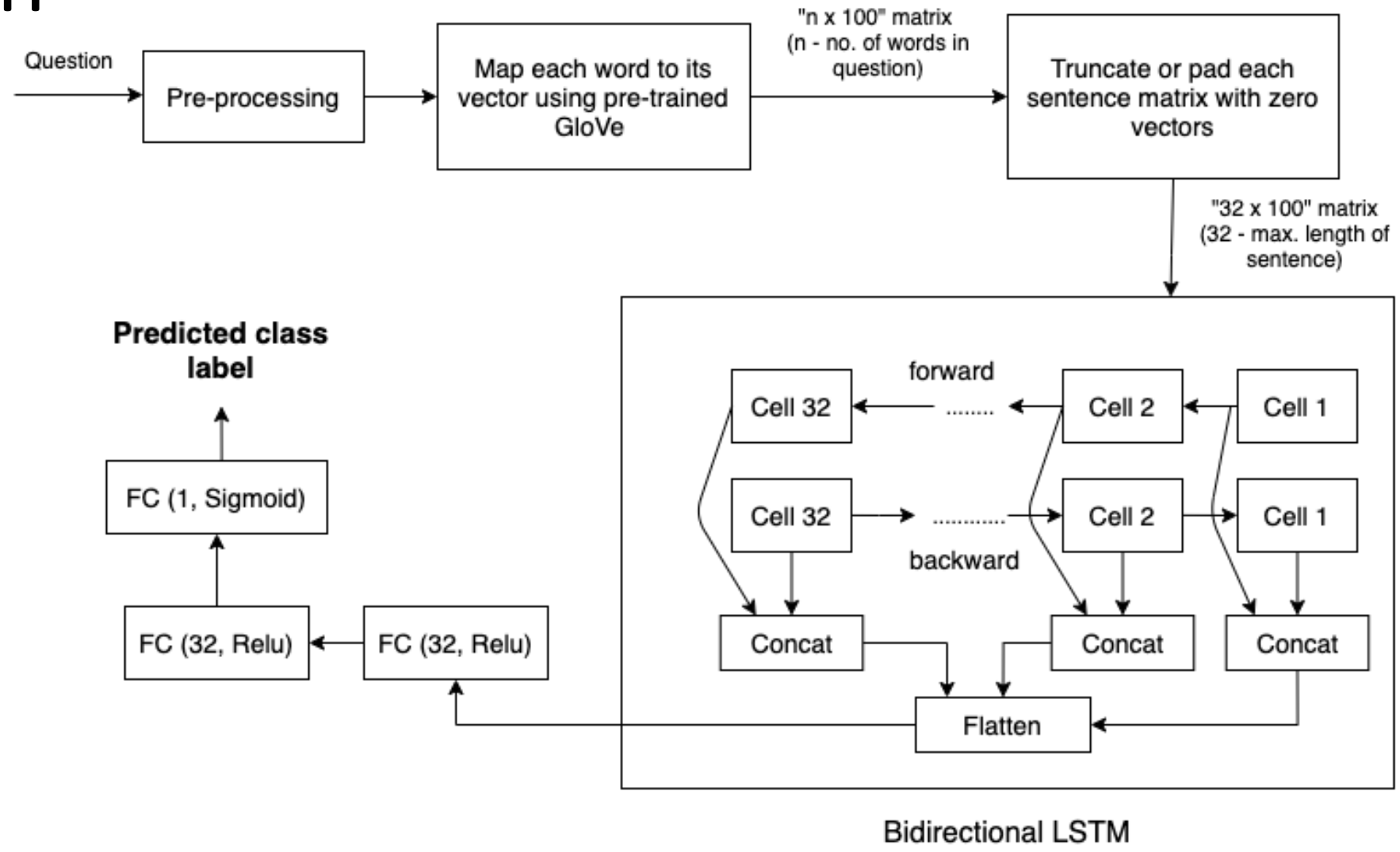


Fig: Basic Structure of the BiLSTM Network[8]

Approach


Question
Pre-processing
Map each word to its vector using pre-trained GloVe
"n x 100" matrix (n - no. of words in question)
Truncate or pad each sentence matrix with zero vectors
"32 x 100" matrix (32 - max. length of sentence)
Predicted class label
FC (1, Sigmoid)
FC (32, Relu)
FC (32, Relu)
forward
backward
Cell 32
Cell 2
Cell 1
Concat
Flatten
Bidirectional LSTM

Model Selection

- Model architecture was fixed at 128 units for RNN cell with the model trained for 10 epochs on a batch size of 64 and default Adam optimizer.
- We see that a bidirectional LSTM outperforms other RNN cell architectures.

RNN Architecture	Val. loss	Val. F1 score
GRU	0.2876	0.6195
LSTM	0.2522	0.6406
Bi-LSTM	**0.2446**	**0.6481**

Model Selection

- GloVe embedding + Bi-directional LSTM + 2 FC (32, RELU) layers
- Performed hyper-parameter tuning (10 epochs, "Adam" optimizer)
  - Batch Size
  - Number of units in LSTM cell
  - Learning Rate

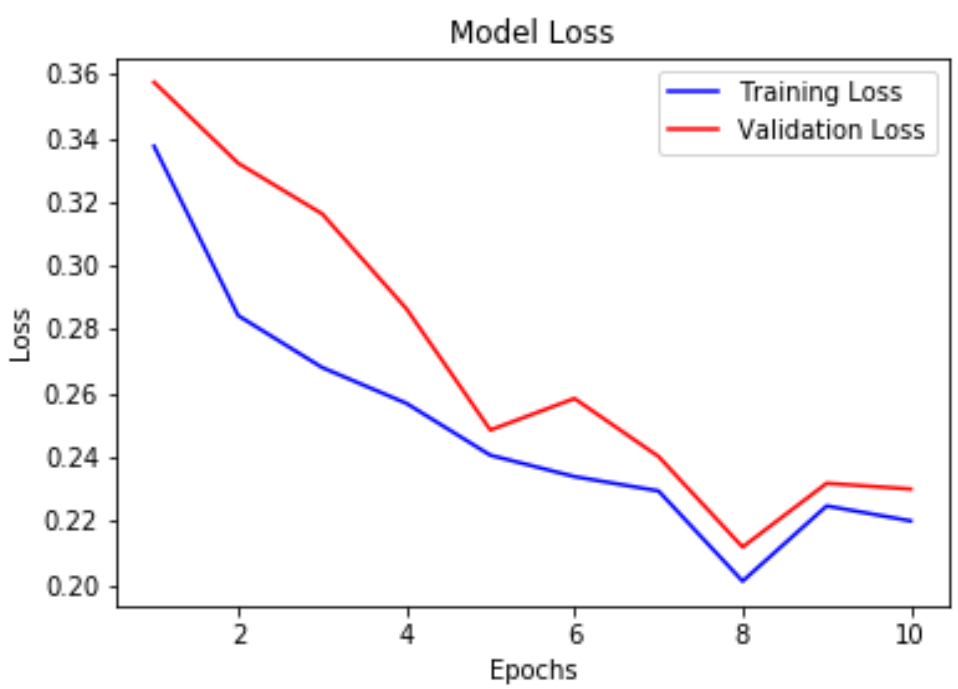
Batch Size	Val Loss	Val F1-Score
32	0.2534	0.6235
64	0.2446	0.6481
128	**0.2118**	**0.6757**
256	0.2634	0.6503

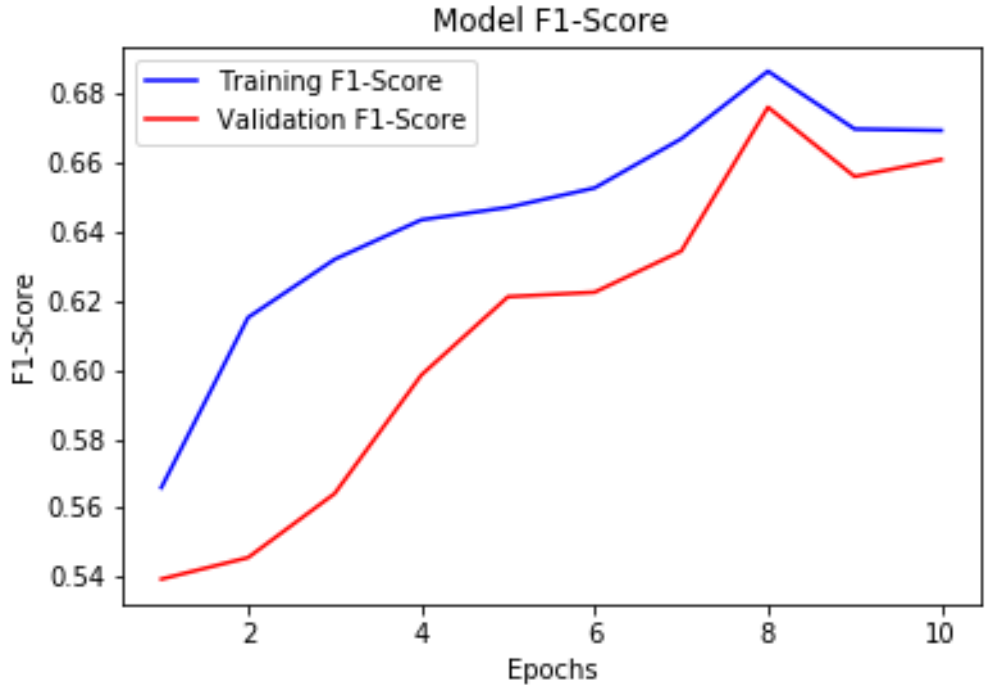
No. of LSTM units	Val Loss	Val F1-Score
64	0.2551	0.6298
128	**0.2118**	**0.6757**
256	0.2358	0.6648

Learning Rate	Val Loss	Val F1-Score
0.005	0.2634	0.6503
0.001	**0.2118**	**0.6757**
0.0005	0.2426	0.66
0.0001	0.2394	0.6606

Results

- **Final Model:** GloVe (100-D) - BiLSTM (128) - FC (32, RELU) - FC (32, RELU) - Output (Sigmoid)





Results

- Class 0 denotes sincere question and Class 1 denotes insincere question.
- We do not have access to labels of test data and hence, we use 20% training data as validation data to report the metrics.
- As we can see, the F1-score for our final approach is 0.67 with a recall of 0.91.
- Recall is more important since we need to make sure that we flag maximum possible insincere questions.

	Precision	Recall	F1-score	Support
0	0.99	0.89	0.93	122532
1	0.53	0.91	0.67	16162
Accuracy			0.89	138694
Macro avg	0.75	0.90	0.80	138694
Weighted avg	0.93	0.89	0.90	138694

Comparison with Baseline

- Our baseline model is a BoW (TF-IDF) model with a SVM classifier which achieves a F1-score of 0.5058 which is less than our best achieved F1-score of 0.67.
- True positives are less and False negatives are more for baseline than the final approach.

Actual	Predicted		
		0	**1**
	0	108891 (TN)	13461 (FP)
	1	1518 (FN)	14644 (TP)

Confusion Matrix for final approach

Actual	Predicted		
		0	**1**
	0	110744 (TN)	11787(FP)
	1	2880 (FN)	13282 (TP)

Confusion Matrix for baseline approach

References

[1] Quora Insincere Questions Classification Challenge on Kaggle. [https://www.kaggle.com/c/quora-insincere-questions-classification].

[2] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

[3] Zhiheng Huang, Wei Xu and Kai Yu. "Bidirectional LSTM-CRF Models for Sequence Tagging". ArXiv:1508.01991v1, 9 Aug 2015

[4] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

[5] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9, no. 8 (1997): 1735-1780.

[6] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014).

[7] Badjatiya, Pinkesh, Shashank Gupta, Manish Gupta, and Vasudeva Varma. ”Deep learning for hate speech detection in tweets.” In Proceedings of the 26th International Conference on World Wide Web Companion, pp. 759-760. 2017.

[8] Yildirim, Özal. (2018). "A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification". Computers in Biology and Medicine. 96. 10.1016/j.compbiomed.2018.03.016.

THANK YOU