Quora Insincere Questions Classification

Junbo Bao, Yabin Guo, Ka Hang Jacky Lok, Katrina Ni, Luna Shen

Simon Fraser University

CMPT 825

Motivation

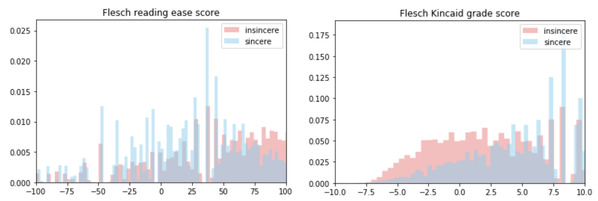
- Quora faces a key challenge to weed out insincere questions -- those founded upon false premises, or that intend to make a statement rather than look for helpful answers.
- The launch of this Kaggle competition, is to help Quora develop more scalable methods to detect toxic and misleading content, thus keeping its platform a place where users can feel safe sharing their knowledge with the world.

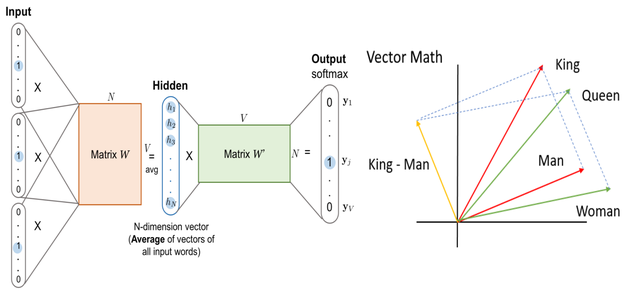
Experiment Setup

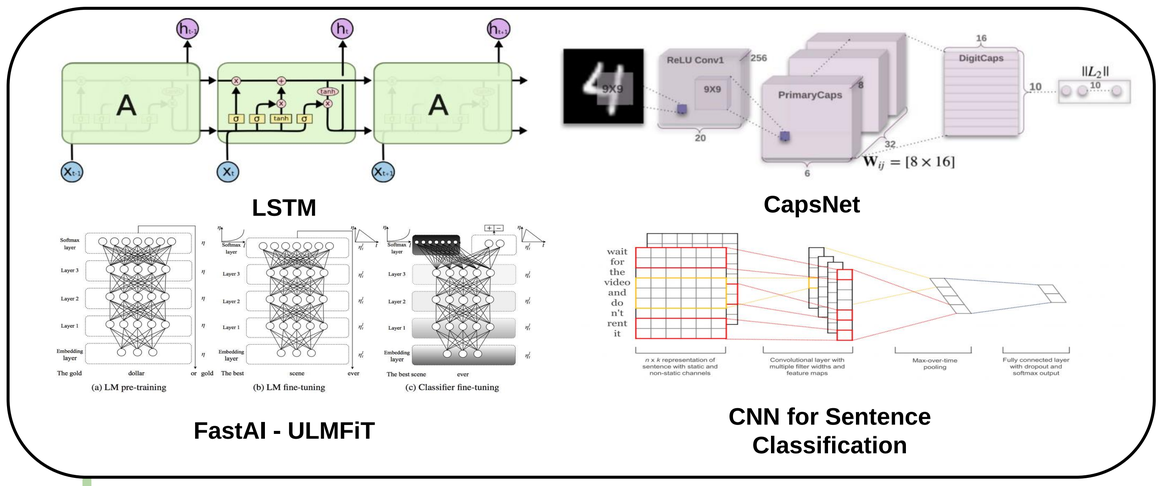
Data:
- Train set: 1306122 sentences, 6.19% positive
- Test set: 56370 sentences

Metrics: F1 Score

Runtime:
- Submissions must be made from Kaggle Kernels, with access to GPU
- Running time is limited to 2 hours to prevent over-complicated models





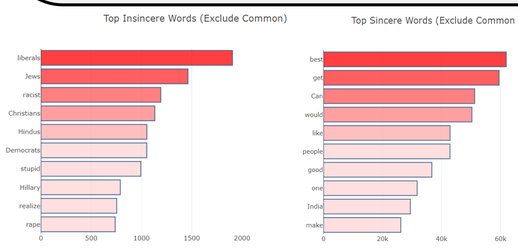


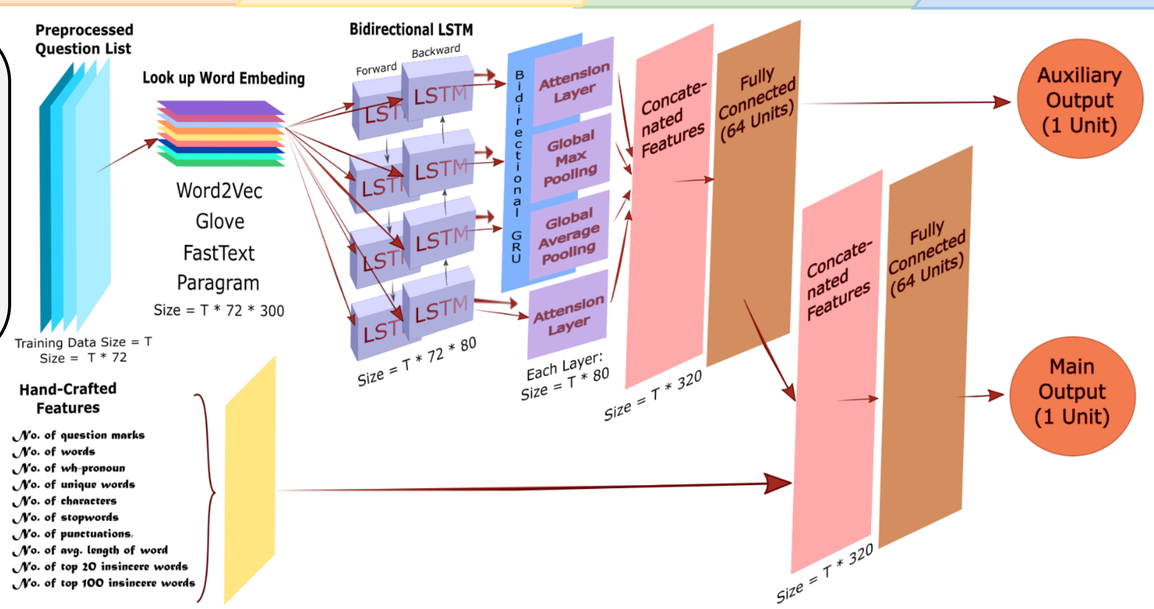
- **Preprocessed data**
  - Correct spelling (Ex: aple-> apple)
  - Split space-less string (Ex: Thisis-> This is)

- **Vector representations** of words, so that words with similar context occupy close spatial positions.
- **Word2Vec, GLOVE, FastText and Paragram**

Preprocessing → **EDA / Topic Modeling** → **Word Embeddings** → **DNN Models** → **Feature Engineering** → **Cross Validation & Ensemble**

- **EDA**
  - Distribution of target values and question length
  - Readability test: Flesch reading ease score and Flesch Kincaid grade score
  - Bad word analysis - Gether banned word list from Google, Facebook -> detect relationship between bad word and targets
- Performed **LDA topic modeling** on insincere questions and sincere questions separately
- Used **perplexity and coherence** to find the best number of topics





- **Cross Validation**: Stratified K Fold
- **Simple Ensembling**: Majority Voting, Averaging
- **Stacking**: Logistic Regression, Random Forest, XGBoost, Multilayer Perceptron

- **10 hand-crafted features**: # question marks, # words, # wh-pronoun, # unique words, # characters, # stopwords, # punctuations, avg. length of word, # top 20 insincere words, # top 100 insincere words.
- **20 topic model factors**.

Experiment Results

Setup ID	Input	Embedding	Model	F1 Score on Validation Set	F1 Score on Test Set	Running Time
0 (baseline)	Original text	Word2Vec	CNN	0.674	0.649	38m
1	Preprocessed	Word2Vec	CNN	0.681	0.654	40m
2	Preprocessed	WikiText-103	ULMFiT	0.651	-	8h
3	Preprocessed	GLOVE	LSTM	0.688	0.675	35m
4	Preprocessed + 10 new features	GLOVE	LSTM	0.689	0.677	40m
5	Preprocessed + topic factors	GLOVE	LSTM	0.692	0.679	1h50m
6	Preprocessed + topic factors	GLOVE	LSTM + CNN	0.694	0.685	57m
7	Preprocessed + 10 new features	GLOVE + Paragram	GRU + CapsNet + cross validation	-	0.690	1h30m

Conclusion and Future Work

- Preprocessing the question text leads to an improvement of 0.005 in f1 score on testset
- In this particular task, GLOVE embedding works best among four embeddings
- Both new numerical feature and topic modeling have positive effect on the f1 score
- With a single model and cross validation, we achieved an **f1 score of 0.690** on Kaggle public leaderboard (top 20%). (**Recall: 74.5%; precision: 63.0%**)
- Todo: try more complex model if there is no 2 hour running time limitation
- Todo: implement the model with Tensorflow instead of Keras to reduce randomness