Quora insincere question classification

Dozent: Stefan Langer

Referentin: Ane Berasategi

28. Januar 2019

Index

kaggle™

Quora

1. Introduction

- **Motivation for this topic**
  - Kaggle competition in NLP classification
  - Difficult task but numerous helpful notebooks available

https://www.kaggle.com/c/quora-insincere-questions-classification

1. Introduction

- **Motivation for this topic**
  - Kaggle competition in NLP classification
  - Difficult task but numerous helpful notebooks available
- **Problem**
  - *"An existential problem for any major website today is how to handle toxic and divisive content."*
  - *"…predict whether a question asked on Quora is sincere or not."*

https://www.kaggle.com/c/quora-insincere-questions-classification

Insincerity

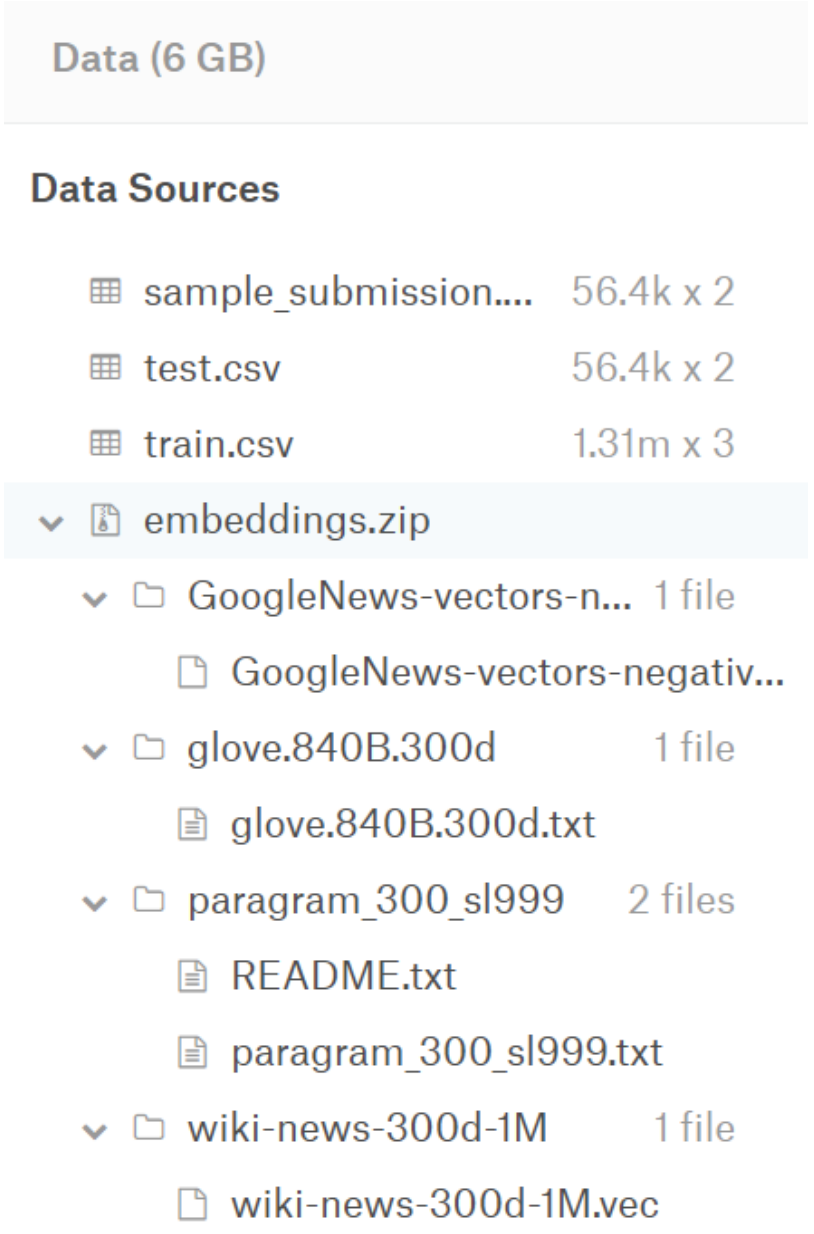
- What constitutes an insincere question?
  - Non-neutral tone: exaggerated tone, rhetorical and implies a statement.
  - Disparaging or inflammatory: attacks/insults, based on an outlandish premise.
  - Isn’t grounded in reality: based on false information, contains absurd assumptions.
  - Uses sexual content for shock value.

Insincerity

- What constitutes an insincere question?
  - Non-neutral tone: exaggerated tone, rhetorical and implies a statement.
  - Disparaging or inflammatory: attacks/insults, based on an outlandish premise.
  - Isn't grounded in reality: based on false information, contains absurd assumptions.
  - Uses sexual content for shock value.
- Examples
  - *"Why is Quora so stupid?"*
  - *"When will Pakistan return Pakistan back to India?*
  - *"Why are religious people so rude?"*
  - *"Why did Brian leave me?"*
  - *"Why are Germans still allowed to have their own country?"*
  - *"How do I teleport from Earth to the Moon?"*
  - And many more…

Data

- train.csv: the training set
  - *qid*: unique question identifier
  - *question_text*: Quora question text
  - *target*: 1 if the question is insincere, 0 otherwise
- test.csv: the test set
- sample_submission.csv
  - sample submission in the correct format
- embeddings/
  - GoogleNews-vectors-negative300 (word2vec)
  - glove.840B.300d
  - paragram_300_sl999
  - wiki-news-300d-1M (fasttext)



Metrics and constraints

- Evaluation metric: F1- score between the predicted and the observed targets
- The code (script/notebook) must be run in the Kaggle website
  - The code must run in <= 2h
  - Disk constraint: 5.2GB
  - RAM constraint: 14GB
  - GPU support (Nvidia Tesla K80)
  - No access to external data
- There is some noise in the training set
  - Some questions are mislabelled → performance won't be excellent

Index

kaggle™

Quora

2. EDA (exploratory data analysis)

- Analyse files, sizes, columns
- Analyse class imbalance between sincere and insincere questions
- N-gram analysis
- Analyse hyperparameters: number of words, number of characters…
- Manual analysis of “extreme” words
  - people, everyone, all, no one, always, never…
  - fight, punish, insult, ruin, castrate…
  - so, such, much, more, than, most…

https://www.kaggle.com/anebzt/quora-eda/

2. EDA recap

- 1.3M training examples (6% of which insincere), 56k test examples
- N-gram analysis
  - The most frequent trigrams appear only 30-40 times (out of 1.3M examples)
- Insincerity depends on context and use of neutral words in a negative way.
- Therefore, pre-processing and architecture are quite general, and focus on the context.

Index

3. Pre-processing

- Prepare data in appropriate format to feed to the model
- Prepare embeddings
  - The 4 embeddings have similar performance
  - Common approach is to average 2+ embeddings
- Approximate the vocabulary to the embeddings as much as possible
  - Otherwise the embedding will not recognize many words → we lose information
- Treat contractions, punctuations, misspellings…

3. Pre-processing recap

- Embeddings are loaded
- 99.56% of all data is covered by the embedding
  - Further correcting misspellings slightly improves performance
- Data is split into train/val sets and shuffled
- Data is in the appropriate format to be fed to a NN

(https://www.kaggle.com/maxjeblick helped me reach these conclusions)

Index

4. Model

- **4.1. Prepare metric**
  - F1 score (trade-off between precision and recall)
- **4.2. Prepare model**
  - LSTM + Attention (good if we want to look for words in context)
- **4.3. Train**
  - Average embeddings (glove and param)
  - Train
  - (Wait 50min)
  - Find best threshold for F1 score

https://www.kaggle.com/anebzt/quora-preprocessing-model

4. Model recap

- Metric: F1
- Average embeddings between GloVe and Param
- Model: Bidirectional LSTM with one Attention layer
- Performance during training: highest F1 score: 0.678
  - top 57% (not so good)
  - Best result in competition is F1 score = 0.712
  - Out of ~3600 teams participating, ~3200 have F1 score > 0.5

5. Conclusion

- **Possible improvements**
  - Add features from the EDA
    - Mean question length
    - Number of extreme words
  - Change architecture
    - Modify LSTM model
    - GRU with Attention
    - Bidirectional RNN/CNN
    - Stratified Kfold (1)
    - Cyclical learning rates (CLR) (2)

(1): Optimizing for Generalization in Machine Learning with Cross-Validation Gradients, Barratt & Sharma, 2018 https://arxiv.org/abs/1805.07072
(2): Cyclical learning rates, Smith, 2015 https://arxiv.org/abs/1506.01186

Quora insincere question classification

Dozent: Stefan Langer

Referentin: Ane Berasategi

28. Januar 2019