Quora Insincere Questions Classification
Detect toxic content to improve online conversations

Haixuan Guo

Introduction

Quora is a website that encourages people to ask questions and learn from each other. However, some insincere questions makes users feel uncomfortable.

An **insincere question** is intended to make a statement rather than look for helpful answers.

- Has a non-neutral tone,
- Is disparaging or inflammatory
- Isn't grounded in reality
- Uses sexual content

Objective

- Identify insincere questions and predict unclassified data.
- Compare the performance of different machine learning algorithms .
- Understand text mining and employ the state-of-art word vectorization techiniques.

Orginality

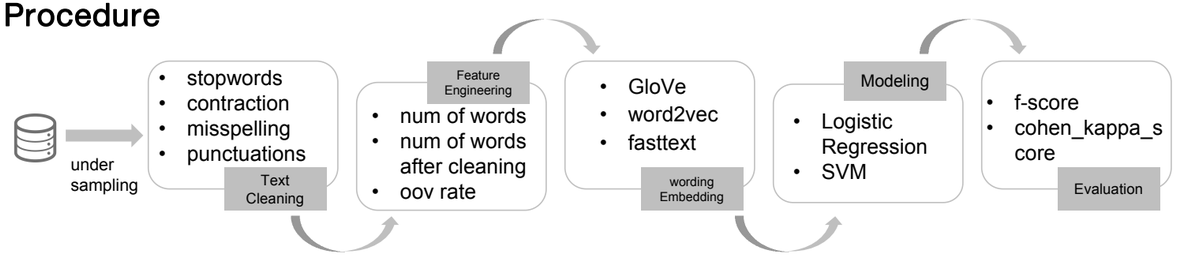
- Use explainable algorithms for this problem instead of NN like most people did on kaggle
- Evaluate word embedding methods
- Employ PCA in text classification problems

Learning Opportunity

Imbalanced data
Text cleaning
Word embedding
Classification algorithms, LR, SVM, NB
Classification_report for evaluation
Visualizing data by matplotlib

Methods and Materials

- **Procedure**

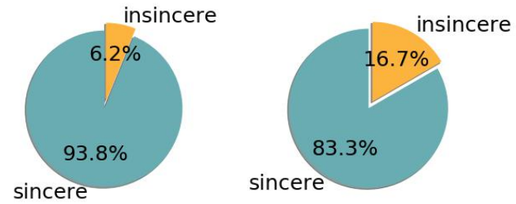


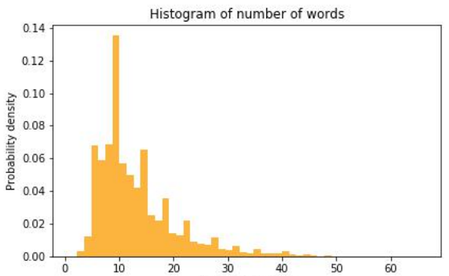
- **Text Processing**

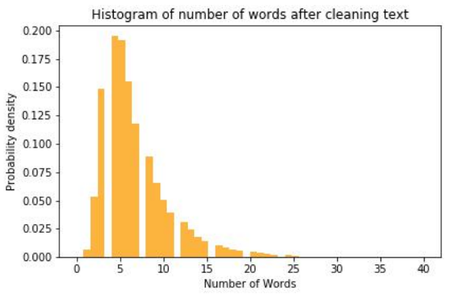
Text	'Whyyyyyyyyy the h*ck do Brits end every sentence with "x"?'
Text cleaning	'why the hack do brits end every sentence with x'
Tokenization	['why', 'the', 'hack', 'do', 'brits', 'end', 'every', 'sentence', 'with', 'x']
Vectorization	**brits**: [-0.332 , -0.04526 , 0.2734 , -0.4175 , -0.6636 , -0.6313 ,0.516 , -0.1755 , -0.0501 , -0.09076 , -0.2147 , -0.06027 ,0.0715 , -0.405 , 0.3286 , -0.3267 , -0.3318 , -1.055 , ...]

Data Exploration

The dataset is from Kaggle
The target value is labeled by human.

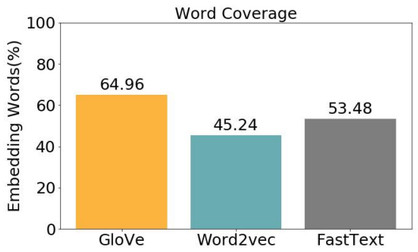




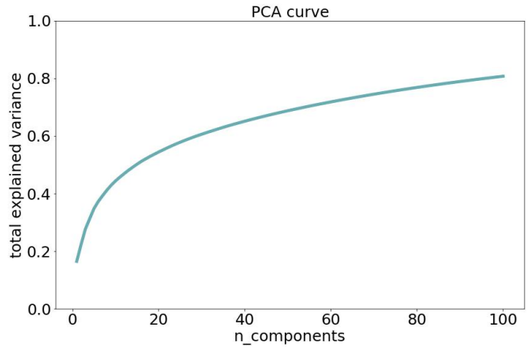


Results

Word Embedding Comparision



Modeling



	F-score	cohen_kappa_score
Logistic Regression	0.54	0.5097
SVM	0.58	0.6083
Naive Bayes	0.49	0.4840
SVM + PCA	0.55	0.3976

Conclusion

- GloVe has the widest coverage of vocabulary, Fasttest is fast of vetorizing words.
- SVM outperfroms Logistic Rgression and Naive Bayes
- After reducing demension by PCA, the performance of SVM model deteriorates.