Kaggle Competition: Quora Question Pairs

Applied Machine Learning 2017

Melania Berbatova, Sophie Kamuf

Introduction

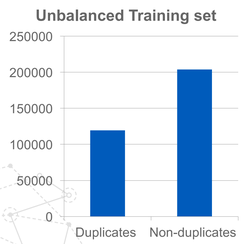
Goal

- Identify duplicate Quora questions with maximum accuracy

Data

- Training data: 323 325 labeled pairs
- Test data: 80 960 unlabeled pairs

Question Pair statistics	
Average words in a question	9
Average length of a question	50
Average # of common words	4
Range of common words	0 - 34
Range of length difference	0 - 872



Data Processing

Text cleaning

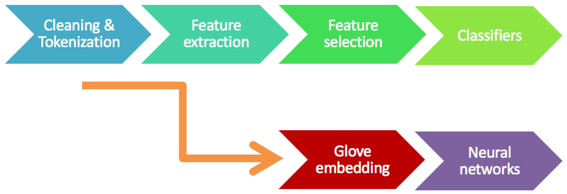
- Remove numbers, punctuation and non-ascii symbols
- Remove stop words and replace misspelled words
- Tokenizing for word embeddings
- Drop NaN questions from Training set

Feature extraction

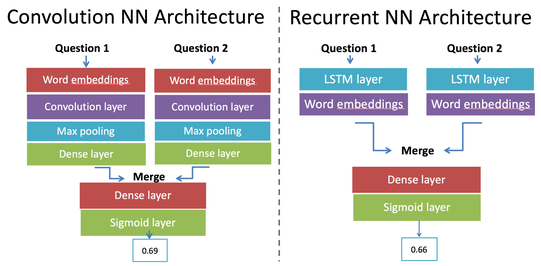
- Feature set 1 (FS1) – Basic Features: Number of words, length of question, difference of questions' lengths, number of common words
- Feature set 2 (FS2) - Fuzzy features : Q-ratio, W-ratio, partial ratio, partial token set ratio, partial token sort ratio, token set ratio, token sort ratio
- Feature set 3 (FS3) – Word2vec distances and statistics: Euclidean, Manhattan, Euclidean, Canberra, Minkowski and Braycurtis distances, cosine similarity, Scew and Kurtosis statistics
- Feature set 4 (FS4) – Same starting word, number of unique words

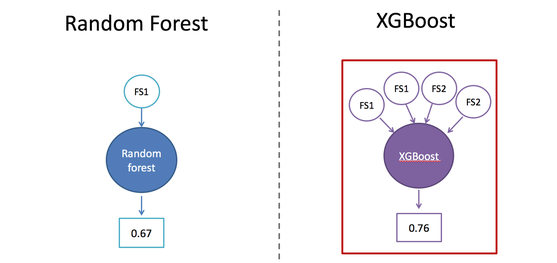
Data Analysis

Process Pipeline

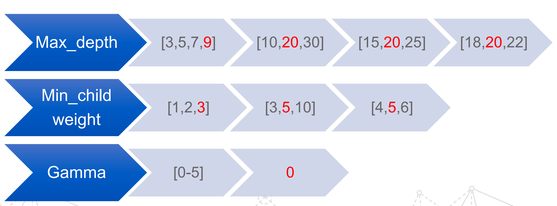


Model Selection





Model Training: XGBoost Parameter Tuning

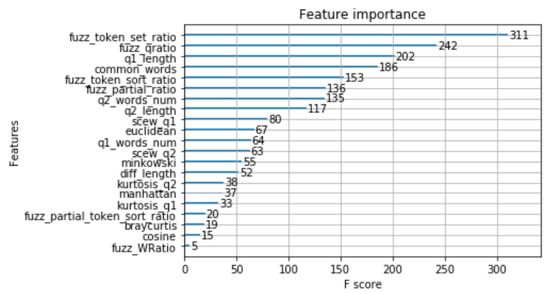


Results

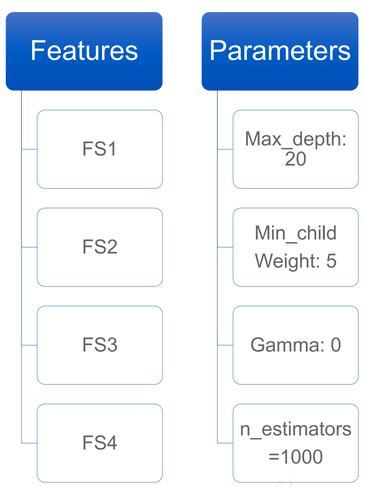
Model Evaluation

- **Kaggle Rank: 9**
- **Accuracy: 0.765**
- **Precision: 0.692**
- **Recall: 0.674**
- **F1 score: 0.683**

XGBoost Feature Importance

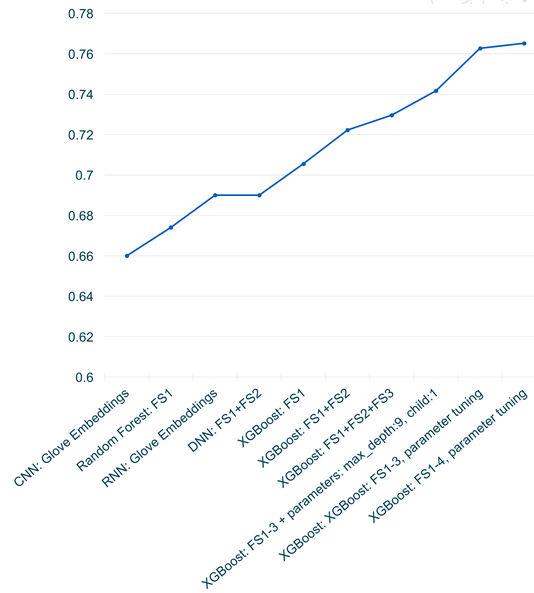


Final XGBoost Model



Conclusion

Accuracy Improvement



References

1. Kaggle user Currie32. "The importance of cleaning text". March 2017. goo.gl/fomeRv
2. Kaggle user Anocas. "Data Analysis & XGBoost Starter". March 2017. goo.gl/3oNZM7
3. Abhishek Thakur. "Is That A Duplicate Quora Question?". February 2017. goo.gl/jz1MPp
4. Hubert Lin. "Kaggle Competition: Quora Question Pairs". June 2017. goo.gl/ccbnqp
5. Kaggle user SRK, "Keras Starter Script with Word Embeddings". February 2017. goo.gl/1SHWRg
6. Kaggle user Lystdo. "LSTM with word2vec embeddings". April 2017. goo.gl/LzCVbo
7. Christopher Olah, "Understanding LSTM Networks". August 27, 2015. goo.gl/hnryCq