Reddit Author Identification with Deep Learning

April Chung, Kendra Gedney, Kate Schulz

ANLY-590 Fall 2018 | MS Analytics | Georgetown University | Dr. Hines

Introduction

Our problem is single-label **multi-classification for 100 authors**, for which we trained a baseline support vector machine (SVM) and four deep learning models.

Since the dataset is cross-topic, we also tested the **impact of removing named-entities.**

Dataset

We collected our own dataset from Reddit.

- 72.6k comments from 100 randomly selected authors from a popular subreddit
- Between 506 and 902 comments per author
- Between 60 and 1,000 characters per comment
- Cross topic: between 8 and 172 subreddits per author

For Experiment 2, we subsetted the above to just 14 authors and 10k comments

Models

We fit a LSTM, CNN, CNN + LSTM, and Simple Pooling Model

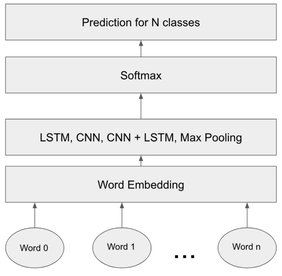


Figure 1: Generalized Model Diagram

Results

Experiment 1: Can we identify the authors?

- The best model was the Simple Pooling model
- The LSTM model overfit, as evidenced by the gap in training and test accuracy, while the CNN models plateaued
- We measured accuracy up to k=10, which means the correct author was within the first k predicted

Results

Model	Training Accuracy	Test Accuracy	Test Accuracy (k=5)	Test Accuracy (k=10)
SVM	**0.99**	0.38	0.59	0.68
LSTM	0.91	0.24	0.44	0.55
CNN	0.50	0.23	0.45	0.58
CNN + LSTM	0.52	0.24	0.47	0.59
Simple Pooling	0.89	**0.42**	**0.64**	**0.74**

Figure 2: Histogram showing distribution of test accuracy by author from the Simple Pooling model. Accuracy rates ranged from 12%-76%.

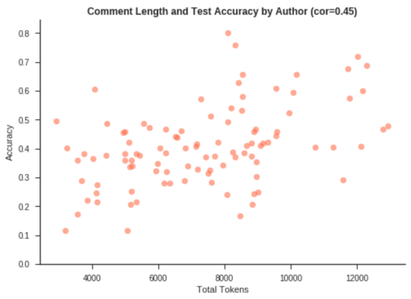


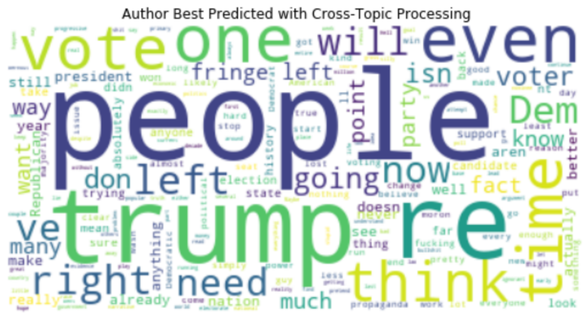
Figure 3: Accuracy increases for more authors who write longer comments for the Simple Pooling model.

Experiment 2: How do named-entities affect results?

- Replacing named-entities, using Stanford's NER, decreased accuracy rates for all models.

Results (Test Accuracy at k=1)

Model	Named-Entites Removed	Named-Entities Intact
SVM	0.55	0.58
LSTM	0.37	0.46
CNN	0.38	0.48
CNN + LSTM	0.34	0.44
Simple Pooling	**0.51**	**0.63**



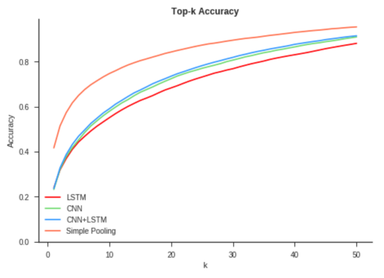


Figure 4: All models achieve high accuracy when looking at top-k, which is useful in author identification applications.

Conclusions

- Though this was a difficult task, **the Simple Pooling model was the best**
- Although accuracy rates at k=1 may seem low, models can reduce pool of potential authors
- In this domain, named-entity extraction did not improve results, which differs from *Markov, et al.*

Reference: Markov, I., Stamatatos, E., & Sidorov, G. (2017, April). Improving Cross-Topic Authorship Attribution: The Role of Pre-Processing.