CS 229 Final Project:

Authorship Attribution with Limited Text on Twitter

Luke Chen, Eric Gonzalez, Coline Nantermoz

{lukech18, ejgonz, coline}@stanford.edu

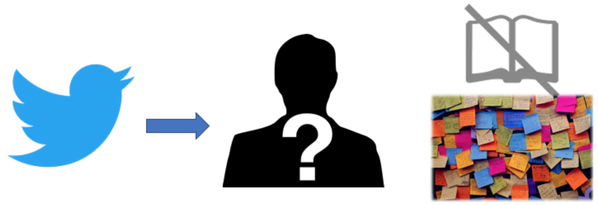
Introduction and Motivation

Question: Can we apply authorship attribution to identify writing by the general public, even with limited data?

➢The rise of social media has allowed billions of people to post their views on a day-to-day basis.

➢Challenges: short, informal and sparse data

➢Applications: personal identity verification, forensics



Dataset and Feature Selection

Dataset:

➢ Dataset of pre-collected tweets: 545 politicians, around 3,200 tweets each [1]

➢ Cleaned and preprocessed data to remove retweets and nonstandard characters, labeled examples

➢ Used a subset of 6 authors for classification

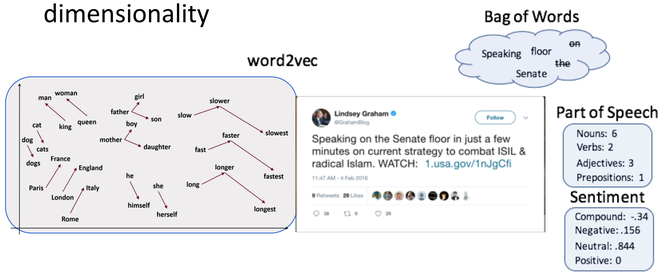
Features:

➢ Lexical, syntactic, and semantic features:

- ➢ **Word frequency (unigram bag of words)**
- ➢ **Part of speech frequency (UPenn Tagset)**
- ➢ **Overall sentiment of tweet (VADER)**

➢ Refined lexical features: removed common stop words

➢ Experimented to replace bag of words features with word embeddings (word2vec) in order to reduce dimensionality



Model Selection

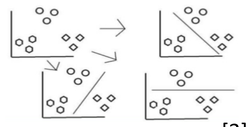
We chose to evaluate 3 types of models:

➢**Naïve Bayes**

- Classic authorship attribution model, used for *Federalist Papers* authorship. [2]
- Multiclass multinomial classifier

➢**SVM**

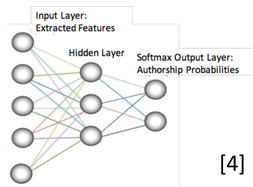
- One-versus-rest model, generates one classifier for each author
- Optimized using L2 regularization, tried RBF kernel for nonlinear SVM



[3]

➢**Neural Network**

- Fully connected feedforward network with a ReLU activation hidden layer and a final softmax output layer
- Optimized using grid search to tune number of layers, number of units per layer, batch size, optimizer



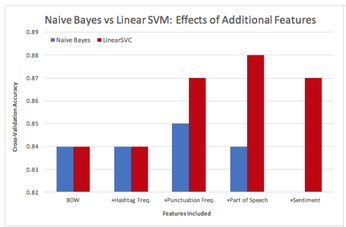
[4]

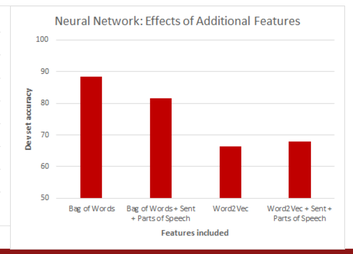
Experiments and Results

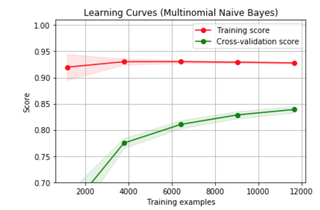
➢ We tested the models above across multiple features to classify 6 authors, using a training set of 12,004 tweets and test set of 2,573 tweets. The best performance achieved by each model is summarized in the table below.

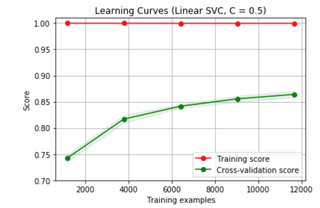
Model	Features	Feature Vector Size	Training Accuracy	Test Accuracy
Naïve Bayes	BOW + Parts of Speech	25231	0.932	0.844
SVM (Linear)	BOW + Parts of Speech + Sentiment	25235	1	0.878
SVM (Gaussian)	W2V + Parts of speech + Sentiment	442	0.996	0.710
Neural Network (1 HL, 100 neurons, ReLU)	BOW	25187	1	0.890

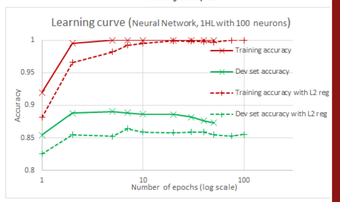
➢ We observed the effects of adding additional features to the various models. Sentiment was not evaluated using NB as it violates the model assumptions. Word2vec features were analyzed using SVM with a nonlinear (Gaussian) kernel and a single-hidden-layer neural network.











Analysis

Naïve Bayes:

➢ The baseline with bag of words achieves higher than expected accuracy, although it decays with more authors.

➢ The model works slightly better even with additional features that are not conditionally independent.

SVM:

➢ Base bag of words features achieves higher training and test accuracy than Naïve Bayes baseline.

➢ Regularization and additional features also bring up accuracy noticeably compared to Naïve Bayes counterparts.

Neural Network:

➢ Achieves perfect training accuracy and the highest test accuracy, with just bag of words features.

➢ Difficulty in hyperparameter tuning (number of layers, of units, regularization): differences weren't significant

Conclusions and Future Work

Conclusions:

➢ An author attribution model for limited text, evaluated on Twitter messages, achieves over 85% categorical accuracy using Naïve Bayes, Linear SVM, and feedforward neural network approaches, with up to 6 authors.

➢ Bag of words features set a reasonable baseline. Additional standard NLP syntactic and semantic features (part of speech, overall sentiment) slightly help accuracy for Naive Bayes and SVM (but not neural networks).

➢ Word embeddings decrease accuracy.

➢ More powerful models (neural networks) do perform better, but not outstandingly so.

Future Work:

➢ Additional model hyperparameter tuning (regularization, network architecture) to improve accuracy using word2vec

➢ More semantic processing: handle negations and phrases

➢ Evaluate the effect of the number of authors on performance

References

1) Dataset: reddit. (2017). *Over one million tweets collected from US Politicians (President, Congress and Governors)*. [online] Available at: https://www.reddit.com/r/datasets/comments/6fniik/over_one_million_ tweets_collected_ from_us/ [Accessed 11 Dec. 2017].

2) Jockers, M. and Witten, D. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25(2), pp.215-223

3) Adapted from Arora, V. (2015). *Binary Class and Multi Class Strategies for Machine Learning.*

4) Adapted from https://sites.google.com/site/mrstevensonstechclassroom/hl-topics-only/4a-robotics-ai/