Who said dat?
Classifying and clustering tweets by authorship

Niklas Bostrom, Jason Ge, Frank Zhou
Harvard University, CS 182, John A. Paulson School of Engineering and Applied Sciences, Cambridge MA 02138, USA

ABSTRACT

- Natural language processing has been an important field in the realm of AI given its varied applications to everyday communication
- Historically, even simple "bag-of-words" models have had good success with tasks such as spam filtering
- Today, the Internet has spawned many new modes of online communication, such as Twitter
- Our goal is to apply machine learning techniques to tweets from various famous personas (e.g. Donald Trump, Elon Musk, Kim Kardashian) in order to be able to 1.) classify unfamiliar tweets and 2.) identify clusters in existing tweets
- We developed a novel feature vector to characterize tweets, including features such as "word-similarity" (determined through a bag-of-words model), "part of speech-similarity", "length of tweet" and so on
- Our Naive Bayes classification algorithm is able to yield an overall 59.6% success rate in assigning authorship to unfamiliar tweets when "word-similarity" is weighted at 0.9 in the feature vector
- Current work is focused on implementing a k-nearest neighbors algorithm, a k-means clustering algorithm, and on using hyperparameters to optimize the relative weights of features in the feature vector

OUTLINE

Goal:

To differentiate unfamiliar tweets by authorship using a Naive Bayes classifier

Method:

- Collect and clean tweet data from 5 authors (n = 52000, training set = 26000)
- Bag-of-Words:
  - Count all words for all people and generate probabilities
  - Implement bag-of-words based on words themselves and their parts-of-speech
- Find average word length, number of hashtags used, and fraction of misspellings per tweet
- Weight all features, pass in a new tweet, and run to generate a prediction

RESULTS

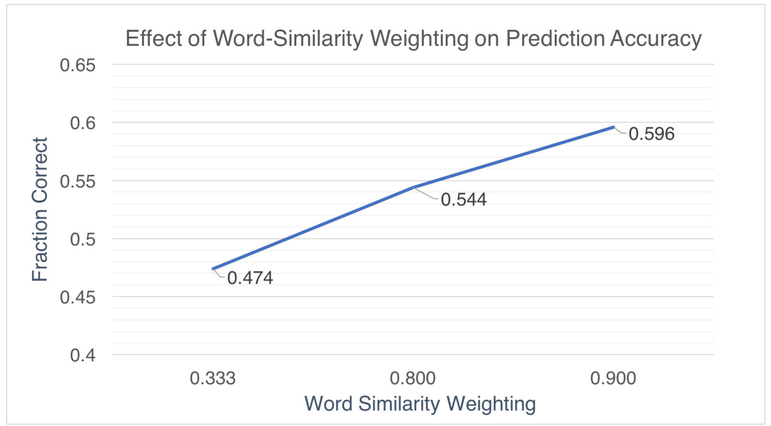
- We passed this unfamiliar tweet into our Naive Bayes classifier
- We tried four different weightings (A–D) of the features in our feature vector

	Weighting			
Feature	**(A)**	**(B)**	**(C)**	**(D)**
Word similarity	0.25	1.0	0.0	0.5
Part of speech similarity	0.25	0.0	1.0	0.5
Average word length	0.25	0.0	0.0	0
Number of misspellings	0.25	0.0	0.0	0

	Prediction %			
Person	**(A)**	**(B)**	**(C)**	**(D)**
E. Musk	12.4	0.0	11.0	5.5
N. Tyson	10.9	0.0	0.1	0.1
K. Kardashian	8.6	0.2	4.2	2.2
B. Obama	10.9	0.1	0.0	0.1
D. Trump	57.3	99.6	84.7	92.1

- We ran our classifier on 500 test tweets for our four different weightings (A–D)
- Word similarity and part of speech similarity are both good predictors of authorship
- The other two features do not have much predictive power

	(A)	(B)	(C)	(D)
Overall predictive accuracy	47.8	69.2	41.4	71.6



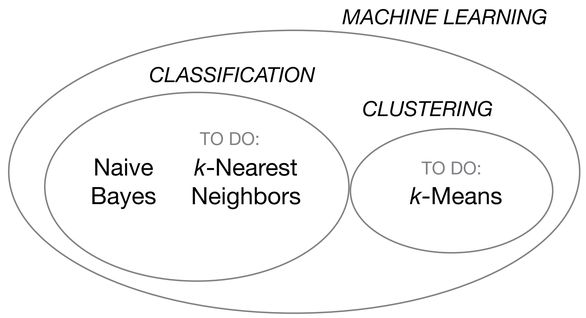
- We ran our Naive Bayes classifier on 500 unfamiliar tweets
- We varied the weighting of the "word similarity" feature
- The weightings of all other features were evenly distributed to overall sum to 1
- As "word similarity" weighting increased, prediction accuracy also increased

CONCLUSIONS

- A simple bag-of-words model for classifying tweets is surprisingly quite effective at correctly predicting authorship of unfamiliar tweets
- In fact, bag-of-words is probably more important than the other features, but we haven't yet figured out how to weight all of our features to increase overall and individual-level tweet-prediction accuracy
- Considering the parts of speech of words in a tweet increases the overall classification efficacy (71.6%) when compared to just considering the words themselves alone (69.2%)

Future directions

- Incorporate more features such as number of mentions and number of links
- Implement and train hyperparameters in order to optimize the weights of features in the feature vector
- Finish implementing a k-nearest-neighbors classifier and compare its efficacy with our Naive Bayes classifier
- Finish implmenting a k-means clustering algorithm and run it to identify patterns in tweet data



ACKNOWLEDGMENTS

We would like to thank Emily Wang and Brian Plancher for their advice and support.

REFERENCES

AdhokshajaPradeep. "President Obama." Kaggle, 2016, www.kaggle.com/adhok93/president-obama/data.

Brown, Brendan. "Trump Twitter Archive." Trump Twitter Archive, 2016, trumptwitterarchive.com/.

Larsen, Liam. "Elon Musk Tweets, 2010 to 2017." Kaggle, 23 Apr. 2017, www.kaggle.com/kingburrito666/elon-musk-tweets/data.