Hw1: Presidential Speech Classification with Word Embeddings

Sam Showalter
University of California, Irvine (showalte)
showalte@uci.edu

Abstract

this is a testThis document contains the instructions for preparing a manuscript for the proceedings of ACL 2020. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used for both papers submitted for review and for final versions of accepted papers. Authors are asked to conform to all the directions reported in this document.

1 Introduction

2 Supervised Learning

2.1 Feature Engineering

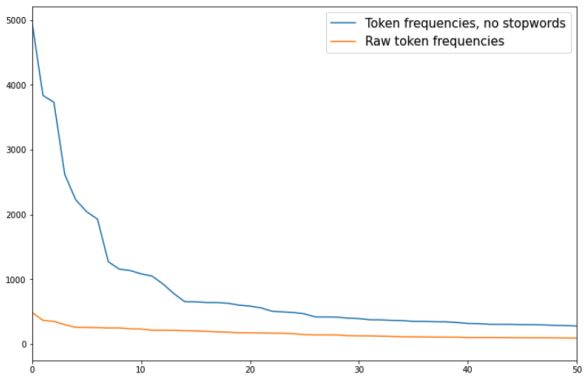

Figure 1: Token frequencies in training data before and after stopwords removed from the dataset

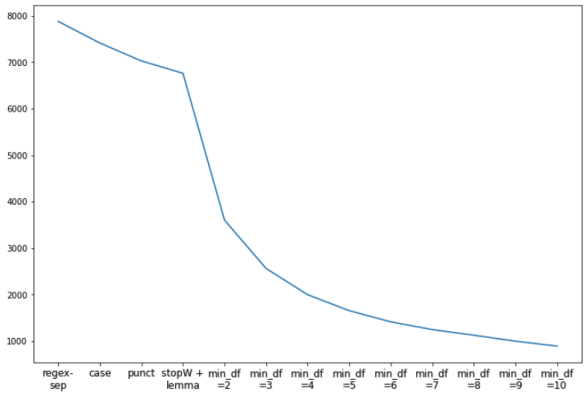

Figure 2: Vocabulary size as successive filters applied to input corpus. From left to right, the strings are regex separated, then case and punctuation is resolved, with final filters trimming rare tokens

Table 1: Feature Engineering Grid Search

		Language Feature Filters			
Token Engine	Tok. Sep.	None	Case+ Punct.	Lem.	Stop Word
CVect	nltk	0.399	**0.435**	**0.435**	0.401
CVect	reg	0.396	0.413	0.413	0.399
Tf-Idf	nltk	0.382	0.382	0.382	0.386
Tf-Idf	reg	0.360	0.374	0.374	0.381

2.2 Dimensionality Reduction

this should word (Pennington et al., 2014) this should word

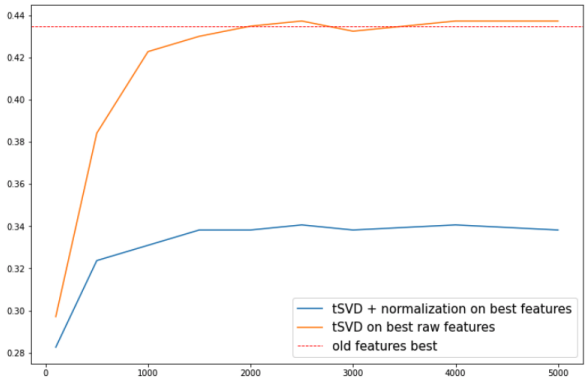

Figure 3: Classification performance applied to best feature selection after Truncated SVD applied with varying numbers of components

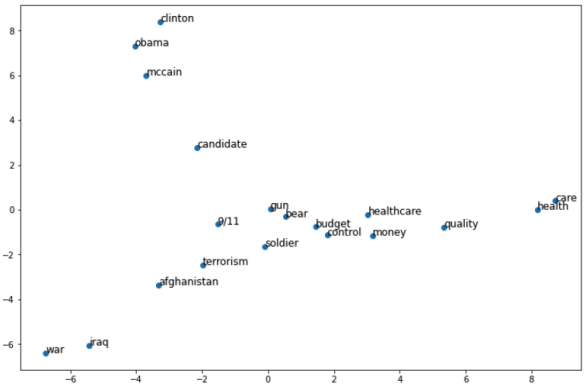

Figure 4: Visualization of Word2Vec embeddings with PCA with a selection of common political terms. Embeddings appear to capture some level of semantic meaning

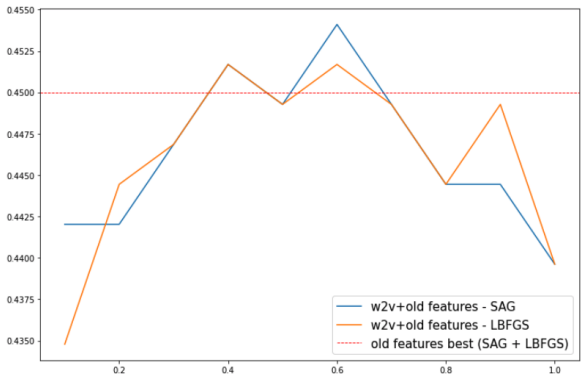

Figure 5: Classification performance of Word2Vec document embeddings with and without original SVD features concatenated. The x-axis represents the fraction of unlabeled documents leveraged to make the Word2Vec model

Table 2: Model Tuning: Logistic Regression

	Logistic Reg. Solver				
Reg. Pen.	lbfgs	lib-linear	newton	saga	sag
-	0.399	-	0.401	**0.454**	0.447
L1	-	0.423	-	0.428	-
L2	0.447	0.432	0.450	0.447	**0.450**

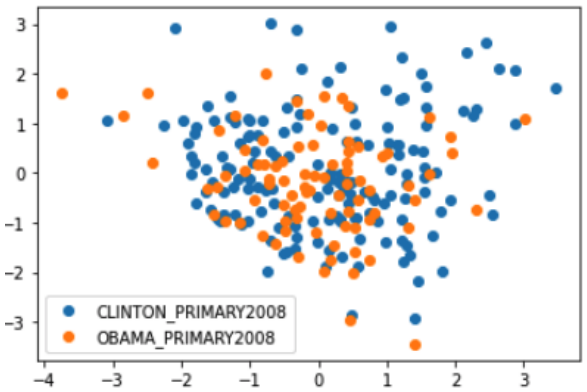

Figure 6: Visualization of political candidate embedding profiles as defined by their speech embeddings. Little separation can be cleaned, likely because order is not considered in embedding creation

2.3 Model Tuning

3 Semi-supervised Learning with Word Embeddings

3.1 Word2Vec for Text Classification

4 Experimental Results and Discussion

5 Conclusions and Further Exploration

References

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

A Statement of Collaboration