The Center for Cyber Defenders
Expanding Computer Security Knowledge

SAND2014-15957D

Valence Bag-Of-Words Text Classifier

Project Mentor: Jeremy Wendt: 5632 | Zachary Benz: 5635

Johahn Wu | University of Texas at Austin

Problem Statement:

The bag-of-words (BOW) classification approach is a simple, commonly implemented model. However it often cannot achieve accuracies over 80%.

Objective:

Improve BOW model utilizing the Laplacian Smoothing algorithm to spread valence through pre-processing and post-processing of corpus terms.

Approach:

Implemented 4 BOW classifiers:

1. *Standard*: Basic BOW implementation with Porter Stemming and Single Word Occurrence Deletion. Supports Term Occurrence, Term Frequency (TF), TF-IDF, and Log Entropy as term weights.
2. *ModTuple*: Extends Standard. Looks for modifier words (e.g., not, very, really) and adds the modifier and the following word pair as a unique term.
3. *SubAdd*: Extends Standard. Subtracts or adds from a term's frequency depending on the value of any modifier words preceding it (e.g., "not good" results in a -1 TF value for "good").
4. *Tuple n*: Extends Standard. Adds top n 2-tuples to list of unique terms.

Testing:

- 10-fold Cross Validation on a movie review corpus of 1000 positive and 1000 negative documents.
- Random sampling for varying size of labeling set

Impact and Benefits:

Unfortunately, none of the modifications made to the BOW model made a significant difference to accuracy. However, between stemming and single occurrence deletions, execution time and classifier size were improved.

Results:

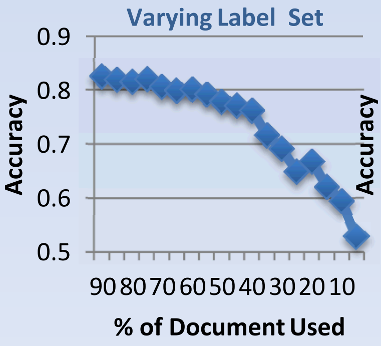
Term Weights: Using this corpus, Term Occurrence outperformed TF, which outperformed both TF-IDF and Log Entropy.

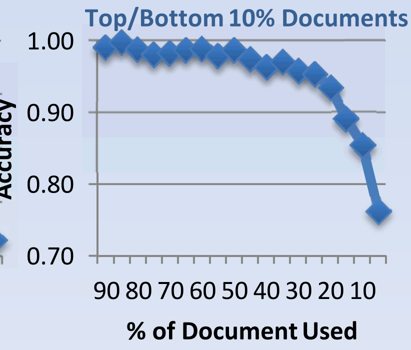
	Accuracy	Precision	Recall	F1 Score
Occ	0.835	0.864	0.794	0.828
TF	0.826	0.848	0.796	0.821
TF-IDF	0.796	0.815	0.765	0.789
Log	0.795	0.815	0.762	0.788

Bag-of-Words: The highest performing BOW models were ModTuple and Tuple 375. Since the n value of Tuple is extremely corpus dependent, utilizing ModTuple when using other corpora might be advantageous.

	Accuracy	Precision	Recall	F1 Score
Tuple 375	0.850	0.874	0.818	0.845
Mod	0.847	0.871	0.815	0.842
SubAdd	0.840	0.864	0.795	0.828
Standard	0.840	0.864	0.794	0.828

Varying Label Set: Often the corpus being analyzed does not have 90% of its documents labeled, so the corpus was tested utilizing varying label set sizes. If only the top and bottom results matter, the accuracy becomes nearly perfect.





Sandia National Laboratories operated by LOCKHEED MARTIN
Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin company, for the U.S. Department of Energy's National Nuclear Security Administration under con-