SAND2017-7537C

The Center for Cyber Defenders

Expanding computer security knowledge

Exploring Parameter Adjustments in Text Classifiers

Melissa Bain, Carleton College

Project Mentors: Tom Brounstein, Org. 5853 and Seth Decker, Org. 5852

Problem Statement

We want to be able to better classify text documents. Specifically, we used posts surrounding the topic of abortion to classify each stance as pro-choice or pro-life.

Objectives and Approach:

We explored the K Nearest Neighbor and Random Forest Algorithms, using the F Measure to determine success.

- Frequency Counts
- Log Entropy
- TFIDF

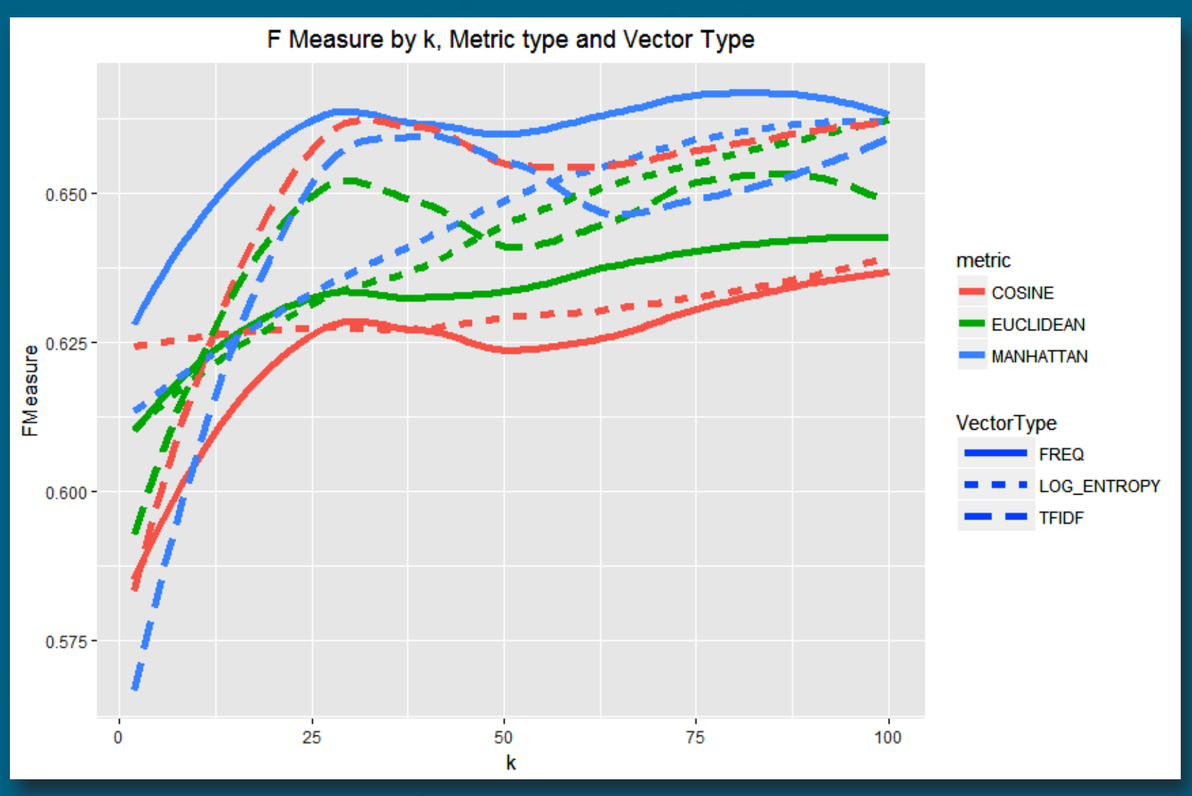
KNN Parameters:

- Number of Neighbors (k)
- Distance Measure
  - Euclidean
  - Manhattan
  - Cosine

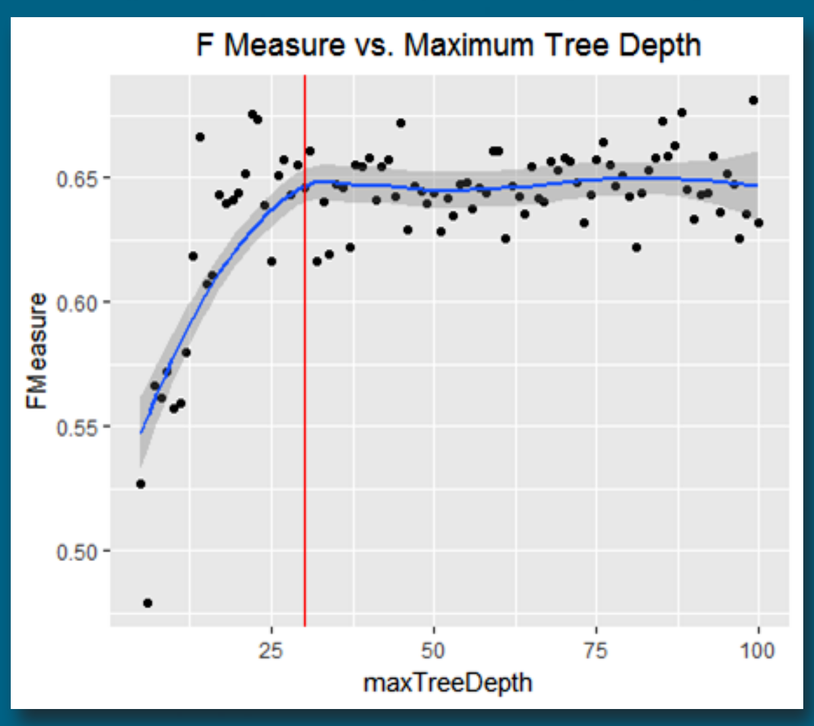
Random Forest Parameters:

- Max Tree Depth
- Ensemble Size
- Bagging Fraction
- Dimensions Fraction

KNN works best with more than 25 neighbors. TFIDF is the most consistent vectorization method, but Frequency works well with Manhattan distance. Cosine distance should only be used with log entropy vectorization.



Random Forest works best with a max tree depth of at least 30.



Conclusions:

KNN outperforms **Random Forest** with a higher mean F Measure and smaller variance.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2017-XXXXC.