CS6120 – NLP

Spring 2022

Final Project

Group Members: Kimberly Boynton O’Brien, Bill Collins

Russian Troll Tweet Identification

Introduction

What is the task

Why is it relevant

How will we approach it

Data

Source and Description

2,973,371 tweets from 2,848 Twitter handles from a Russian Troll Factory

Troll Tweet dataset: <https://github.com/fivethirtyeight/russian-troll-tweets>

Normal Tweet dataset: <https://archive.org/details/twitter_cikm_2010>

Surface observations

Hypotheses about key features

Cleaning and processing to get features

Data Statistics

Stats and charts about these features showing differences between the classes

Data Representation for Models

Will try different types along dimensions: text processing and vectorization methods

Text processing

tokenization, lemmas, POS

Vectorization

**Word Level**

Onehot, Wordvectors

**Document Level**

count vectorizer, tfidf, LSA, weighted wordvecs, docvecs , FSE

Visuals

T-SNE of Wordvecs

T-SNE of Document Vectors

Models

We will try different variations along dimensions: model type, metaparameters

Hypotheses about models

Model Training

Data will be split into train/cross validation/test sets of size:

Metaparameters: batch size, learning rate, optimizer, early stopping, boosting ensembles etc

Perceptron, logistic regression, SVM , MLP, LSTM, CNN

Model Visualization

learning curves, learned embeddings

Reflection

Explain strengths and weaknesses of each feature type with each model.

Evaluation of hypotheses

Evaluation of hardest data

Evaluation of key features via model (logistic regression)

Possible Improvements

Key takeaways and conclusion

Acknowledgements/Citations

https://fivethirtyeight.com/features/why-were-sharing-3-million-russian-troll-tweets/

Z. Cheng, J. Caverlee, and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In Proceeding of the 19th ACM Conference on Information and Knowledge Management (CIKM), Toronto, Oct 2010. (Bibtex)