# Neural Network Initialized with a Decision Tree for Document Categorization

Ethan Fulton, Dr. Girard

## Problem Description

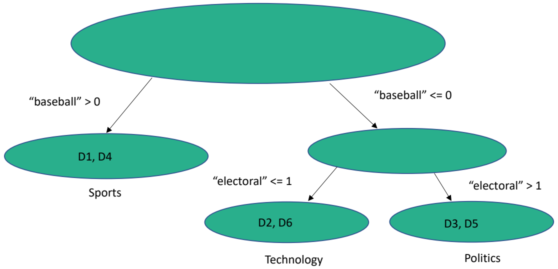
How does the accuracy of a Neural Network initialized with a Decision Tree (NNIDT) compare to the results found by Justin Rebok using Deep Belief Networks and how does the introduction of stemming affect the accuracy?

## Literature Review Highlights

- Porter's Stemmer
- TF-IDF important words
- Decision Tree / Information gain
- Neural Network Initialized with a Decision Tree (NNIDT)

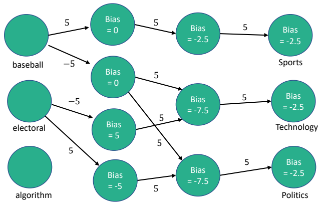
| | |
|---|---|
| *PRIVATE -> C* | *PRIVATE -> CVCV* |
| *PRIVATE -> CV* | *PRIVATE -> CVCVC* |
| *PRIVATE -> CVC* | *PRIVATE -> CVCVCV* |

Porter's Stemmer consonant and vowel grouping for pattern evaluation.



Decision tree creating paths from root to leaves by using rules that split into a more homogeneous subgroup.

$$(baseball > 0) \vee (baseball \leq 0 \wedge electoral \leq 1) \vee (baseball \leq 0 \wedge electoral > 1)$$



First hidden layer constructed from literals in DNF and second hidden layer constructed from conjuncts of DNF. All other edges set to +/- β.

## Primary Objective

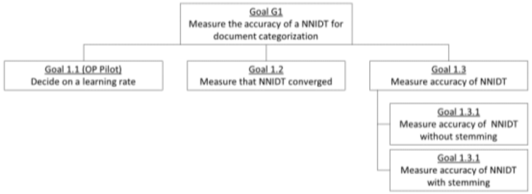
To compare the classification method presented in "Text Categorization Using Neural Networks Initialized with Decision Trees" (N. Remeikis, I. Skucas, and V. Melninkaite) for categorizing articles with the method used by Justin Rebok in "Deep Belief Network to Classify Documents". (1.5 person weeks over 1 semester)

## Solution Description

- Tools: C, Linux Environment
- BBC Dataset – 2225 documents in 5 categories
- Create tool to calculate 50-most important words per category
- Select Samples as per experiment design
- Create DNF from decision tree
- Initialize Neural Network with DNF
- Train NNIDT with Back Propagation

## Hypotheses

- The NNIDT will perform as well as a Deep Belief Network with three RBMs.
- Introducing stemming before the 50 most important words are picked will increase the accuracy of the NNIDT.



Goal G1
Measure the accuracy of a NNIDT for document categorization

Goal 1.1 (OP Pilot)
Decide on a learning rate

Goal 1.2
Measure that NNIDT converged

Goal 1.3
Measure accuracy of NNIDT

Goal 1.3.1
Measure accuracy of NNIDT without stemming

Goal 1.3.1
Measure accuracy of NNIDT with stemming

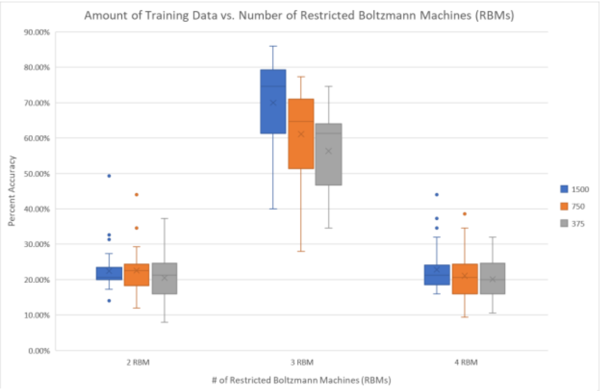
## Experiment Design

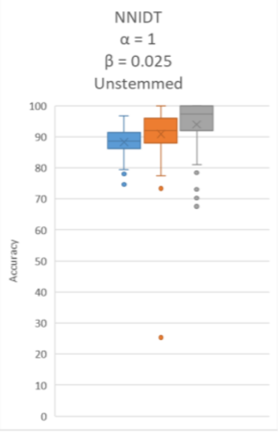
| Block Design | | Stemming | |
|---|---|---|---|
| | | Yes | No |
| Number of Documents in Sample | 1500 total – 1350 training, 150 test | X | X |
| | 750 total – 675 training, 75 test | X | X |
| | 375 total – 338 training, 37 test | X | X |

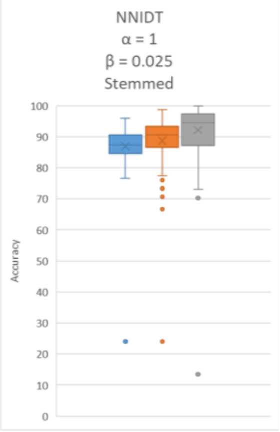
* This t-test would be considered inaccurate if the samples tested were all the same, which is a possibility due to the randomization of sample input during testing.

## Results

NNIDT performed significantly better than a DBN with t-values of -42.896 for 1500 documents , -64.1647 for 750 documents and -87.2938 for 375 documents , but stemming had no effect with p-values of 0.1115, 0.0796 and 0.1611 at a 95% confidence interval.*


Amount of Training Data vs. Number of Restricted Boltzmann Machines (RBMs)


NNIDT
α = 1
β = 0.025
Unstemmed


NNIDT
α = 1
β = 0.025
Stemmed

## Future Work

- Adjust α and β values to tune the NNIDT for faster convergence
- Modify NNIDT to not use words that do not appear in DNF from decision tree as input