

Classification of Political News with NLP

Magdaleno, Alejandro
Lyles-Woods, Quin'darius

Kennesaw State University

December 1, 2021



Outline

- 1 Proposed Research
- 2 Goals
- 3 Relevance
- 4 Data
 - Data Source
 - Tools
 - Stages
- 5 Models and Algorithms
 - Options to Implement
 - Model Structure
- 6 Results
- 7 Questions and Answers



Proposed Research

- The study of an aggregation of news articles.



Proposed Research

- The study of an aggregation of news articles.
- Using various Natural Language Processing Techniques.



Proposed Research

- The study of an aggregation of news articles.
- Using various Natural Language Processing Techniques.
- Detection of bias with news outlets for the reader.



Goals

- To determine where bodies of text come from algorithmically.



Goals

- To determine where bodies of text come from algorithmically.
- To have a success rate about 70% with said goal.



Goals

- To determine where bodies of text come from algorithmically.
- To have a success rate about 70% with said goal.
- To understand the success and failures and reaching the goal.



Relevance

- Uncovering bias may get us closer to the truth.



Relevance

- Uncovering bias may get us closer to the truth.
- Will be one of the first tools to give a metric to news articles.



Relevance

- Uncovering bias may get us closer to the truth.
- Will be one of the first tools to give a metric to news articles.
- If utilized correctly could be used to save some time when there is more subjective bias than there is fact.



Data about our Data

- Approximately 30,000 articles from CNN and Fox News.



Data about our Data

- Approximately 30,000 articles from CNN and Fox News.
- Each article is self contained within a text file.



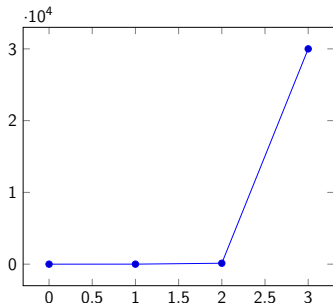
Data about our Data

- Approximately 30,000 articles from CNN and Fox News.
- Each article is self contained within a text file.
- All articles are from the politics section.



Data about our Data

- Approximately 30,000 articles from CNN and Fox News.
- Each article is self contained within a text file.
- All articles are from the politics section.
- Contains about a year of news articles from both institutions.



Data Source

The Data was sourced from CNN and Fox News Sites directly. There was no data source appropriate for our means so we gathered it ourselves. The sites presented their own individual challenges.



Tools

- Curl is a tool for transferring data from or to a server.



Tools

- Curl is a tool for transferring data from or to a server.
- Grep searches for PATTERNS in each FILE. We used it for its ability to search via REGEX expressions.



Tools

- Curl is a tool for transferring data from or to a server.
- Grep searches for PATTERNS in each FILE. We used it for its ability to search via REGEX expressions.
- Pipes are a builtin within most shells allowing for manipulation of text streams.



Tools

- Curl is a tool for transferring data from or to a server.
- Grep searches for PATTERNS in each FILE. We used it for its ability to search via REGEX expressions.
- Pipes are a builtin within most shells allowing for manipulation of text streams.
- Cat allows for the catenation of files.



Tools

- Curl is a tool for transferring data from or to a server.
- Grep searches for PATTERNS in each FILE. We used it for its ability to search via REGEX expressions.
- Pipes are a builtin within most shells allowing for manipulation of text streams.
- Cat allows for the catenation of files.
- Pup is a tool used for parsing and extracting information from html.



Tools

- Curl is a tool for transferring data from or to a server.
- Grep searches for PATTERNS in each FILE. We used it for its ability to search via REGEX expressions.
- Pipes are a builtin within most shells allowing for manipulation of text streams.
- Cat allows for the catenation of files.
- Pup is a tool used for parsing and extracting information from html.
- Node is a JavaScript runtime that allows for JavaScript to be ran on a server.



Stages

- Constructing a file with all the links to the respective articles.
 - ZSH Shell Scripting to algorithmically call websites.



Stages

- Constructing a file with all the links to the respective articles.
 - ZSH Shell Scripting to algorithmically call websites.
- Downloading all the HTML articles from the links.
 - Using the wget (similar to curl) to download each article from text file.



Stages

- Constructing a file with all the links to the respective articles.
 - ZSH Shell Scripting to algorithmically call websites.
- Downloading all the HTML articles from the links.
 - Using the wget (similar to curl) to download each article from text file.
- Parsing the data for only the article bodies.
 - Using pup to get the article body and writing the output to a correct file.



Stages

- Constructing a file with all the links to the respective articles.
 - ZSH Shell Scripting to algorithmically call websites.
- Downloading all the HTML articles from the links.
 - Using the wget (similar to curl) to download each article from text file.
- Parsing the data for only the article bodies.
 - Using pup to get the article body and writing the output to a correct file.
- Cleaning up the folder structure.



Models and Algorithms

Tensor Flow is utilized initially to create TF and IDF for the algorithm. Then the model takes in the raw text files and puts them together vectorize the text into integer vectors. This is utilized with a Sequential Feed-Forward Network to make the base model.



Options to Implement

- TF-IDF



Options to Implement

- TF-IDF
- FastText



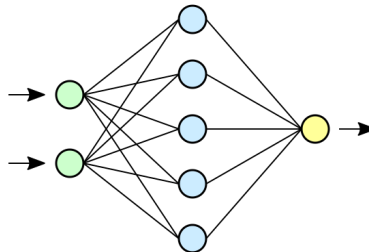
Options to Implement

- TF-IDF
- FastText
- Deep Learning
 - Binary
 - Multiclass



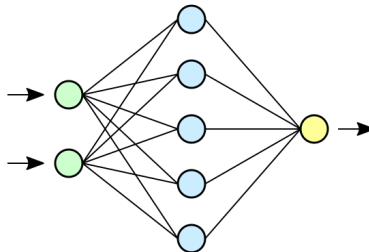
Model Structure

- Embedding Layers



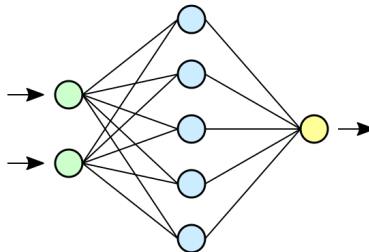
Model Structure

- Embedding Layers
- Dropout Layer



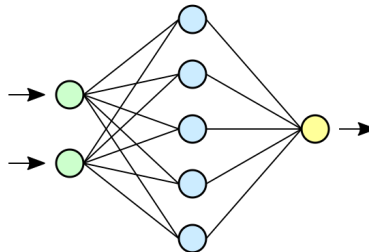
Model Structure

- Embedding Layers
- Dropout Layer
- Dense Layer



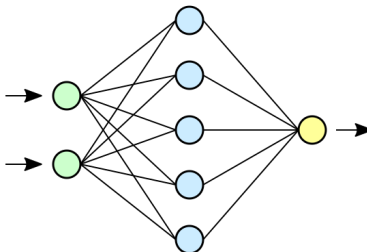
Model Structure

- Embedding Layers
- Dropout Layer
- Dense Layer
- Global Average



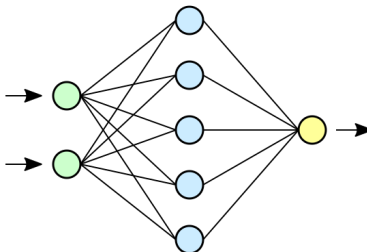
Model Structure

- Embedding Layers
- Dropout Layer
- Dense Layer
- Global Average
- Dense



Model Structure

- Embedding Layers
- Dropout Layer
- Dense Layer
- Global Average
- Dense
- Final Dense



Results

We ran the model for at most 25 epoch. The loss values started very high around 40 percent but as the model ran it went below zero for the loss function and thus the accuracy was around the same and achieving around 99.96 percent accuracy.



Questions and Answers

Please ask some questions about anything related to the project.

