Executive Summary

Predicting Media Bias in News Articles

Melissa Lee

Data Scientist

BrainStation - Data Science P/T

Email: melissa@leemail.ca

LinkedIn: www.linkedin.com/melissa-ws-lee

13th, May 2023

This executive summary report was prepared and submitted by Melissa Lee on May 13th, 2023 to fulfill BrainStation's Data Science Diploma Capstone requirement.

BrainStation®

Introduction

This report outlines an executive summary detailing the main problem statement, some context and background on some of the historical work that has been conducted on this subject area, limitations, methodology, description of the dataset and definitions, the cleaning and preprocessing process, modeling, results, and conclusions. Please see full code on the attached Jupyter notebook.

Background on the Subject Matter

In modern day news media, there is great concern around misinformation, propaganda, and media bias that is polarizing our communities, especially as it relates to politics. The growing emergence of the natural language processing field within data science and its applications have demonstrated impressive abilities to "discover hidden patterns and capture linguistic regularities that reflect human biases, such as racism, sexism, and ableism" (Caliskan et. al, 2017). Word embeddings are numerical representations that are able to capture the grammatical and semantic meaning of the text. We can analyze the patterns in word embeddings using NLP techniques to better understand how bias is formed in the human language. There have been many studies conducted by researchers and academics on how bias and unfairness is shaped in text at various granular levels including at a word, sentence, and paragraph level (Chen et. al, 2020). As a result, for example, applications like bias or fact checkers have emerged for users to utilize on the everyday media that they consume and tools have been shared with news outlets to better inform their editorial decisions.

Problem Statement

*Can we predict **media bias** in the news articles that we read every day using machine learning?*

We know that news portals play a central role in our society in many ways, including how they keep citizens informed and how they bring important topics into public discussion. With that, media organizations bear great responsibility because of their considerable influence on communities in how they shape our values, norms, attitudes, and behaviours. Any form of news media can contain some level of biased content. We are particularly concerned with "media bias" as defined as a bias towards a Left, Right, or Centre political viewpoint.

By exposing media bias and imbalance news coverage, we can:

1. Increase our understanding of how political bias can be manifested linguistically. For example, is it the choice of words? Is it the context that is created by a sequential combination of words?
2. Learn to better identify political bias in the media we consume every day and build applications and tools to detect misleading information and build critical thinking.

Groups who may be of concern are news consumers who want to be more aware, media publishing companies who want to assess their own biases and provide a more balanced news coverage, and policymakers with interest in media regulation to build more equitable, democratic communities.

Dataset

The dataset is a corpus containing 7,775 news articles from 113 unique U.S. news sources published between June 2012 and March 2019. 6,447 articles were first scraped by Chen et al. (2018) from a website called ***Allsides.com***, a news aggregator that collects news articles on American politics. The data was scraped to fill a gap in analysis on what makes up bias itself on different granularity levels of the English language. Later in 2020 the corpus was extended to a total of 7,775 articles by Chen et al.

(2020) as they scraped more recent articles from the same source. A fairness label was added which measured how fair (neutral) or unfair (biased) an article was.

The fairness label was extracted from ***Adfontesmedia.com***, a website that maintains a "bias scale" through assessments from media experts who annotate each portal with fairness labels.

The political bias label was extracted from ***Allsides.com*** which has its own Media Bias Chart that positions online news portals on a Left-Centre-Right spectrum. The chart was developed and updated regularly through various methodologies and scientific analyses from a multi-partisan lens.

The file came as a JSON file, located on GitHub, so it required loading it into a workable Pandas dataframe. Schema:

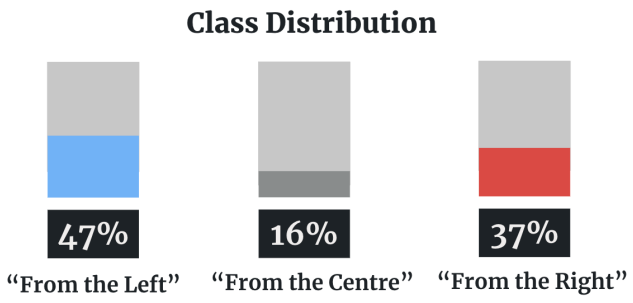
- **title**: the title of the news
- **content**: the content of the news
- **source**: the news portal
- **allsides_bias**: the bias indicated in the allsides.com (Left, Center, or Right)
- **misc**: other information, such as author, date, and topics
- **adfontes_fair**: labels from adfontesmedia.com, whether the article is fair or not (bias, neutral, unknown)
- **adfontes_political**: labels from adfontesmedia.com, whether the article has political bias or not (bias, neutral, or unknown)
- **event_id**: the event id. Articles with the same event have the same id

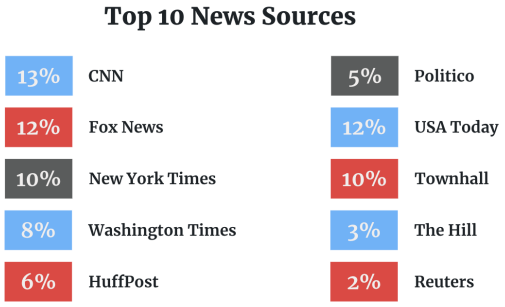
Cleaning and Preprocessing

Python was used to perform exploratory data analysis, data manipulation, text preprocessing, and modeling and evaluation. The specific columns of interest were the title and content columns as our dependent variables, as well as the allsides_bias column as our independent variable containing three biases of "From the Left", "From the Center", and "From the Right". These biases are portal-specific, not article-specific. I decided to use this label to classify the bias that the article contains to train my models with. I use label encoding to encode the classes into 0, 1, and 2 respectively.

The data had an imbalanced class distribution as seen below, therefore I was concerned with recall as the main metric for my models' performances as recall measures the proportion of correctly predicted instances among the instances of a specific class. The top 10 news sources captured some of the most well-known news portals from each side, including CNN, Fox News, and New York Times.

The gensim package was used to perform basic preprocessing of the text data, including converting the text to lowercase to ensure consistency and removal of English stop words. The text was then vectorized into averaged word embeddings of 300 dimensions through loading and passing a pre-trained word embedding model by **LexVec and Wikipedia** which contains a vocabulary of approximately 1 million words in the English language. The embeddings were then converted into arrays and stacked so that each news article would be represented by a long array of 600 dimensions.



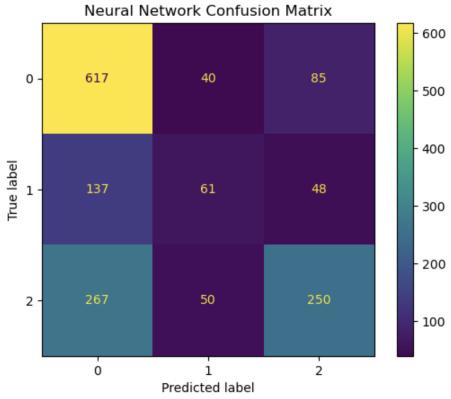


Insights, Modeling, and Results

A logistic regression model performed at a weighted average recall score of 59% and was used as the baseline as our basic classifier. Other models were fitted to further establish our baseline and we performed hyperparameter tuning throughout to optimize our performances. Below shows the results of all our models where we discover that **a neural network performed the best with a weighted recall score of 60%**, specifically able to recall 83% of our majority class ("From the Left"), 25% of our minority class ("From the Centre"), and 44% of our third class ("From the Right"). The recurrent neural networks were only able to make predictions for our majority class at a suspicious 100% recall, failing to predict any instances at all for the other two classes. Further analysis and tuning is needed with the RNNs.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	59%	58%	59%	58%
KNN	47%	46%	47%	44%
Decision Tree	49%	41%	49%	45%
Random Forest	55%	62%	55%	48%
XGBoost	57%	55%	57%	54%
Neural Network	59%	60%	60%	57%
Recurrent Neural Network (GRU)	44%	48%	100%	65%
Recurrent Neural Network (LSTM)	44%	48%	100%	65%

**all scores are weighted averages across all three classes



Findings

Neural networks and deep recurrent neural networks are known to perform very well in various NLP tasks as they are able to capture complex patterns and contextual meaning in text. It is unsurprising that the other models performed on average below 60% recall as they have more difficulty doing this. An 83% recall by our neural network on the majority class is impressive. Returning to our problem statement, we can infer that there may be strong patterns within the text in articles from Left-labeled news sources that construct and indicate a Left-leaning bias. Our ability to predict Right-labeled articles is below 50% recall which is worse than randomly guessing. Despite Centre being the minority class, they may have been harder to predict because of their neutral perspective. Further analysis should look at upsampling the other classes to see if our models can perform better on them.

Recommendations

1. **Develop a Bias Checker:** build an app to measure how Left, Centre, or Right leaning an article is.
2. **Collaborate with Fact-Checking Organizations:** deploy bias checkers to help users identify misleading content and promote media literacy.
3. **Provide Data Insights to News Outlets:** share anonymized and aggregated data insights on media bias to promote internal learning of own bias tendencies, to make informed editorial decisions, and to overall promote a more balanced news coverage for their readers.