**Predictability in Journalism**

The intention of this project is to quantify predictability in journalism. We will accomplish this by analyzing the relationships between the headlines of articles, the bodies of articles, and their respective editorial sources.

The first part centers around the relationships between the articles (both headline and body) to their respective sources. This exploration is straightforward, given a headline/body, which newspaper source is likely to have produced it? This will provide concrete objective measures to quantify style predictability from a broad editorial perspective.

Secondly, we will explore the remaining relationship between the headlines and the bodies of the articles themselves. This is intended to quantify style predictability - how well does the article match the headline? This has significant potential applications for detecting ‘clickbait’ or inaccurate headlines.

This project is important because it can provide insight that is potentially of great interest to newspapers and editors. It will give valuable insight into how effective headlines are for capturing an audience’s attention.

This exercise is challenging because of its uniqueness. Existing literature on text analysis methods centers on more common practices such as sentiment analysis or author-detection for fraud. Our project borrows elements from these practices but with the unique goal of measuring creative predictability, which can be broadly thought of as ‘style analysis’.

Our main data set will be the ‘All the news’ data set from Kaggle. It contains 140,000+ articles from 15 American publications such as the New York Times, CNN, etc. from 2016-2017. A secondary data set has 4.5 million headlines from 10 publications ranging from 2007-2022. We might use this for the first part of the first portion of the project.

The first part of the project can be structured as a multi-class classification problem. The second part of the project can be structured as a summarization problem, where we will use summarization procedures and metrics such as BLEU scores for evaluation.

References

Machine Learning Media Bias MIT study: <https://arxiv.org/pdf/2109.00024.pdf>

Abstract evaluation: <https://www.frontiersin.org/articles/10.3389/frma.2018.00016/full>

Author detection: <https://www.nature.com/articles/s41598-022-13690-4>

Another authorship detection study from Stanford: <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760185.pdf>

Fake News Detection with LSTM + Attention: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8075658/>

Sentiment Analysis on headlines: <https://towardsdatascience.com/sentiment-analysis-on-news-headlines-classic-supervised-learning-vs-deep-learning-approach-831ac698e276#:~:text=We%20created%20sentiment%20predictions%20using,and%20deep%20neural%20network%20respectively>.

[Submit Here](https://docs.google.com/forms/d/e/1FAIpQLScTX8AaUW0mMYywtB6ygIBgKYItHpxXZPKbMlA1aFSnpOfodQ/viewform)

**To do**

11/15/23

* For each article-outlet group:
  + Summarize each article in an article group using Pegasus to generate a “Short Reference”
  + Compare this Short Reference with that article’s headline and check the BLEU score between them
  + Calculate the average of each news outlet’s BLEU scores
  + Compare the average BLEU scores of each one against each other to see how well their headlines summarize the articles
* For each article:
  + Classify each article according to HW 3’s code
    - Try headlines as input first, check for performance issues
      * If performing well, do body as well. If not, proceed
    - Replace topic with news outlet source
  + Generate confusion matrix for all outlets and see if you need to cut down
  + Check how often some news outlets were confused as other ones to see if any of them have a distinct style