Design Document

Team Name: Midnight Sons
Vinnie Khanna, Maanas Purushothapu, Raj Srivastava, and
Sahil Sudhir
Georgia Institute of Technology

October 2020

1 Overview

News is the accurate reporting of current events that are important and relevant to the general public.

The influence of news articles on a population is deeper than the information given in an article. How the articles are presented to users is equally if not more significant. Many hidden factors, including the prioritized listing structures developed by media outlets, sway viewers intentionally and provide confirmation bias towards ill-informed assumptions. In an effort to define a fair analysis and improve user interaction, our group has developed an algorithm using Natural Language Processing (NLP) and web scraping techniques to restructure a sample news feed. The incorporation of our machine learning model into the algorithm provides a unique implementation that incorporates many factors of the problem at hand.

2 Scope and Informed Factors

To aid in making a model within the time constraints, we wanted to target an English-speaking country. As our dataset was in English and as English speakers ourselves, we could do qualitative analysis of the articles more easily. We chose Australia due to its common language and similar issues to the US, with our chosen topics being General News & Health. Our dataset would likely be categorized under the topic of General News.

Our solution for news ranking consists of using the News API[7] to retrieve articles by specifying country and category and an algorithm to score each article, which is what articles are ranked on. Higher scores put articles closer to the top. The composition of the ranking algorithm consists of three measurable factors: credibility, readability, and time since publishing.

Credibility, the most important factor, is determined by evaluating the article with an NLP ML model trained by the "Fake News Inference Dataset," which classifies collections of statements as "real" or "fake" and has over 30000 training samples.[3] The model returns a decimal between 0 and 1, which is the probability that the article is credible. This factor is the most important in making our news ranking system democratic, as its sole purpose is to filter out articles that are probably untrustworthy.

The second factor, readability, determines how difficult the article is to read, based on a ratio of words to sentences and syllables to words. This factor within the news source is calculated using the Flesch Reading Ease Equation.[5] After this function returns a value located between 0 and 100, we further constrain it with ideal and absolute ranges that determine the overall readability score used in the final function. The readability within the range of a 10th- to 12th-grade level allows for inclusivity towards individuals of various reading levels that could have been discriminated upon by unintentionally ranking more difficult articles higher on the list. Readability will also exclude or discourage articles with abnormally basic language, as they can be a sign of low-quality information.

The third factor is time since publishing, to keep top ranked articles relatively up-to-date. A three-day soft threshold was established in creating the algorithm, so articles less than 2 days old would have a slight to moderate preference compared to articles between 2 to 3 days old, and articles 3 days old would have a very significant preference compared to articles older than 3 days.

3 Algorithm

After news articles are fetched from Australian news sources using the News API, they are fed to an algorithm to determine their rankings. This algorithm is based off of three unique factors: credibility, readability, and time since publishing. These three factors combine into the equation:

$$Score = Credibility^2 * Readability * Time$$

where *Credibility* is the probability returned by the credibility model, *Readability* is the result of the Flesch readability model, *Time* is a custom time score, described below, and *Score* is the final score of the article, used to determine ranking. Articles with higher scores are ranked higher (closer to 1) than articles with lower scores.

Real is the probability that the article is real or fake news. The credibility model is a text categorizer model from the SpaCy library[6] that uses a bag-of-words architecture. It was trained using the "Fake News Inference Dataset."[1] Ultimately, the model is a binary classifier that was able to predict if news was real or fake with around 85% accuracy. This could likely be improved using neural networks or an ensemble of neural networks and bag-of-words, but unfortunately the runtime of training those models was too high. *Real* is squared

in the equation to give it that much more importance over the other factors in determining a ranking score.

Readability is determined using the Flesch readability ease test when articles are first fetched, which is defined as:

$$206.835 - 1.015(\frac{words}{sentences}) - 84.6(\frac{syllables}{words}) = Flesch$$

This is then used to define the *Readability* Score:

$$Readability = \begin{cases} L - (55 - Flesch) & Flesch < 55 \\ L & 55 \leq Flesch \leq 75 \\ L - (Flesch - 75) & 75 < Flesch \end{cases}$$

where L is defined as difference between the ideal and absolute ranges, or distance from the preferred readability range to the absolute minimum or maximum. A value of 15 was used for our algorithm.

The ideal Flesch score was determined to be between 55 and 75, which is of a 10th- to 12th-grade level. This range excludes both low-level articles, which are mainly either opinion articles or advertisements, and higher-level articles which tend to be more science-based and specific certain fields. As such, the *readability* score determines how for from this ideal range the article is in. The absolute range is between 40 and 90.

The score for *Time* was based on the *arctangent* function, and was used to prefer newer articles, specifically articles around 3 days, or 72 hours, old. The equation was:

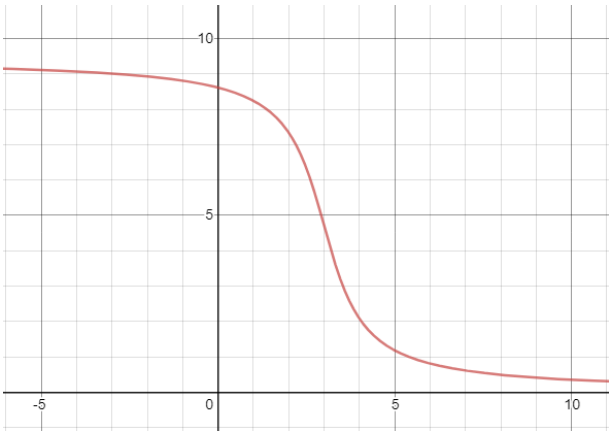
$$Time = S * arctan(-g(\frac{time}{24} - d)) + \frac{S\pi}{2}$$

where *time* is how long the article has been published in hours, S is the scaling factor, d is the number of hours since the article was published, and g is gravity, which is used to determine how fast the *Time* score falls as an article gets older. Testing showed that constant values of $S = 3$ and $g = 1.2$ were sufficient with a limit of $d = 3$ days, or 72 hours, meaning that articles 3 days or newer would be ranked, all other factors held constant, approximately the same.

4 Additional Factors of Interest

Although our algorithm touches upon and improves key problems to many news feeds, we have also considered additional variables of interest.

One such factor is the concept of *time relativity*. For a given news event n, there exists a certain amount of time t that passes after that event occurs and a news article gets published for event n. Although this time t has an influence towards the quality of the news article, its magnitude of importance is relatively small compared to other factors. Instead, we chose to emphasize



A graph of the *Time* equation for the specified constant values.
The x-axis is in days for this particular graph.

the depreciation of an article's value due to time since older articles tend to be less reliable than newer ones on the same event n. With this in mind, incorporating the concept of time relativity into our equation would "double-dip" and exaggerate on the importance of time. Time is significant towards our algorithm, but the extent of its influence should not overpower the algorithm.

In addition, our group took into account the bias associated with certain news outlets. However, instead of emphasizing this notion within our equation, we chose to ignore it in our back-end implementation. At first we thought our algorithm should emphasize removing bias by favoring news articles with a center bias, as we originally believed center bias means no bias. Although the center of a spectrum seems like the "best of both worlds", it does not imply it is neutral, unbiased, or perfectly reasonable.[1] Since the center may miss important perspectives and neglect valid arguments that right- and left-leaning media outlets present, we decided to not prioritize any part of the spectrum. In turn, this allows us to simplify the equation while taking into consideration the bias associated with each news article. However, we still believe bias should be transparent to readers as hidden bias can mislead and divide people in a democracy.[2] Front-end displays should have an indicator that tells the direction and magnitude of bias of the article. This bias rating can be decided by a group of experts based on the media outlet, or it can allow users to easily access information about the ownership of the media outlet as the Carter Center's election standards recommend.[4]

Lastly, we considered incorporating hate speech, incitement to violence, and discrimination in our algorithm. According to the election standards from the Carter Center, controlling these three are vital to promote safety and equality.[5] Articles with hate speech or incitement to violence would be removed from the list of articles. Also, we would have liked to increase the ranking of articles that

promote female and male candidates receiving equal coverage, while reducing the ranking of articles that include discrimination. Unfortunately, we did not have time to find or create a good machine learning model that could predict if an article has hate speech, incitements to violence, and discrimination to incorporate those into our algorithm.

References

1. "A 'Center' Media Bias Rating Doesn't Mean Neutral." AllSides, 13 Aug. 2020, www.allsides.com/blog/center-media-bias-rating-doesnt-mean-what-you-think-it-means.
2. "AllSides Media Bias Chart." AllSides, 23 Sept. 2020, www.allsides.com/media-bias/media-bias-chart.
3. Bidgoly, Amir Jalaly. "FNID: Fake News Inference Dataset." IEEE DataPort, IEEE, 18 Aug. 2020, ieee-dataport.org/open-access/fnid-fake-news-inference-dataset.
4. "Election Obligations & Standards Database." Election Standards — The Carter Center, eos.cartercenter.org/parts/14.
5. Flesch, Rudolph. "How to Write Plain English." Guide to Academic Writing Article - Management - University of Canterbury - New Zealand, 12 July 2016, web.archive.org/web/20160712094308/www.mang.canterbury.ac.nz/writing-guide/writing/flesch.shtml.
6. Matleonard. "Text Classification." Kaggle, Kaggle, 1 Oct. 2020, www.kaggle.com/matleonard/text-classification.
7. "Python Client Library." News API, newsapi.org/docs/client-libraries/python.