News Bias Proposal

Alex Lambert, Casey Mathews, and Shivam Patel

***Abstract*—This document details what our project is about, how it will be built, and what it hopes to accomplish. The goal of the project is to find patterns in the words of both politically biased and scientifically accurate/inaccurate articles.**

I. INTRODUCTION

Our project's primary goal is to find similarities and differences between word usage and patterns in both politically biased and scientifically accurate/inaccurate articles. Due to the current political climate, ease of making internet sites, and the politicization of science, it is becoming increasingly difficult to find unbiased and factual articles. Sites like mediabiasfactcheck.com can greatly help in this process, but they discuss the entire web sites instead of individual articles. If patterns and ways of detecting bias or inaccuracy from then articles content can be discovered, a person's ability to discern unbiased factual articles from the rest could be greatly improved. If the patterns are highly evident, automated checkers could even be made. Regardless of being manual or automatic checking, a guide for determining bias and scientific validity would greatly benefit everyone.

II. METHODS

A. How we will obtain our data

We will obtain our data by using the process of web scraping on predetermined articles. One potential way to accomplish this is to use Python since there are many useful libraries to make this process easy. We could use Selenium as a driver for the web, which is useful for obtaining the html from a webpage and storing it. We could also use the library BeautifulSoup4 to parse the html once it is in a string. Once the data is gathered from the web page, a filter can be applied to get rid of everything except the main text, article headings, and certain HTML tags. We can store the data in a database such as mongodb and categorize it based on each type of article.

B. Data Representation

We can also use Python to create stunning visual representations of our data with libraries like matplotlib to generate graphs and charts to represent the data we collect. Along with the graphical representation, tables will be used to present interesting information in its raw form.

C. What we will be analyzing

- Emotional words: This method involves counting the number of "emotional words". Emotional words are words like "freedom" and "dream" or phrases like "Think of the children."
- Word Length: As we read in the words, we will be grabbing the length of each word and determining the average word length. This may indicate that jargon or fake scientific words are used.
- Headline vs. Content: Using other data gathered and word context, we will try to see how related the article is to the header. This may be used to check for which types of articles are more likely to be clickbait.
- Fact vs. Opinion: We can analyze which types of articles are more likely to use statements such as "I think" or "I believe" over absolute statements that may begin with "I know.".
- Positive vs. Negative Words: We can analyze word connotation to determine if an article is primarily using positive, negative, or neutral language.
- Buzz word count: By counting buzz words we can figure out if certain types of articles use buzz words over other articles..
- Point of view: By looking for and counting words like "I" and "you" we can try to see if the author is presenting information from their own viewpoint. This could indicate a higher level of opinionation over pure fact discussion, or attempts to hide opinion as fact.
- References and citations: By counting references and citations we can see which types of articles are more likely to show where they are getting their information from, if anywhere at all.
- Picture and its caption: By counting the number of pictures we can determine which articles are more frequently accompanied by photographs. When possible, we will examine alt text to determine what the picture is about.

III. TEAM ORGANIZATION

A. Responsibilities

- Alex will be responsible for setting up the database which includes adding the word filtering groups.

- Casey will be responsible for web scraping the articles that are chosen, and parsing the data received
- Shivam will be responsible for examining the parsed data and using the proper measurement metrics to determine results. He will also create visual representations of our results.
- All group members will contribute by picking the articles we want to use, programming the algorithms we need for our comparisons, and working together on the final report.

B. Timeline

The order of work we expect to get done is as follows:

- Determine articles to use for content comparisons.
- Web Scraping the chosen articles.
- Set up our word categories and qualifiers in a database.
- Parsing the data into our word categories and qualifiers.
- Examine the parsed data.
- Program the algorithms on comparisons of our identifiers.
- Use measurement metrics determine the level of bias
- Visual Representation.
- Final report on evaluation and analysis.

IV. EXPECTED OUTCOMES

- We expect articles with stronger biases to have higher word frequencies for emotional language.
- We expect more conspiracy based or scientific articles to use longer words and jargon throughout the text.
- We expect unbiased articles to have a lower usage of 1st and 2nd person pronouns since they will most likely be trying to report fact not opinion.
- We expect biased and conspiratorial articles to lean more heavily on either a high usage of positive or negative words.
- We expect biased articles to have a more consistent use of buzzwords.
- We expect a higher number of references and citations from articles that claim to be scientific.
- We expect the headlines of articles to match the main content less with more biased media.
- We expect biased articles to have a fewer number of pictures.