Latent Probabilistic Model of News Sources

Project Alpha Prototype Report

Army Cyber Institute

Machine Learning for Media Bias

William Hiatt, Gabriel Matthew, and Deven Biehler

02/21/2023

TABLE OF CONTENTS

I. Introduction

This document serves as a guide to the Latent Probabilistic Model of News Sources. It will present new and updated details of the current state of the project as well as the direction the project is going. It will include information such as the social science aspects of the project, the open source models we will be using, models we will be creating, research that we will be using and research we will be doing internally, and any requirements that we currently have. Finally this document will quickly go over the bio's of group members on the team that are working on developing the model as well as all the stakeholders that the project will effect.

I.1. Project Introduction

The US Army Cyber Institute needs a model used for evaluating bias within individual news articles. This model will be used to give a bias rating on a specific news article. The model will grade the article on a scale based on the different bias's that the model looks for. With this scale the Army will be able to fight information warfare more efficiently, both on US based news sites and foreign news sites. The end goal of this project is to create a new model for the Army that will rate individual news articles on their bias and misinformation. This rating will be displayed in a way that is easy to read and follow. The model will use a mixture of currently available software as well as software created by the team.

I.2. Background and Related Work

Concerns about media bias and its potential impact on society have grown in recent years. Misinformation, polarization, and even discrimination can result from biased news. Journalists and media companies can become more aware of their own biases and work toward creating more objective and balanced news content by using a machine learning model that detects bias in news reports. Furthermore, news consumers can use such a model to critically evaluate the news they consume and make more informed decisions.

Due to large amounts of misleading information or complete misinformation, the US military has grown increasingly interested in fighting back. With their help, we are hoping to build a model to decrease the amount of misinformation and the speed at which it spreads.

There are some sites that currently provide a similar service such as allsides.com, mediabiasfactcheck.com, and adfontesmedia.com. These sites currently just look at the media outlet instead of breaking down individual articles. These outlets also tend to use humans to help determine this bias and that brings in another potential bias.

I.3. Project Overview

The media has a huge duty in the form of distributing news to the American public. The information plays a huge role in how people vote. The current offerings in finding media bias aren't as detailed as they should be and still contain a bias in themselves as it's often a human who is assigning these bias's. For example, allsides.com has CNN ranked as far left while adsfontesmedia.com has them as slightly left leaning. This is just a single example, but it shows that there is even a bias within deciding bias. In addition one post from a media outlet could contain no bias and be factually correct while another may contain a lot of bias and not be factually correct. The goal of the team is to combat this by creating a model to find the bias, without any hint of human bias, of a single news article.

The team, with the help of mentors at the Army Cyber Institute, will create a new model that will address the issues presented in current media bias tools. Through research, designing, building, and testing the team will deliver a media bias tool to help fight information warfare.

The project is being built from the ground up. There are no limitations on how we create the model. The team has decided to use Python to create the model. Python is extremely popular for machine learning and will be a great platform to use. The code will be housed on GitHub, this allows for easy version control and collaboration.

The team is also in contact with two Social Scientists at Army Cyber Institute to help in finding the most important bias's to use as well as the best way to display the output. The team will be using sentiment analysis and fact selection as the primary drivers in deciding a bias. As the project progresses, we plan to include additional secondary drivers.

I.4. Client and Stakeholder Identification and Preferences

Our client is the US Army Cyber Institute with Senior Research Scientist Iain Cruickshank as our mentor and primary contact for the project. The product will be used and maintained by the Army. There are several stakeholders within the Army Cyber Institute including Iain and his colleagues.

II. Team Members - Bios and Project Roles

William Hiatt is a 4th year software engineering student at Washington State University. His skills include C#, Python, SQL, Java, C++, and GoLang. He has prior experience working as a software engineer intern at Kochava as well as a junior web developer at Washington State University. For this project, William will act as a team lead, main developer and designer of the model, and work with his team to create a viable machine learning model.

Gabriel Sams is a 4th year Computer Science major at Washington State University. His skills include C#, Python, Database development/management, SQL, machine learning, software development, agile process, test-driven development, and data science. His prior experience includes mobile application development, database development and management, and unit, end-to-end, and functional testing. For this project, Gabriel will act as a main developer and designer of the model, and work with his team to create a viable machine learning model.

Deven Biehler is a 4th year Computer Science major at Washington State University. His skills include C/C++, Python, SQL, machine learning, HTML/CSS, agile process, Haskell, and data science. For this project, Deven will act as a main developer and designer of the model, and work with his team to create a viable machine learning model.

III. Project Requirements

III.1. Spike Stories

III.1.1

Determine the tools to be used for sentiment analysis/textual analysis. Users will be able to view per-article data gathered from NLP and sentiment analysis. The data will be clearly displayed and comprehensible so the user can interpret it in a professional setting. Our team will need to determine open-source tools that will provide the highest accuracy and reliability to our data.

III.1.2

Develop an algorithm to compare per-article emotionality and bias. Develop an algorithm to compile those per-article results into a per-source/per-topic bias for each source. Develop an algorithm that combines per-source data into an ecosystem of sources. The user will be able to gather useful data from three compounding levels of analysis. The tool itself will focus on displaying the "news ecosystem" for users, so the previous layers need to provide useful data to construct it. We will explore and develop algorithms to compile meaningful results at every level of news bias modeling.

III.1.3

Find the most meaningful ways of measuring bias in text. Users will need the most accurate bias data for each topic. By researching and talking with social science professionals the team will create the most effective way at measuring bias so that the users have the most accurate information available. In the beginning only a couple different forms of bias will be evaluated with plans to expand in the future.

III.1.4

Explore methods of fact-checking as well as opinion recognition. Users will need to see when an article is falsifying data or otherwise lying to promote a bias. Fact-checking is an essential point of data when determining the motives of a particular article or source. We want to research possible ways of assessing the validity of a fact. Fact-checking will allow us to determine if a piece tries to persuade a user into a specific worldview. If we can find the lies, we can detect future lies that could be told in articles yet to be published.

III.2. Functional Requirements

III.2.1. Sentiment Analysis

Per Topic: Using a pre-trained natural language processing model, we receive an emotionality score towards an array of topics. Sentiment analysis is crucial to predict the emotional bias towards a specific topic.
Source: Senior Program Manager with Army Cyber Institute originated this requirement. The requirement is necessary for the bias analysis.
Priority: Priority Level 2: Essential and required functionality

Per Article: We receive an emotionality score towards the entire article using a pre-trained natural language processing model. The goal is to give the model more crucial data about the article's topic.
Source: Senior Program Manager with Army Cyber Institute originated this requirement. The requirement is necessary for the bias analysis.
Priority: Priority Level 2: Essential and required functionality

Per Source: Using a custom-built machine learning model, we can determine where the source will lay in the entire ecosystem of media news sources. The goal is to allow a user to understand the worldview of a source within the media ecosystem.
Source: Senior Program Manager with Army Cyber Institute originated this requirement. The requirement is necessary for the bias analysis.
Priority: Priority Level 2: Essential and required functionality

III.2.2. Topic Modeling

Modeling Relevant Topics: Determining the topics is essential to model a per-topic bias analysis. The model learns bias towards an array of topics to portray the media ecosystem to the end user.
Source: Senior Program Manager with Army Cyber Institute originated this requirement. The requirement is necessary for the bias analysis.
Priority: Priority Level 2: Essential and required functionality

III.2.3. Similarity Analysis

Similar Article Analysis: When learning the bias of specific sources, we can use their similarities to other sources to predict missing values. The model will look for any blatant copying from sources and take it as valuable data to make accurate predictions.
Source: Senior Program Manager with Army Cyber Institute originated this requirement. The requirement is necessary for the bias analysis.
Priority: Priority Level 1: Required for an accurate model

III.3. Non-Functional Requirements

Reliability/Accuracy

This tool should most importantly provide accurate and reliable information based on the input given. The tool should provide the same results for the same input every time, as well as provide data that is usable in a professional news media environment.

Extensibility

This tool will be extensible to other domains within online media, including other languages and other political environments. While the tool itself will only work for English language and American news sources, it will allow for the insertion of new tools and libraries to adapt it to these new domains as its users see fit.

Usability

This tool will aim to be usable by industry professionals and those who have domain knowledge of American news, but not necessarily any knowledge of Computer Science or how the tool functions in terms of programming. This means that the tool will include an instruction manual or other documentation that helps users work with the tool. The tool will be designed in an intuitive way.

Content-Focused

This tool will focus on analyzing the actual content of news articles rather than the context surrounding them. The tool will analyze the text itself as well as relevant data and sources to make its calculations–it will not focus on the author, the timing of the article, or other contexts involving the news piece being analyzed.

Comprehension

This tool will provide highly comprehensible data that can be interpreted by industry professionals to make professional decisions in a business setting. It will provide any relevant data to the user and will document any unreliable data or data that could be up to interpretation. Our goal is to be as transparent as possible about the reliability and usability of the data.