Project Proposal – Revised

LeTicia Cancel

DATA 698

October 2, 2022

Research Question

For this project I chose to analyze news datasets to identify true versus fake information, or as it is sometimes described on social media, "Real" News vs "Fake" News. People spend most of their time on the internet so we are more likely to get our news from online articles instead of television. Information is spread quickly and easily through social media but how can we tell if the information we are reading is accurate? Is there a way to flag an article as misinformation? What are the consequences of an article being misrepresented as true? For this paper I will use the term Fake News in reference to articles that are or are suspected to be misinformation and Real News in reference to articles with factual information.

Newspaper articles hold a high standard for the accuracy of information that is published. Authors are required to find reliable sources, backup claims with quotes from individuals or find documentation that can be used as proof that was they write is correct. The author must research their topic using books, online articles, and interviewing individuals1. The author must include their sources so the article can be checked for accuracy by anyone who reads it. When we read articles from major newspaper companies, there is a trust, an expectation that we can believe what we are reading without doing the research ourselves.

Information online is abundant through personal and professional blogs, local and global news websites, and free video services like YouTube. According to Siteefy.com, there are 197, 046, 670 active websites as of 9/18/2022 and "175 new websites created every minute"2. If the information we are reading is presented as factual, how do we know that the source is reliable? Can we read something and trust that the author did their research prior to writing the article? Are all articles published on the web held the same standard as a company such as the New York Times?

When thinking about misinformation, how can we determine the intention of the author? Was it the writer's intention to be deceptive and publish information? Did they intend on writing an accurate, factual article or blog but not do their due diligence when researching the topic and can we clearly know their intention without actually speaking to the person?

Facebook has been accused of creating an algorithm that prioritizes negative posts to a user's feed since people are more likely to interact with content that sparks a strong emotional reaction3. Using the pandemic as an example, this was such a scary time for all and the spread of misinformation about a new virus was dangerous and potentially deadly. Facebook updated their system to compare information against a fact-checker and flag posts as false4. Twitter has also attempted to stop the spread of misinformation by asking users to flag posts that "seem misleading"5. The existence of fake news is not new and is also not unique to the pandemic information that has been shared on social media in the past two years. What makes this so important today is just how easily information is shared to a large group of people. A system is needed to accurately identify misinformation as quickly as this information is spread and is needed across the web, not just on social media platforms.

I am glad to see these social media companies attempting to identify and stop or slow the spread of misinformation by using fact-checkers and flagging by the community. I would like to learn how the fact-checker and other methods of identifying misinformation work. What is the common thread between these "fake" articles and how accurate is the algorithm used to catch the misinformation? I understand that the most accurate way to determine if an article is fake is to run it through a fact checking system or to have a professional editor check the author's sources for accuracy. Most individuals, including myself, do not have access to a fact checking system and are not professional editors who would check the sources of an article we are reading. So what can we do instead?

I plan to build a model that can categorize the information as Real News or Fake News. The purpose of the model is not to check an article for factual accuracy but instead flag an article as possible misinformation or Fake News. This flag can help the reader to make an informed decision with what they are reading. This model will be used with public article datasets found on Kaggle that are assumed to be "Real News" to determine accuracy.

Research

This academic paper about the Development of Fake News Model Using Machine Learning through Natural Language Processing is another resource for this project6. It breaks down fake news into 6 categories from False Connection which is described as "When headlines, visuals or captions don't support the content", to Fabricated Content which is described as "New content that is 100% false, designed to deceive and do harm." Naïve Bayes, Support Vector Machine, Passive Aggressive, and Logistic Regression models were used in this paper split and test the data in the 6 categories.

Data

My main data source will come from Kaggle's [Fake and Real News Dataset](about:blank)7, a dataset that contains two files already categorized as real and fake. Each file has 4 columns, Title of the article, Text of the article, Subject of the article, and Date is was published. The article text will be mined to create features for our model. Natural Language Processing (NPL) can be used to find common words and phrases used in the Fake News dataset from Kaggle. This will mainly give us the frequency of words so we will also use the Random Forest model. Decision Trees will be used first on the article text to help us group the text. The Decision Trees will then be used in the Random Forest Classifier to predict of the article is real news or fake news. The model will be tested on Fake and Real News Dataset to test for accuracy.

The model will be used against the [All the News](about:blank)8 dataset from Kaggle which contains articles that we expect to be classified as true to test the accuracy of the model. I expect a high percentage of the articles to be labeled as True after it is passed through the model created from the Fake and Real News Dataset.

Data Flair gives an example of a machine learning model built using a Passive Aggressive Classifier9. Data Flair describes this algorithm as remaining "passive for a correct classification outcome, and turns aggressive in the event of a miscalculation, updating and adjusting." I will experiment with a Passive Aggressive Classifier with the Fake News dataset and compare the accuracy against Random Forest Classifier.

I will also use the LIAR Dataset from Activeloop10. It contains 12,800 classified short phrases and will be used to build the model for fake news detection. The Kaggle data only has two discrete categories, True or False. The LIAR dataset uses six categories - pants-fire, false, barelytrue, half-true, mostly-true, and true. It will be interesting to see how the accurate the model will be with 6 categories instead of just two. If we think about the impact of more categories, I think it makes more sense to have more nuanced categories instead of a Boolean category.

Summary/Conclusion

A fact-checker would be most accurate in identifying Real vs Fake news but this is not something we have access to. Instead we will use packages in R to find patterns in the news articles so we can accurately categorize them. We will look for patterns in words used, phrases, and tone within the article text. This will be an experimental and exploratory project so I will use multiple text mining techniques in an attempt to get the most accurate model possible.

Sources

1. Fleming, G. (2019, October 7). Tips for writing an effective news article. ThoughtCo. Retrieved September 18, 2022, from <https://www.thoughtco.com/how-to-write-a-news-article-1857250>
2. Huss, N. (2022, August 22). How many websites are there in the world? (2022). Siteefy. Retrieved September 18, 2022, from <https://siteefy.com/how-many-websites-are-there/>
3. Chakradhar, S. (2021, October 26). More internal documents show how Facebook's algorithm prioritized anger and posts that triggered it. Nieman Lab. Retrieved September 10, 2022, from [https://www.niemanlab.org/2021/10/more-internal-documents-show-how-facebooks-algorithm-prioritized-anger-and-posts-that-triggered-it/](about:blank)
4. Mukul, P. (2021, May 28). Explained: How new facebook feature flags misinformation. The Indian Express. Retrieved September 10, 2022, from [https://indianexpress.com/article/explained/facebook-misinformation-fake-news-tool-7332659/](about:blank)
5. Spangler, T. (2021, August 17). Twitter is asking users to flag misinformation, including about COVID and elections. Variety. Retrieved September 10, 2022, from https://variety.com/2021/digital/news/twitter-users-flag-misinformation-covid-elections-1235043215/
6. Kaggle dataset: [https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset?select=Fake.csv](about:blank)
7. Kaggle dataset: [https://www.kaggle.com/datasets/snapcrack/all-the-news?select=articles1.csv](about:blank)
9. Team, D. F. (2021, March 31). Detecting fake news with Python and Machine Learning. DataFlair. Retrieved September 10, 2022, from [https://data-flair.training/blogs/advanced-python-project-detecting-fake-news/](about:blank)
10. Liar dataset. Activeloop. (n.d.). Retrieved September 10, 2022, from [https://docs.activeloop.ai/datasets/liar-dataset](about:blank)

GENERAL COMMENTS

LeTicia,

You’ve clearly picked a fascinating topic (can one viably automate the differentiation between real and fake news?).

My main concern about your proposed is its (expansive) scope. For years, researchers in this field have stressed manual intervention (researching other sources and exercising “good judgment” – in addition to NLP models) – so the prospects of finding a reliable way to completely automate this process seems overly ambitious.

In other words, while I agree that “A system is needed to accurately identify misinformation as quickly as this information is spread and is needed across the web, not just on social media platforms,” I don’t think that’s viable, based on training models on the Kaggle Fake and Real News and All the News datasets. News or information by definition is highly situational and context-driven (unless you can convince me otherwise).

You’ve clearly picked a fascinating topic; my advice is to do a more in-depth literature review and give more thought as to defining the scope of what specifically you plan to do (within the semester timeframe allowed), with what models and datasets, to achieve what types of results (i.e. how you measure success or accuracy).

Please see my comments on the attached draft, which I’ve also posted in the “Draft Proposals with Track Changes” folder in the “Feedback on Drafts” section on the course site.

And please let me know if you’d like to discuss.

A