The Detection and Mitigation of Mis/Disinformation through the Development of Predictive and Generative AI Machine Learning Methods

Abstract:

The issue of fraudulent and misleading news has been an issue for as long as humans can remember, but today it is more rampant than ever with social media giving everyone the chance to stay up to date with news worldwide. With the development of AI there is the opportunity to extenuate the increasing amount of misinformative news, or to try and mitigate it. Previous approaches have brushed the surface, with many fact checking websites leading the way in debunking non-factual statements made by various sources, but the development of generative AI has opened new doors which have allowed for new techniques to develop. This project I am working on is going to continue to push the needle towards removing misinformation and disinformation from media sources through predictive AI and generative AI machine learning methods. By finding new data that is specific to this task, along with developing models addressing many different factors that go into detecting misinformation, my project will be a culmination of the most important factors contributing to mis/disinformation which is unlike many other projects previous. The result will be a generative AI model that is able to intake some information, and return how false it is by returning text describing the falsities along with a graph showing what areas it is the most inaccurate in.

Introduction:

Throughout the internet there are countless sources of news that people use every day to keep themselves updated on current events. The most prevalent being news websites and social media apps. These platforms have grown consistently as web applications have continued to develop, making them the primary news sources for people worldwide. The problem is that it has become easier than ever for misinformative posts and news to spread rapidly with the increase of news sources and the usage of the web. It is extremely difficult to detect wrong from right through a computer screen when the credibility of a piece of news, among other factors, cannot be easily detected. The goal of this project is to consider as many of the factors present, which include clickbait, source reputation, textual analysis among others and use these to paint a picture of where the article may be truthful and where it may be misinformative. This will allow the user to take this information and make their best judgement on whether it is valid information or not, as making that decision for them may be irresponsible. Thus far, my process has led me to getting much larger amounts of usable data through web scraping that I can train models on, which is an area that is still improving regarding this topic. The models developed so far have focused on using NLP methods that create embedding vectors of the text and title of an article to detect textual trends that may determine the level of misinformation. This misinformation level is scaled using PolitiFact’s Truth-o-meter scale which has a range of classifications which are, True, Mostly-true, Half-true, Barely-true, False, and Pants-on-fire. This gives a much more accurate description of the truth value of an article in comparison to a simple binary classifier stating whether something is true or false. Thus far the performance of these predictive models isn’t strong enough to validify deployment, but the work of others and the combination of models will lead us to a much more accurate final model that will help people determine the validity of a piece of news.

Previously there has been a large amount of research and work put into predictive tasks using data and well-known machine learning algorithms to try and detect misinformation. A novel study was conducted by William Yang Wang in which he created a new benchmark dataset to be used for fake news detection. Wang drew information from PolitiFact.com to create a dataset with over 12,000 data entries of the statement, speaker, context, truth-label, and justification for some factual/non-factual statement (Wang, 2017). The development of large datasets like this emphasizes how early on we still are in this process of quality fake news detection methods. This study also goes on and uses this dataset to try and classify similar statements into their correct truth-value category. Unfortunately, the best results Wang got were from Hybrid CNN models that used all the data present but still only got 27.4% accuracy (Wang, 2017). A separate study by Bourgonje, Schneider, and Rehm looked into the clickbait feature of articles through stance detection. Stance detection is a process of taking in the titles of a news article and the text and seeing if the title agrees, disagrees, discusses, or is unrelated to the text. This is an indicator of whether an article may be clickbait or not as most truthful articles will be more focused on just discussing the topic at hand and reporting facts on it. If the title is unrelated or disagrees with the text this may be a red flag for truth validity. The study focused on using n-gram matching of the CoreNLP Lemmatiser to achieve this labeling (Bourgonje et. al, 2017). This process of clickbait detection is just one of the many factors that go into determining the validity of an article. By referencing and using research like this, we can accumulate best practices for several different factors and in turn combine them to create an ensemble model that acknowledges and addresses all the aspects of fake news detection. This is different than many of the studies and projects done previously, as often researchers will focus on one individual aspect of fake news detection instead of combining many.

The data I am using is a combination of readily available data from previous studies, as well as data I was able to scrape and collect specially for usage on my own project. The data that I have used so far has simply had the title of an article, the text of the article, and an indicator of whether that article is classified as clickbait or fake news. This has allowed my models to learn what themes are present in clickbait titles and false news articles. This data I found from the web is primarily binary though. This is problematic because the truth value of an article is never just black or white. It falls on a scale that has a much wider range. This is why I have attempted to implement into my models a way of turning this binary prediction into a multi-class prediction using the PolitiFact.com six-class classification. This provides a much better idea of the truth-value of an article. This leads me to the description of the data I collected from web scraping. Although I haven’t been able to fully leverage this data I’ve scraped yet, I intend to use it as I develop more models and share it with others trying to achieve the same goal. The data I collected was from PolitiFact.com. Initially I scraped all the truth-o-meter data present on the website, giving me a dataset of 22,065 entries that have the claimer of the statement, the statement itself, the truth-value of the statement, the summary of how PolitiFact came to that truth value, and the full text describing how they came to that truth value. This large dataset will be extremely useful in training predictive and generative AI models as it will provide a strong-base data set for them to learn from. Additionally, I also scraped data from PolitiFact.com by looking through all their recent news articles dating back to mid 2019. In this collection I scraped the statement, summary of the text, and the full text. This data will be extremely useful in the training of my generative AI model by giving it highly accurate information about current and previous events to draw from. These data sets combined should allow my predictive and generative AI models to have an excellent performance once the models have been properly selected and tuned.

References

Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. 2017. [From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles](https://aclanthology.org/W17-4215). In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 84–89, Copenhagen, Denmark. Association for Computational Linguistics.

William Yang Wang. 2017. [“Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection](https://aclanthology.org/P17-2067). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.