Cecilia Bell
5.19.2021

Business Project Proposal

Facebook Non-Factual News Detection

The Question

In recent years, tech companies have come under increasing scrutiny for allowing misinformation to spread online. At Facebook, a fake news detector was developed to fight this misinformation, which flags posts with fake news articles. However, a problem has arisen: posts with non-factual articles, including satirical, comic, and opinion articles, are also being flagged as fake news.

The question I want to address, then, is: how can non-factual articles be identified and distinguished from fake news articles (an identification problem)? My first proposed solution path is to build a natural language processing classification model, which classifies articles as mostly true, mostly false, or non-factual (these categories are subject to change).

The desired impact of this project is to improve the accuracy of Facebook's current fake news detection system. Further desired impacts include increasing Facebook's credibility and increasing Facebook user satisfaction, in particular the satisfaction of media companies that are currently seeing some of their non-factual articles flagged as fake news. The impact hypothesis, then, is that classifying posts with news articles into more granular truth categories will improve the accuracy of Facebook's fake news detection system. The hypothesis assumes that such classification is possible; it is important to note that the distinction between true, false, satirical, comic and opinion statements can be difficult to distinguish, even to a human being.

To validate the project, the model will be tested on a test set of articles. If the model achieves a specific R2 score (to be determined), the project will be considered successful.

The Data

For the preliminary exploratory data analysis, I plan to use the dataset compiled for the *BuzzFeed News* article, "Hyperpartisan Facebook Pages Are Publishing False And Misleading Information At An Alarming Rate," published October 20, 2016. It is open source and can be found here: https://www.kaggle.com/mrisdal/fact-checking-facebook-politics-pages. The raw data consists of over 1000 rows and 12 columns. Each row represents an individual post and the columns include Facebook account ID, Facebook post ID, political category, Facebook page name, post URL, date published, post type, truth rating, debate, share count, reaction count, comment count.

The Tools

Cecilia Bell
5.19.2021

I plan to use Excel to perform a preliminary exploratory analysis of my data and Tableau to visualize my preliminary findings.

The MVP Goal

My MVP goal is to create an early iteration of a final draft of the project proposal. This will include a preliminary impact hypothesis, one or more solution paths, and a preliminary supporting visualization. This will require project scoping, data acquisition and cleaning, and some preliminary exploratory data analysis.