Cecilia Bell 5.28.2021

Business Project Write-Up

Fake News Detection

Abstract

The goal of this project was to pitch Facebook a data science solution to a problem facing their fake news detector: posts with no factual content were being flagged as fake news. The data science solution proposed was to perform natural language processing on a large dataset of facebook posts and feed the results into a classification model, which could classify posts into more granular truth-rating categories. In order to demonstrate the importance and viability of the project, I used data on a set of Facebook posts from 2016 to explore the differences between posts with different truth ratings. Notably, posts with no factual content see the highest engagements. After exploring and analyzing the data in Google Sheets, I visualized some preliminary insights with Tableau.

Design

Opportunity

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 misinformation, which flags posts with fake news. However, posts with no factual content, including satirical, comic, and opinion content, are also being flagged as fake news. (See here for an example case.) The problem, then, is: how can posts with no factual content be identified and distinguished from posts with fake news (an identification problem)?

Impact

The desired impact of this project is to offer Facebook users more accurate truth-rating content warnings. 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 posts flagged as fake news. (When this happens, the media company's page administrators are notified that their reach and monetization could be reduced.)

Data science solution path

My proposed solution is to perform natural language processing on facebook post captions and then feed the results into a classification model, which classifies articles as mostly true, mostly false, or non-factual (these categories are subject to change). The solution requires a large dataset of Facebook posts, including their captions and truth ratings. The truth ratings will be determined and cross-checked by a group of fact-checkers given a specific set of guidelines. Other potential solutions include:

- Using third-party (human) fact-checkers
- Allowing users to flag misinformation
- Combining my proposed solution with one of the two above

Impact hypothesis

Using a model to classify posts into more granular truth categories will give Facebook users more accurate truth-rating content warnings.

Measure of success

To validate the project, the model will be tested on a test dataset of articles. If the model correctly classifies a certain percentage of posts, the project will be considered successful. The specific percentage will be determined later, but the new model must outperform the old fake-news detector.

Assumptions

The proposed solution path assumes that the classification of posts into different truth-rating categories is possible. It is important to note that it is difficult to distinguish between true, false, satirical, comic and opinion statements, even for a human being. Certain circumstances make this especially difficult. For example, some fake news pages claim that their content is satirical to evade social media misinformation guidelines.

Data

For the preliminary exploratory data analysis, I used 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 2283 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. I subsequently found a secondary dataset which also included the types of reactions that each post received (e.g. angry reaction, sad reaction, etc.)

Algorithms

In Google Sheets, I cleaned the data, imputing missing values, and aggregated certain data. I then performed some preliminary exploratory analysis and visualized my preliminary insights in Tableau.

Tools

- Google Sheets to clean, aggregate, explore and analyze the data
- Tableau to visualize preliminary insights

Communication

My pitch is presented in a PowerPoint presentation, which includes the visualizations created in Tableau. All colors used are color-blind friendly.