Fake News Risk Identifier

Members: *Shivansh Bansal, Eesha Sharma, Jake Frischman*

Date: *7/18/22*

Introduction

Fake news, misinformation, and disinformation plague our online websites and mislead millions of people who get their news online. Using Data Science, we aim to evaluate the level of risk to which a web page may contain false information, based on text variables such as the number of grammatical errors. Then we will assign the web page a numerical "risk" factor from 0, where 0 is almost certainly false and 1 is almost certainly true. This will help the reader judge what they are reading and determine whether or not the page contains misleading information.

End Goal:	Train a model to scan a website's text and predict (0-1) if the website contains fake news, misinformation, or disinformation.
Use:	A tool to help judge how much a reader should trust a website.
Problem solved:	People blindly trusting information found on websites that may be misleading or give false claims.

Data Source

Dataset Description

a. https://drive.google.com/file/d/1IoTRrJNDJqvaG3hnUpnHQyGvPAJbO8y3/view

- This folder has two different datasets, “true” and “false.” They each contain the bodies of articles that have been classified as trustworthy or untrustworthy. Both datasets have a large amount of content related to American politics.

b. https://www.kaggle.com/datasets/jruvika/fake-news-detection

- This dataset contains around 3,745 different websites that are classified as true or false. This is a raw data set that includes the URL, headline, body, and label (classification) of each dataset.
- Each data set is classified as a “1” or a “0.” A “1” website is considered true and a “0” website is considered false.
- 100% of the URLs and Headlines are valid; however, 1% (21) of the articles have missing body sections. These articles must be cleaned from the dataset.

c. https://www.hindawi.com/journals/complexity/2020/8885861/#materials-and-methodshttps://www.hindawi.com/journals/complexity/2020/8885861/#materials-and-methods

- This research article breaks down article classification and the linguistic variables behind it. It also provides the raw data used in the study, one of which was listed above (letter “b”) and another Kaggle dataset.
    1. This dataset contains further datasets that have an id (unique id for each website), title, author, text within the article, and a label of either “0” or “1.”
    2. In this case, a label of “0” means the website is true and a label of “1” is considered false.

d. https://www.kaggle.com/datasets/hassanamin/textdb3

- This dataset contains around 6,256 different websites that are classified as real or fake. This is a raw data set that includes the headline, body, and label (classification) of each dataset.

- Each data set is classified as a “1” or a “0.” A “1” website is considered true and a “0” website is considered false.
- __ of the URLs and Headlines are valid; however, __ (21) of the articles have missing body sections. These articles must be cleaned from the dataset.

Questions

1. What linguistic variables matter most in a website's level of risk to which a web page may contain false information?
2. What variables does a web page of high risk contain in large degree compared to a web page of low risk?
    a. E.g Does a web page of high risk contain more grammatical errors compared to a web page of low risk?
3. What variables does a web page of low risk contain in large degree compared to a web page of high risk?
    a. E.g Does a web page of low risk contain more verbs compared to a web page of high risk?
4. How accurate is the machine learning model's risk factor for containing false information?
5. What websites are most likely reliable? What websites are least likely reliable? What websites are somewhat likely reliable?
6. What is the True Mean Risk Factor of All Websites on the Internet?
7. What websites and publishers are reliable? What publishers are unreliable?

Analysis

Our plan is to build a supervised ML ensemble XGBoost classifier to produce a risk number for false information in the webpage being read. This is a natural language processing problem so the documents will be cleaned up into one solid page then data points (#of verbs, punctuation, author's other written pages etc.) will be extracted and tokenized. The data will be split 70/30 and the trained model accuracy rate (takes both the false positive and the false negative observations into account) will be aimed for 85% accuracy or greater.

Our first step to answering these questions will be to clean the dataset by, for example, filtering out entries with missing data or with links instead of text.

Secondly we will process the text to extract many variables (0-1) on which we will train our model.

Finally we will test our model on the datasets until we achieve sufficient accuracy. The model will then be tested on websites of our choice published recently and see how it performs.

If we have time we will write a website scraper to collect the current web page for a user and have the model assess the risk of false information then display it to the user. Conditionally we can create a data scraper to collect a couple hundred different web pages and create our own dataset to apply our model to.

Our analysis will consist of creating a ML model to classify text and make a visual representation of the model's output.

1. Data collection
2. Data Cleaning and formatting
    a. Reduce all pages to one page
    b. Bring values like title, author, etc. out
    c. Remove formatting
    d. Remove pages with <20 words
3. Extract Linguistic variables
    a. Research important variables

    b. Use NLTK

4. Choose features/variables to train on

5. Split training and testing data

Train model using training data

Evaluate model

Hyperparameters and model tuning

User website >> trained model >> Output (Classification * certainty)

Visualizations

1. Question 1 can be visualized using a pie chart with the color key corresponding to the different linguistic variables. The size of each piece signifying the degree to which each variable impacts the website's risk factor.

2. Question 2 can be visualized using a pie chart with the color key corresponding to the different linguistic variables for high risk websites. The size of each piece signifying the degree to which each variable impacts the website's risk factor.

3. Question 3 can be visualized using a pie chart with the color key corresponding to the different linguistic variables contained within a low risk website. The size of each piece in proportion to the entire chart can signify the amount that each variable impacts a website's risk factor.

4. Question 4 can be visualized using a single percentage value, a confidence interval of the error in the risk factor (Error = ML Risk Factor - Actual Risk Factor), and the Accuracy of ML * Prediction of ML.

5. Question 5 can be visualized using a table or spreadsheet of common websites that have been classified as either reliable or unreliable (based on the 0-1 risk assessment). Additionally, a scanner could be possibly developed in order for users to enter their own websites to be scanned and determined reliable or not.

    - What websites are most likely reliable? What websites are least likely reliable? What websites are somewhat likely reliable?

6. Question 6 can be visualized using a 95% Confidence Interval of the Mean Risk Factor of a Sample of 1000 websites on the Internet.

7. Question 7 can be visualized using a single percentage of the number of reliable web pages versus the total number of web pages published, for a specific website publisher.

Marketing Paragraph

We plan to add a screenshot of your site on our website and we need you to write the paragraph that will go with it. We only need one submission from each team. Please collaborate on the paragraph and do not use personally identifiable information (like names). Feel free to use the royal we.

Please start with the following information: Team name (optional), topic of site, how/why topic was chosen, who the site is for (audience). After that feel free to write about how you did it, what you learned, how you learned, etc.

Fake news, misinformation, and disinformation plague our online websites and mislead millions of people who get their news online. Using Data Science, our team aims to evaluate the level of risk to which a web page may contain false information, based on text variables such as the number of grammatical errors. Then we will assign the web page a numerical “risk” factor from 0, where 0 is almost certainly false and 1 is almost certainly true. This will help the news readers judge what they are reading and determine whether or not the page contains misleading information.