Using NLP to Understand Hyperpartisanship in Media

Group 18: Mark Glinberg, Stephan Tserovski , Justin Huang
CS 4650 - Spring 2024

Proposal Challenges Overview

	Free	Basic	Pro
Getting access	Get Started	Get Started	Get Started
Price	Free	$100/month	$5000/month
Access to X API v2	✓ (Only Post creation)	✓	✓
Access to standard v1.1	✓ (Only Media Upload, Help, Rate Limit, and Login with X)	✓ (Only Media Upload, Help, Rate Limit, and Login with X)	✓ (Only Media Upload, Help, Rate Limit, and Login with X)
Project limits	1 Project	1 Project	1 Project
App limits	1 App per Project	2 Apps per Project	3 Apps per Project
Post caps - Post	1,500	3,000	300,000
Post caps - Pull	✗	10,000	1,000,000
Filteres stream API	✗	✗	✓
Access to full-archive search	✗	✗	✓
Access to Ads API	✓	✓	✓

Initial problem statement:

- Generate a fully labelled and robust dataset of Tweets pertaining to the upcoming presidential election
- Use an NLP model to predict political leaning of tweets and extrapolate the public's favored candidate

Issue:

- Using the X API costs $100/month, and offers only a minor subset of data (not enough to train our model)
- Feasibility of defining proper annotation guidelines and measuring labeling consensus proved too time-consuming for this project's scope

How We Pivoted

Problem:

- Election season, many polarizing political articles on the internet
- Can be difficult for voters to inform themselves without knowing if the information they are reading is presented neutrally

Solution:

- We found a dataset that includes 750,000 articles, their titles, and labels about whether they are hyperpartisan
- There is also a model which uses DistilBert with a 99% accuracy (theoretically) on that training dataset

Objective

Understanding the Dataset

- What are the general trends we can observe?
- Can we use the provided labels for a more complete picture of the predictions?

Understanding the Model

- What motivates the predictions?
- What are the limitations of its predictions?

What can be done to create a more complete solution to this problem?

Dataset

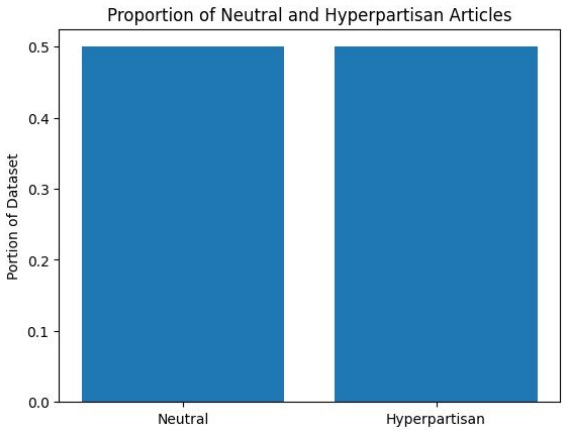
text string · lengths 98 37.8k	title string · lengths 6 198	hyperpartisan bool 2 classes
Money (<a href="https://farm8.static.flickr.com/7020/65515348…	Kucinich: Reclaiming the money power	true
Donald Trump ran on many braggadocios and largely unrealistic campaign promises. One of <a…	Trump Just Woke Up & Viciously Attacked Puerto Ricans On Twitter Like A Cruel Old Man	true
In response to Joyce Newman’s recent letter about a conversation about guns: According to the…	Liberals wailing about gun control, but what about abortion?	true
After Colin Kaepernick rightly chose to kneel during the national anthem before NFL games, many…	Laremy Tunsil joins NFL players in kneeling during national anthem	true
Almost a half-century ago, in 1968, the United States seemed to be falling apart. The…	It's 1968 All Over Again	false
Briefly: In our opinion, full (150% of the regular full position) speculative short positions…	Gold Price in December 2017 - Myriads of Signals and Analogies	true
When the Graham-Cassidy bill failed to reach the Senate floor last week, the media wanted to put a…	Conservatives & Trump Can Still Stop the Worst of Obamacare	true
In recent months, late-night talk show host Jimmy Kimmel has taken to scaremongering his audience wit…	Don't Fall for Jimmy Kimmel's Cheap Zero-Sum Emotionalism	true
Days after Steven Bannon’s blustery, accusatory interview on “60 Minutes,” i…	Tell Us, Mr. Bannon -- Just What Is Trumpism?	true
During the campaign, Donald J. Trump made lots of	THEY DON'T CALL IT 'THE GREAT TWEET OF CHINA'	true

Hugging Face dataset:

[hyperpartisan_news_detection · Datasets at Hugging Face]

Fields:

- article title
- article text
- hyperpartisan (boolean)
- bias (political leaning)
- publishing date
- article ID
- article URL



Prediction Distribution

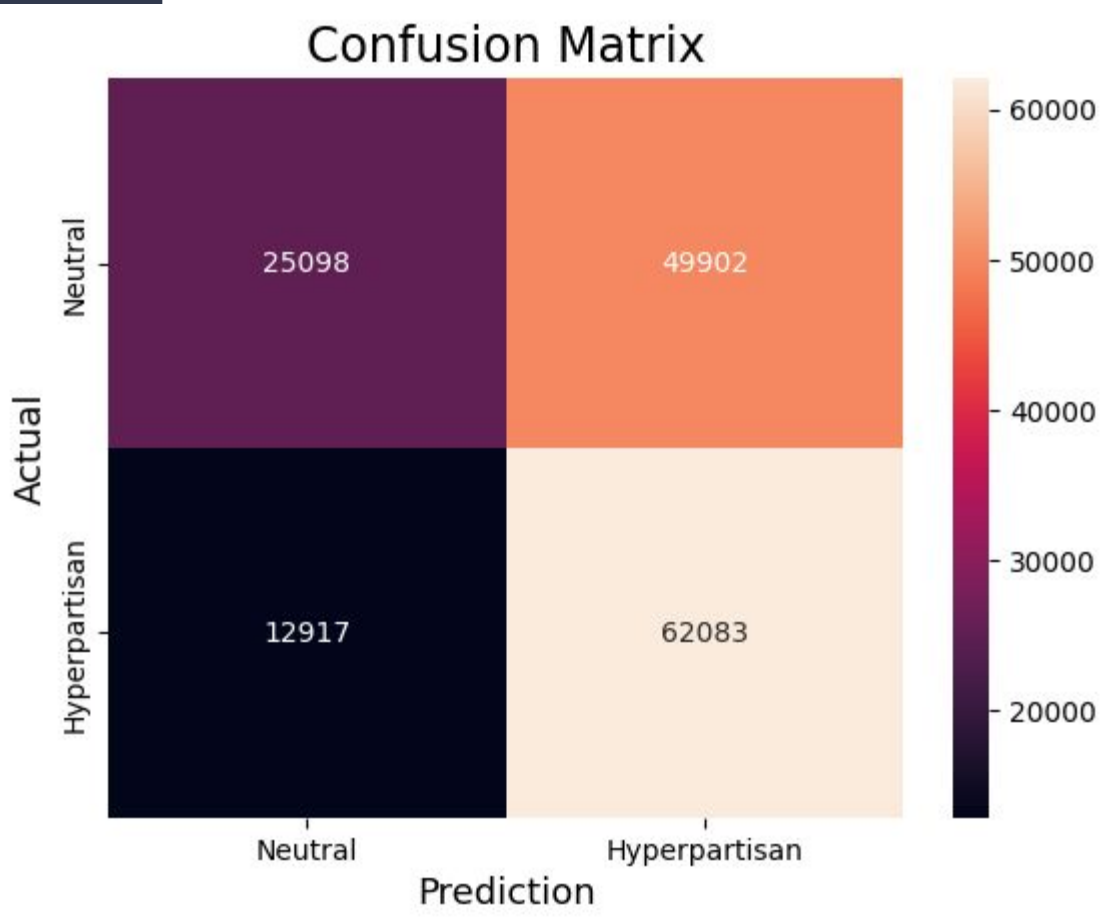
Evaluated on the article title and body of validation entries

Accuracy: 0.58

Precision: 0.55

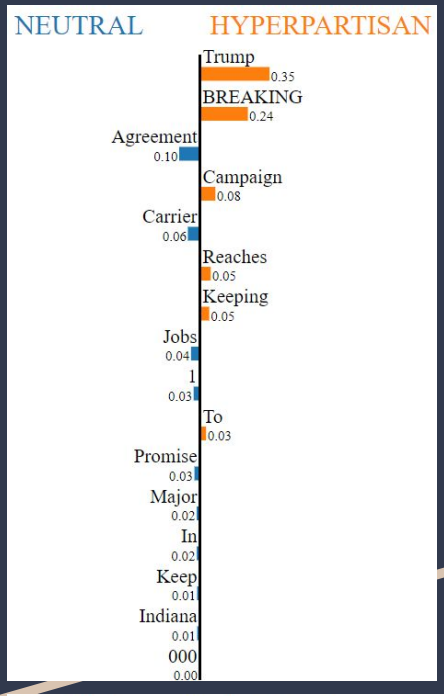
Recall: 0.83

F1 Score: 0.66



Understanding Predictions

Used LIME and transformers-interpret

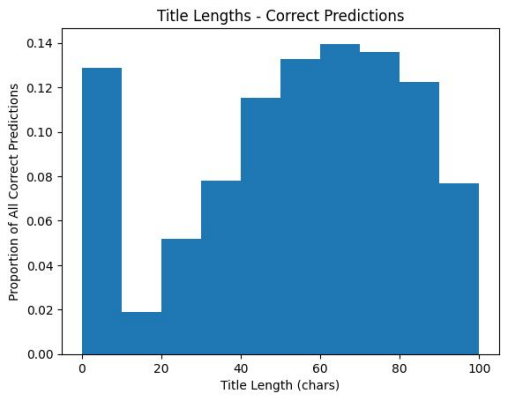


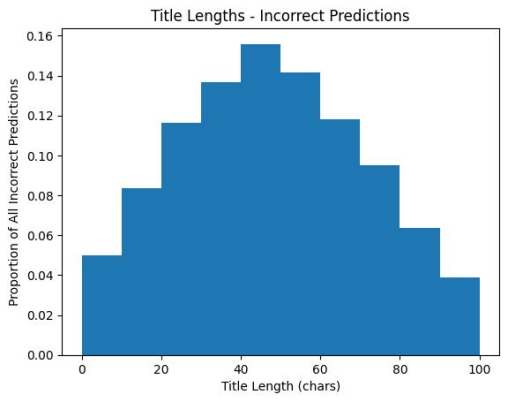
Text with highlighted words

BREAKING: Trump Reaches Agreement To Keep 1,000 Carrier Jobs In Indiana, Keeping Major Campaign Promise

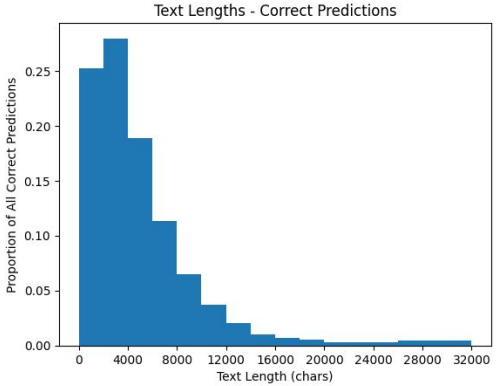
True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	HYPERPARTISAN (1.00)	HYPERPARTISAN	2.58	[CLS] breaking : trump reaches agreement to keep 1 , 000 carrier jobs in indiana , keeping major campaign promise [SEP]

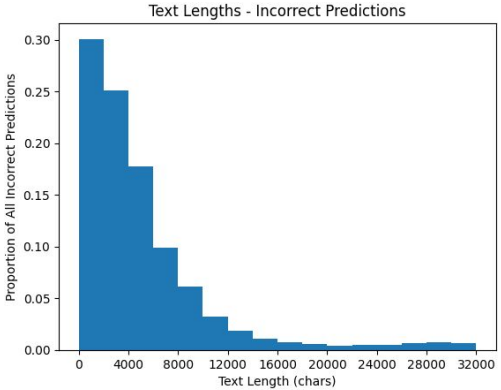
Text-length





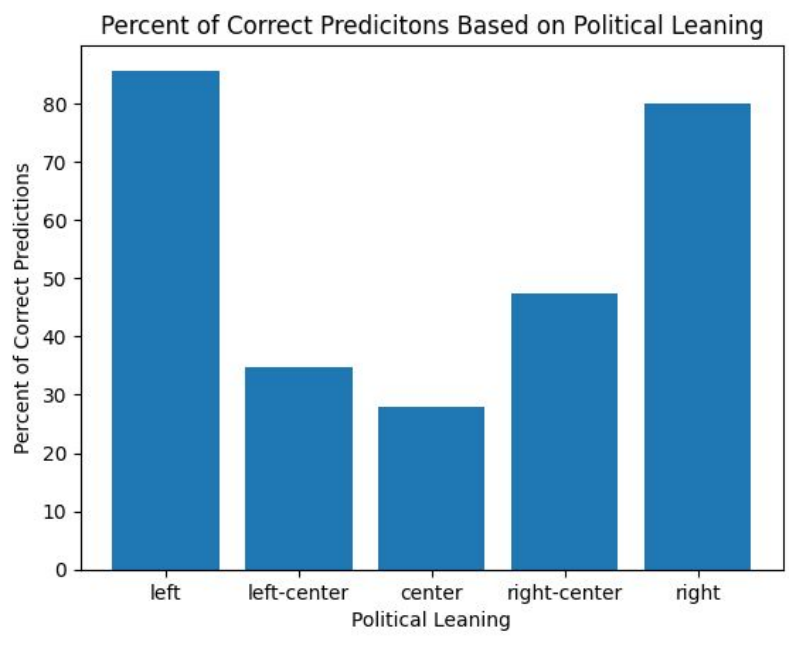
The longer the title or text, the more accurate the predictions are





Bias/political stance

More likely to label left leaning articles as hyperpartisan, followed by right leaning articles



Conclusion & Evaluation

Conclusion:

- The accuracy of the proclaimed 99% model drops significantly when given limited information
- There are multiple instances where certain more sentimental vocabularies are given too much weight

Going Forward:

- Contrastive learning has found success in similar applications
- Increase dropout to reduce overfitting
- Test the versatility of different models