BAIT ‘N’ SWITCH

An AI powered API for optional content moderation.

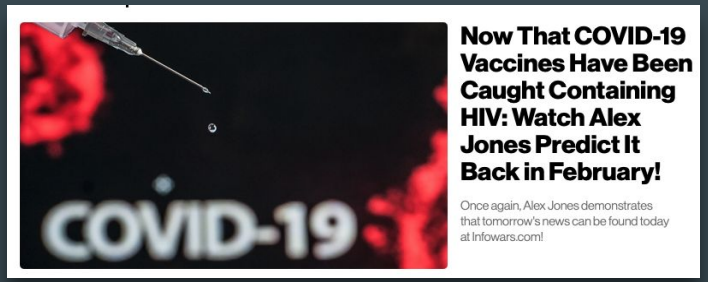
By Tim Hintz
Flatiron Data Science Bootcamp
Capstone Project

BUSINESS CASE

- Online media operates on a pay per click economy
- Harnesses our curiosity for profit
- Create a backend solution for optional in-browser content moderation

WHAT IS CLICKBAIT?

- Exploits our Curiosity Gap*
- Asserts “expert” Knowledge
- Hyperbole/Superlatives



*https://www.theatlantic.com/technology/archive/2014/03/the-curiosity-gap-is-closing-says-upworthy/359541/

A CASE FOR CURIOSITY

- Clickbait erodes trust in the website*
- Clickbait damages perceived integrity of journalism**

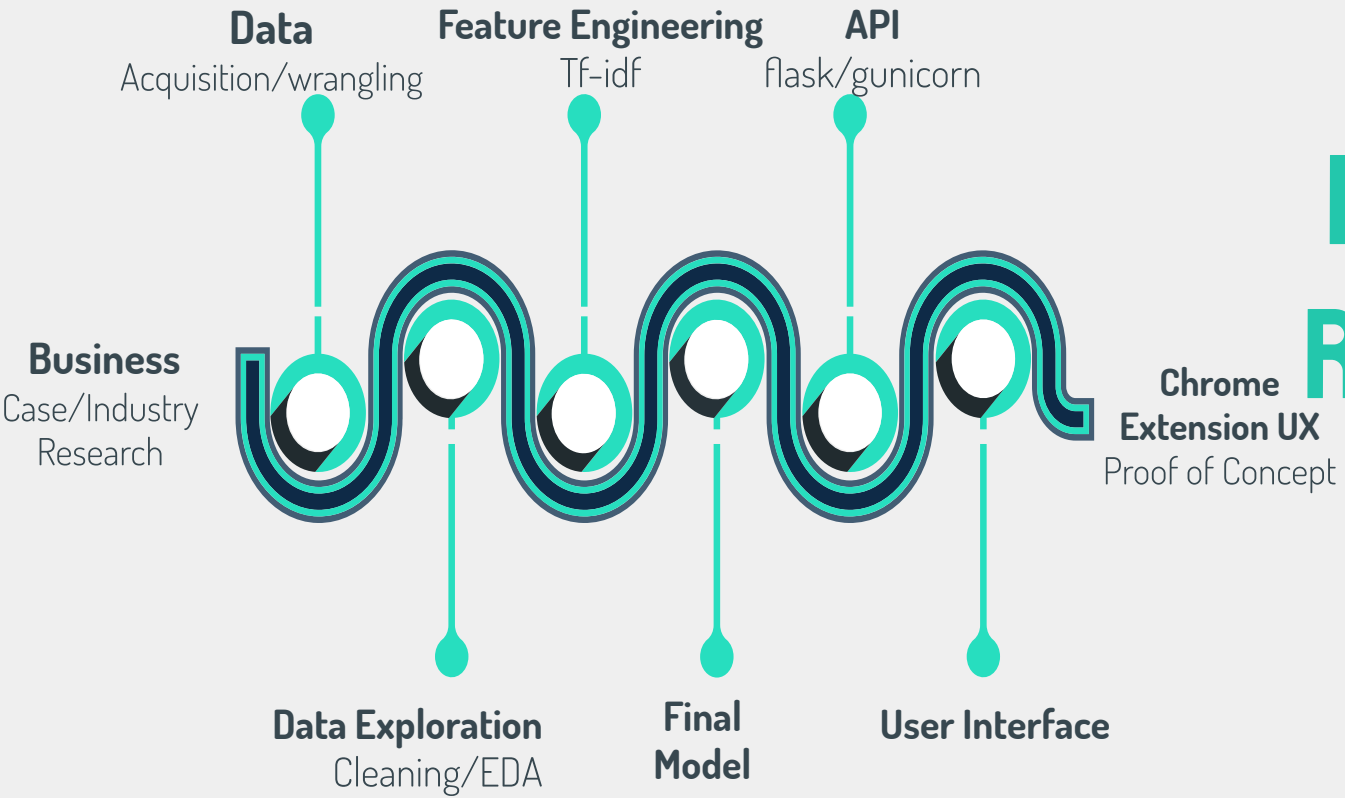
"Curiosity is The Impulse Towards Better Cognition"

—WILLIAM JAMES

*Fred Vultee, G. Scott Burgess, Darryl Frazier, Kelsey Mesmer. (2020) Here's What to Know About Clickbait: Effects of Image, Headline and Editing on Audience Attitudes. Journalism Practice 0:0, pages 1-18.

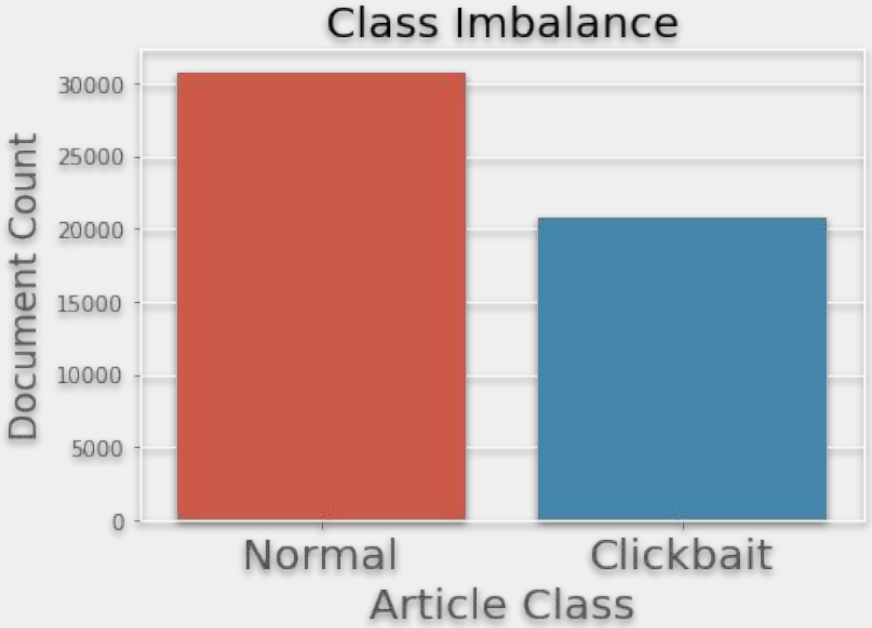
**https://www.bbc.com/news/uk-wales-34213693

PROJECT ROADMAP



pandas NumPy NLTK TextBlob matplotlib seaborn scikit learn Flask HEROKU Bootstrap JS

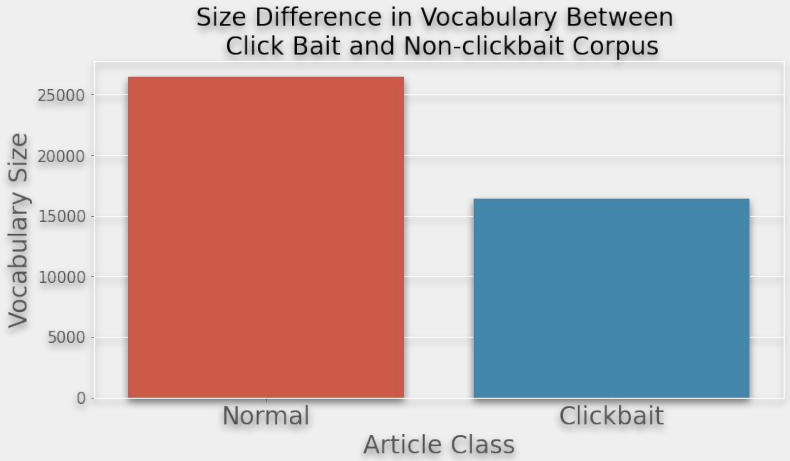
DATA

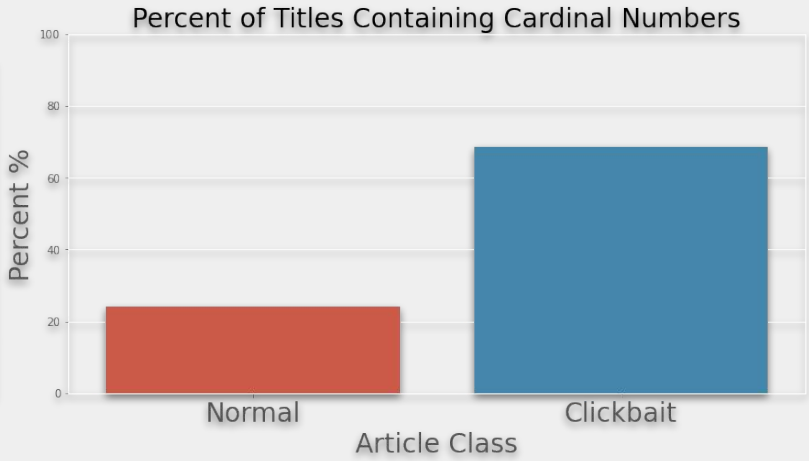


- Final Datasets came from Indian Institute of Technology (2016) and Webis, 2017
- Final dataframe had 51538 news headlines: 20760 clickbait and 30778 normal
- 3:2 class imbalance in the final dataset ameliorated via upsampling

* Amazon Mechanical Turks. A Crowd-sourcing marketplace. For more information visit https://www.mturk.com

DATA



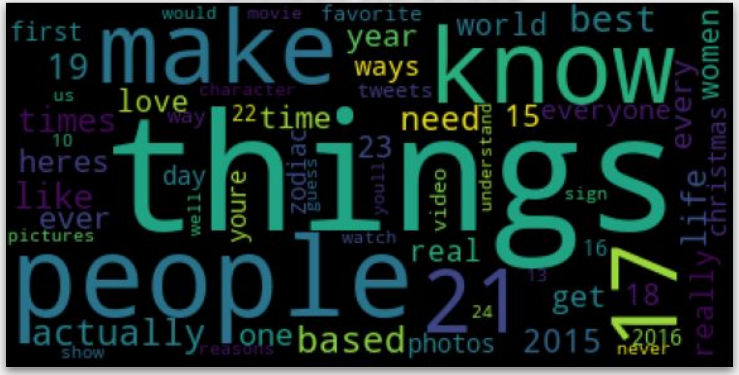


- 10000+ unique normal words and 5000+ unique clickbait words.
- 70% of all clickbait headlines in the corpus contain cardinal numbers

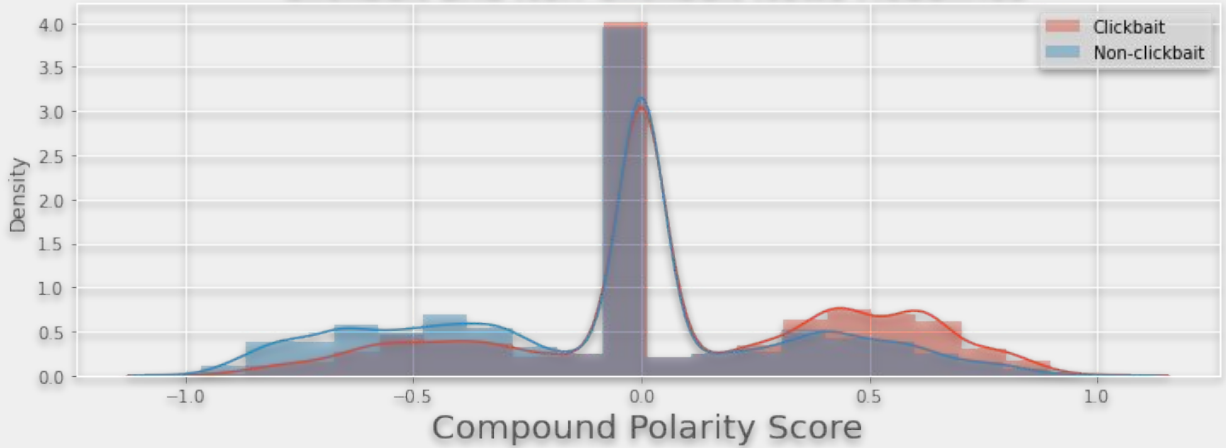
DATA

Normal WordCloud

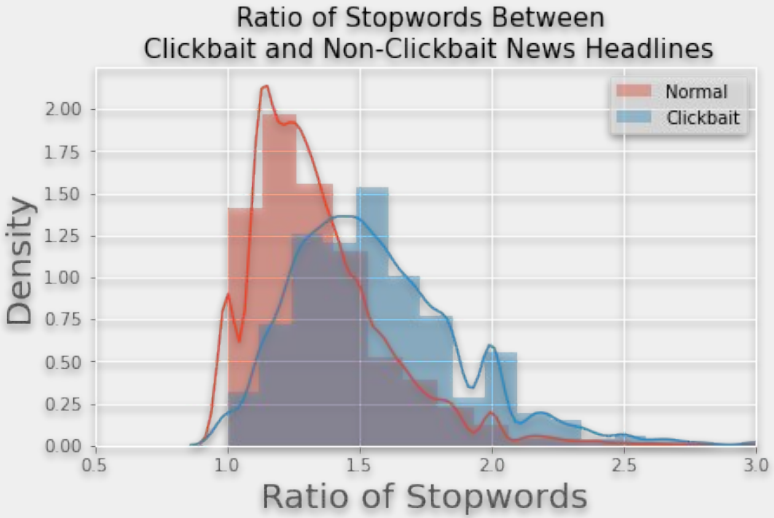
Clickbait WordCloud

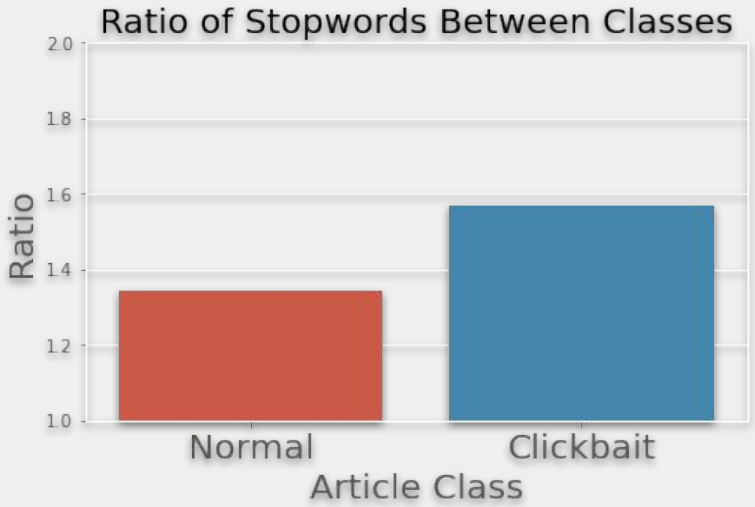


Distribution of Polarity Scores Between
Clickbait and Non-Clickbait News Headlines



DATA

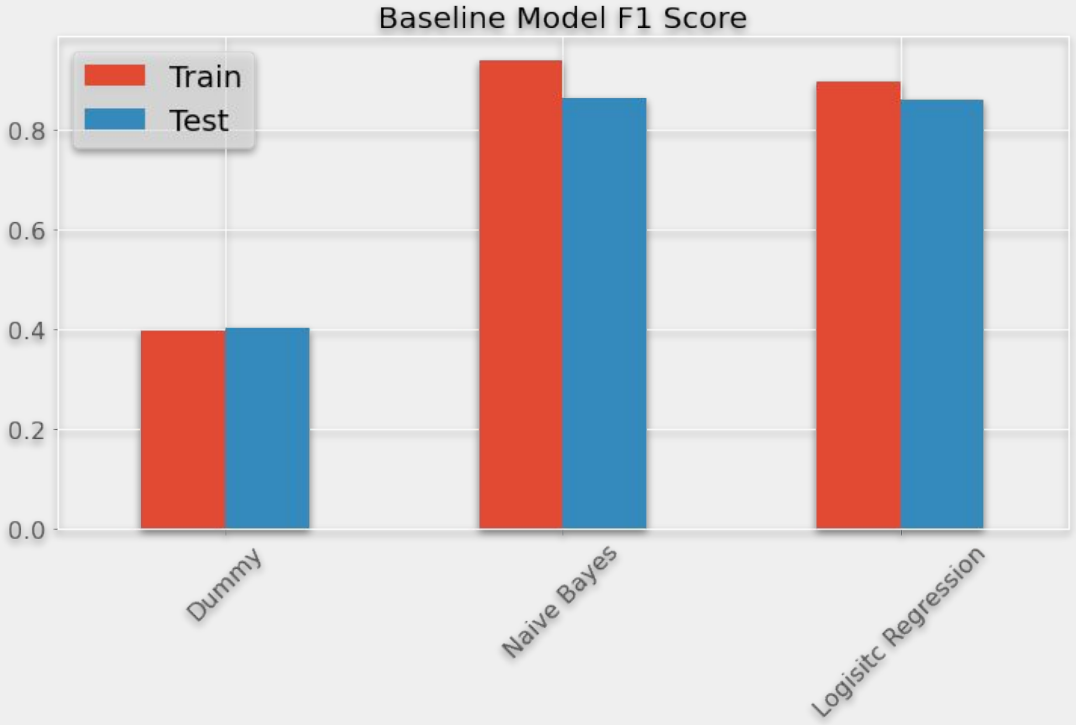




- Stop words are common english words that don't confer a lot of meaning*
  - "The, and, do...etc"

*NLTK has a set of stopwords that I used in addition to some others that I discovered through EDA. More information at: https://www.nltk.org/index.html

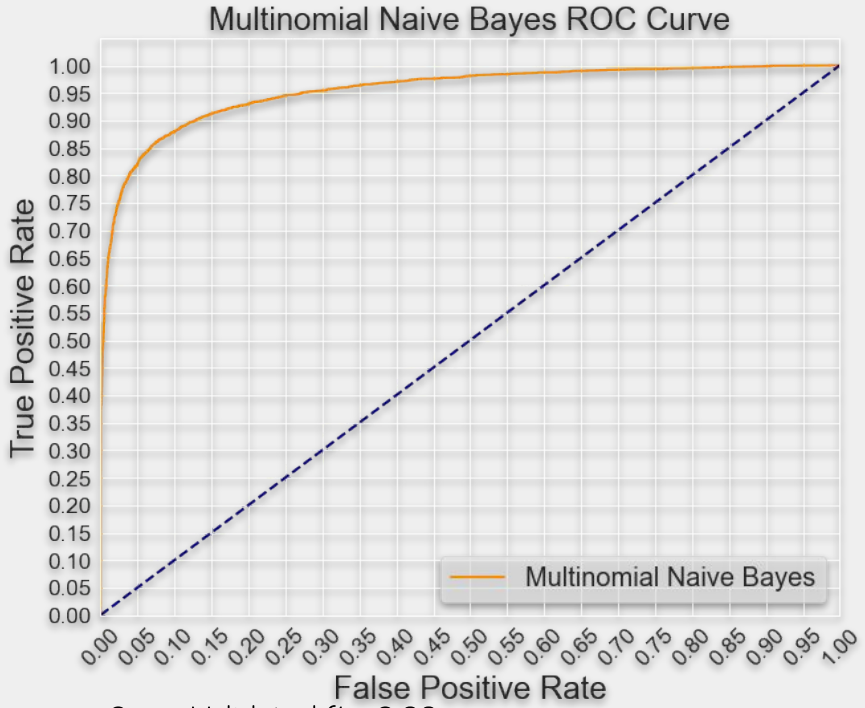
BASELINE MODELS



MODELING

- Bag of Words (BoW)
- tf-idf Vectorizer

FINAL MODEL

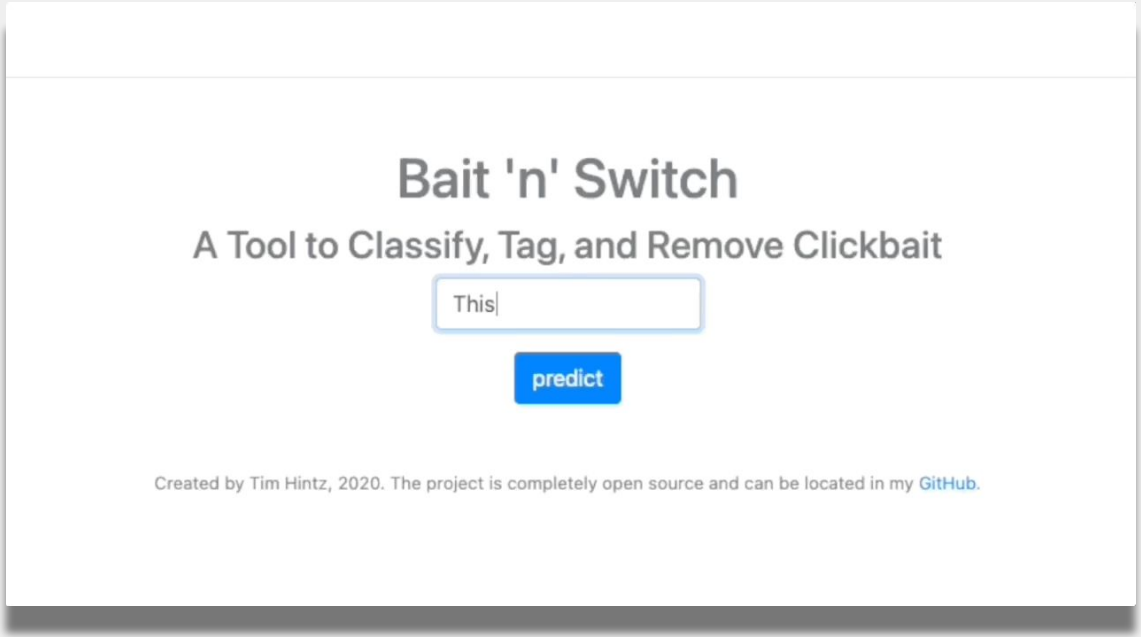


- Cross Validated f1 = 0.88
- ROC-AUC = 0.96
- Accuracy score = 0.89

MODELING

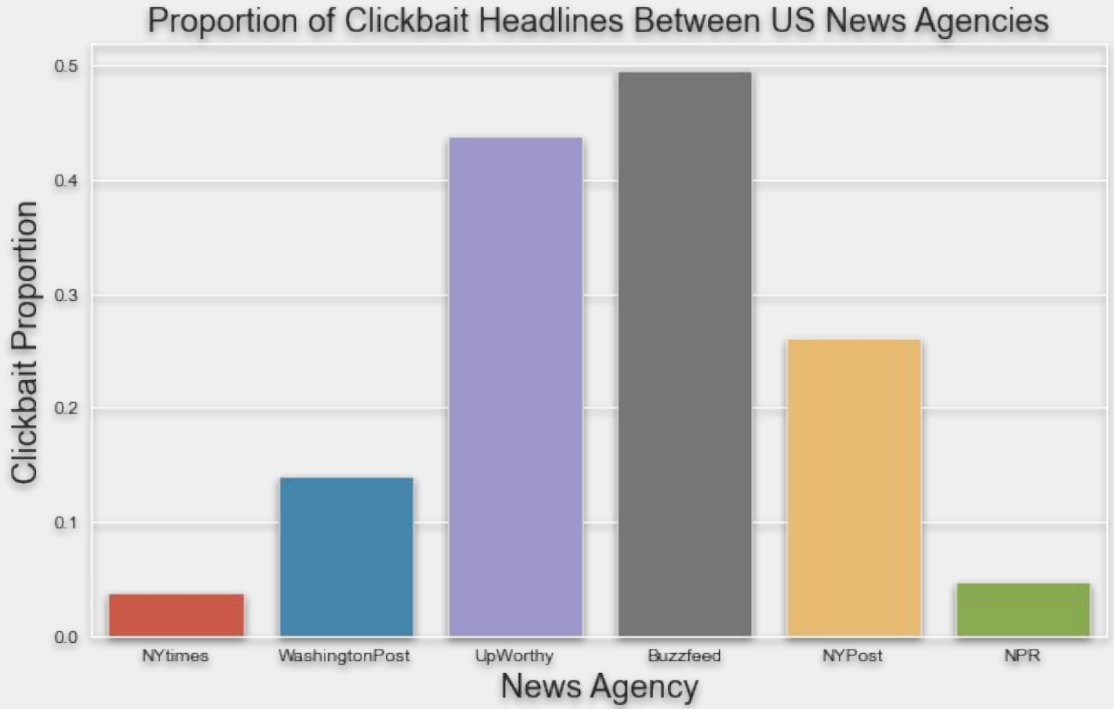
- Multinomial Naive Bayes
- Tf-idf Vectorized

MODEL DEPLOYMENT



Frontend UX (GET): https://baitnswitch.herokuapp.com
API endpoint (POST: "headline"): https://baitnswitch.herokuapp.com/apiendpoint

MODEL DEPLOYMENT



**Article headlines scraped and evaluated by the Bait 'n' Switch API.*

MODEL DEPLOYMENT

- Browser extension is still in prototype phase
- I can establish an event listener but I cannot connect it to my API

NEXT STEPS

FINISH CHROME EXTENSION

- Establish a connection to the Bait'n'Switch API
- Parse Response
- Display predicted results

RECOMMENDATION RATHER THAN CLASSIFICATION

- Clickbait is a subjective classification problem
- Create personalised filtering based on personal preference

RETHINK MY CLICKBAIT DEFINITION

- Rather than a subjective classification, topic model on both the headline and article and assess similarity between the two.

BAIT ‘N’ SWITCH

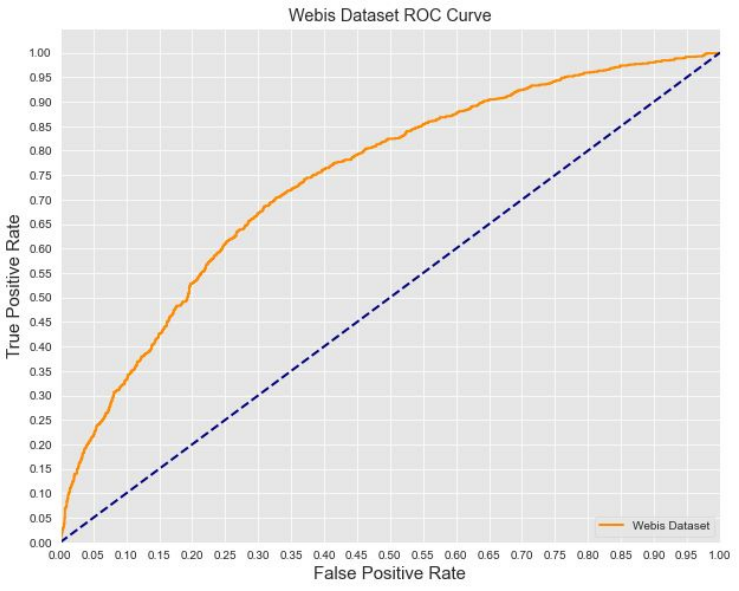
This One Simple Tool That has Marketing Firms Furious

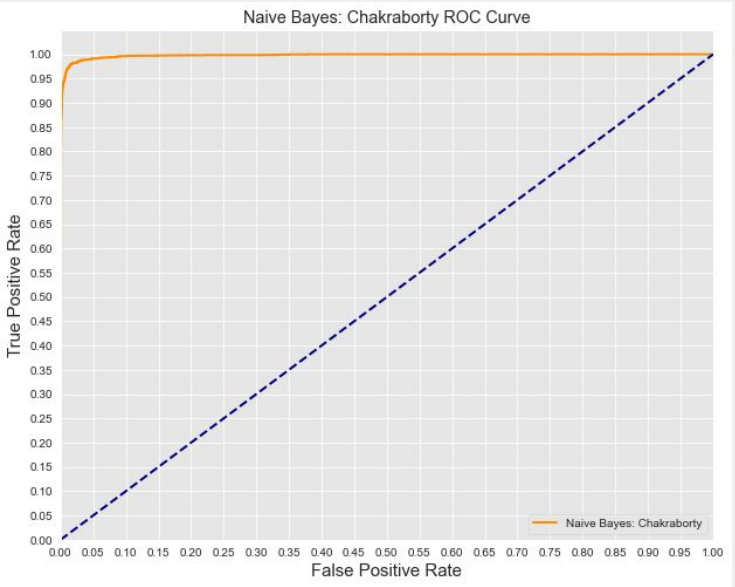
THANKS

Any questions?

Email
GitHub
Linkedin
Bait’n’Switch Web App

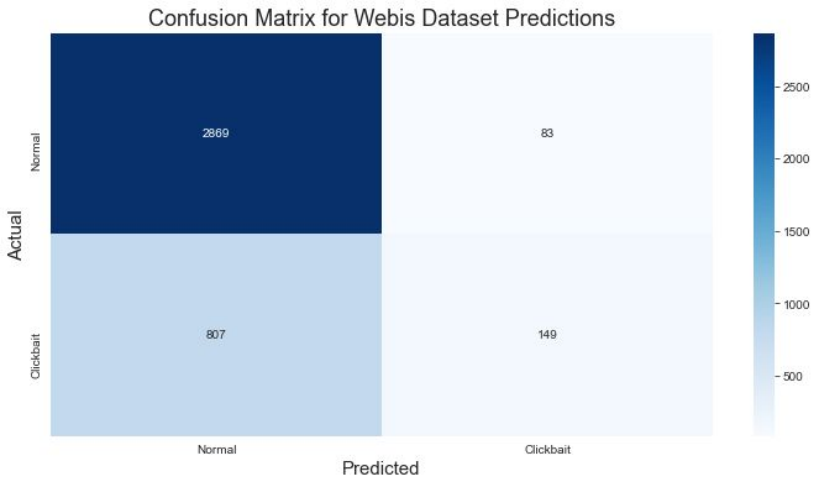
SUPPLEMENTARY SLIDE A

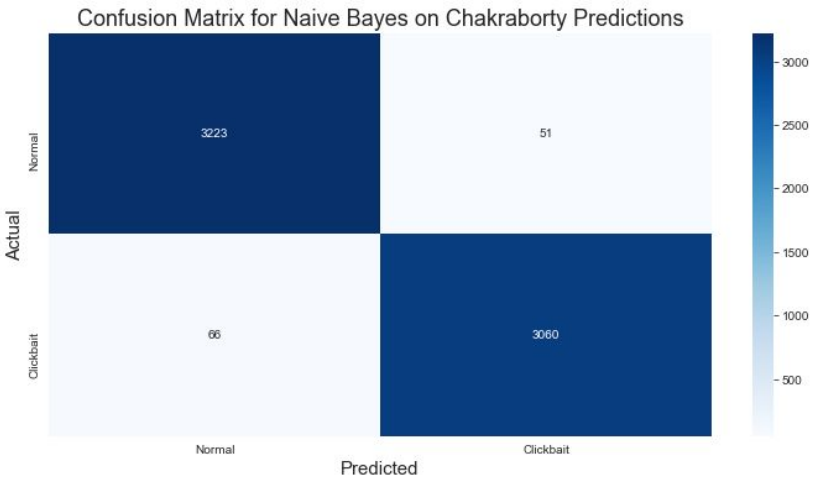




- Webis Dataset was poorly labelled
  - Money versus Self Motivation
  - Cultural Differences
  - Dichotomous versus Likert scale

SUPPLEMENTARY SLIDE B





SUPPLEMENTARY SLIDE C

From Terminal:

```
>>> curl 'http://baitnswitch.herokuapp.com/apiendpoint' --data
'headline="https://nytimes.com"'
{"information": {"date": "2021-01-02", "url": "https://nytimes.com"}, "contents":
{"num_normal": 19, "num_clickbait": 10, "total_headlines": 29}}
```

Python Code:

```
url = 'https://baitnswitch.herokuapp.com/apiendpoint'
postObj = {'headline': 'https://www.nytimes.com/'}
x = requests.post(url, data = postObj)
print(x.text)
{"information": {"date": "2021-01-02", "url": "https://nytimes.com"}, "contents":
{"num_normal": 19, "num_clickbait": 10, "total_headlines": 29}}
```

SUPPLEMENTARY SLIDE D

	Dataset 1 Chakraborty et al., 2016	Dataset 2 Webis, 2016 database	Dataset 3 Webis, 2017 database
Labelling	• 6 Volunteers, • Dichotomous • Hard Voting	• 5 AMT*, • Likert Scale • Soft Voting	• 5 AMT* • Likert Scale • Soft Voting
Content	-16000 clickbait -16000 normal	-767 clickbait -2225 normal	-9276 clickbait -29241 normal
Source	6 US News Agencies; 2016	Twitter Post text; 2016	Twitter Post URL; 2017

- Included Datasets 1 and 3. Dataset 2 was purely twitter posts rather than news headlines
- Final dataframe had 51538 news headlines: 20760 clickbait and 30778 normal
- 3:2 class imbalance in the final dataset