NEWS SATIRE DETECTOR

capstone project standup

08/31/20

Gabriela Tanumihardja

All Generalizations are Wrong, Including This One…

“‘Honesty is the best policy.’ O.K.! Now, what’s the second-best policy?”

News Satire:

- Designed to make commentaries on various topics and their vices and errors through humour
- Uses sarcasms, ironies, parodies, and exaggerations
- NOT intended to mislead

Often mistaken as legitimate news articles…

Data Collection

Satirical and Legitimate News:

- Scraped from four news sources
- Headlines, topic, and published date
- Data scraped using Selenium and BeatifulSoup packages

"I'll have the misspelled 'Ceasar' salad and the improperly hyphenated veal osso-buco."

BEFORE CLEANING

- 55,925 headlines
- Published between 1993 – 2020
- 59% legitimate news, 41% satirical news

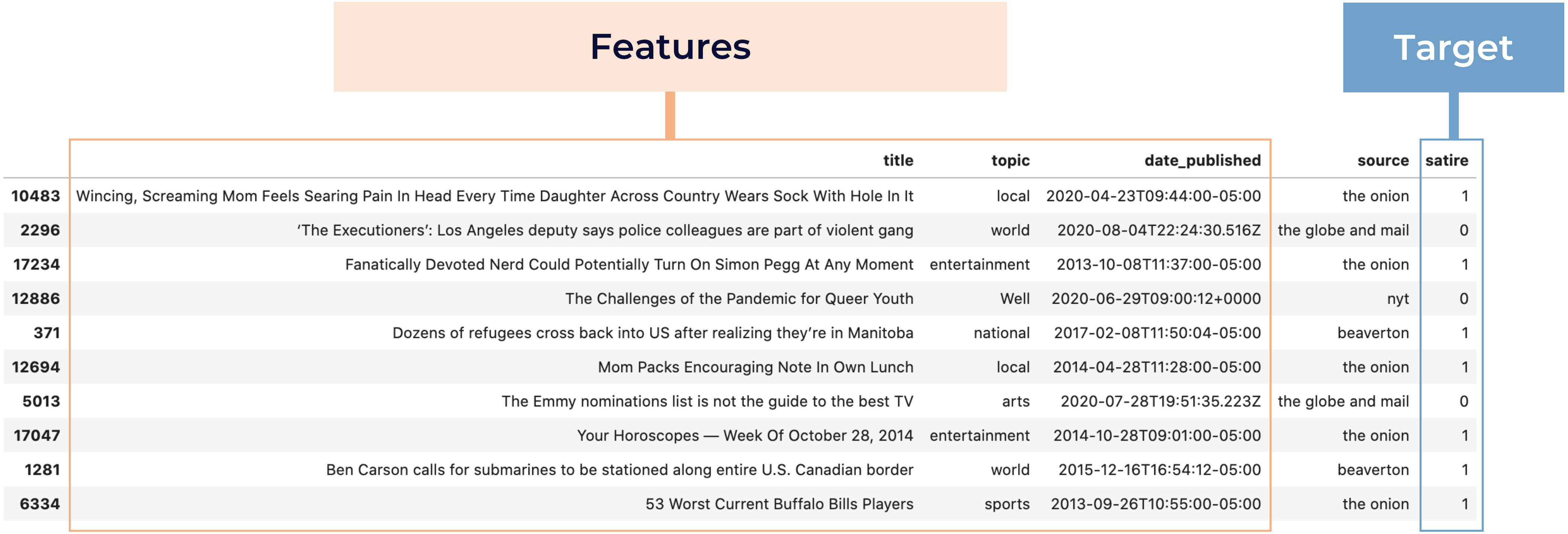
THE GLOBE AND MAIL

THE BEAVERTON

the ONION

The New York Times

Data Description



Features

Target

	title	topic	date_published	source	satire
10483	Wincing, Screaming Mom Feels Searing Pain In Head Every Time Daughter Across Country Wears Sock With Hole In It	local	2020-04-23T09:44:00-05:00	the onion	1
2296	‘The Executioners’: Los Angeles deputy says police colleagues are part of violent gang	world	2020-08-04T22:24:30.516Z	the globe and mail	0
17234	Fanatically Devoted Nerd Could Potentially Turn On Simon Pegg At Any Moment	entertainment	2013-10-08T11:37:00-05:00	the onion	1
12886	The Challenges of the Pandemic for Queer Youth	Well	2020-06-29T09:00:12+0000	nyt	0
371	Dozens of refugees cross back into US after realizing they’re in Manitoba	national	2017-02-08T11:50:04-05:00	beaverton	1
12694	Mom Packs Encouraging Note In Own Lunch	local	2014-04-28T11:28:00-05:00	the onion	1
5013	The Emmy nominations list is not the guide to the best TV	arts	2020-07-28T19:51:35.223Z	the globe and mail	0
17047	Your Horoscopes — Week Of October 28, 2014	entertainment	2014-10-28T09:01:00-05:00	the onion	1
1281	Ben Carson calls for submarines to be stationed along entire U.S. Canadian border	world	2015-12-16T16:54:12-05:00	beaverton	1
6334	53 Worst Current Buffalo Bills Players	sports	2013-09-26T10:55:00-05:00	the onion	1

Cleaning and Preliminary Modelling

CLEANING AND EDA

- Removed duplicate entries
- Removed entries with null values
- Identified some titles that may skew the predictions
  - e.g. ‘Onion’ and ‘Horoscope’

AFTER CLEANING

- 46,098 headlines
- 50% legitimate news, 50% satirical news

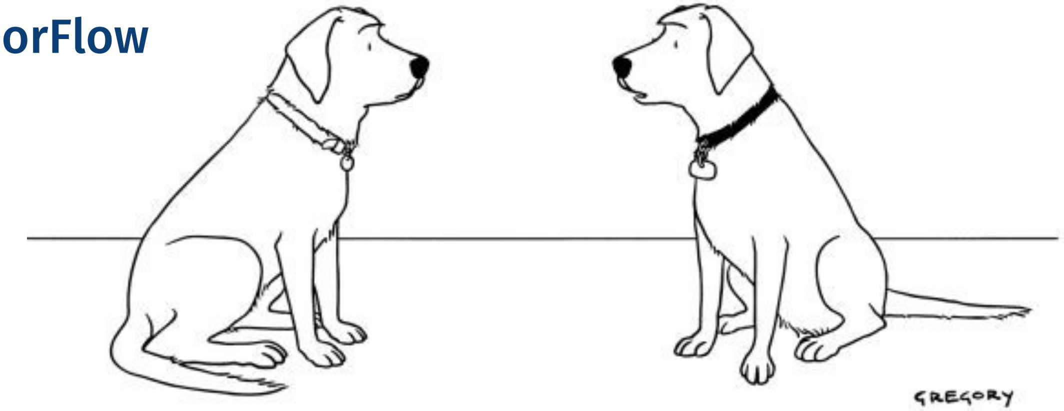
PRELIMINARY MODELLING

- Count vectorizers & TF-IDF
- Logistic Regression
  - 88% training / 83% validation accuracy
- KNN
  - Performed with scaling and PCA
  - 85% training / 70% validation accuracy
- SVM (kernel = linear and rbf)
  - Linear = 84% training / 81% validation accuracy
  - RBF = 92% training / 83% validation accuracy
- Decision tree

What’s Next?

Plans for the next 2 weeks:

- Finesse dataset – replace common terms and add layers of stratification
- Re-run models – logit, KNN, decision tree, SVM
- Fit a random forest model
- Hyperparameter optimization – SVM is showing potential
- Build and fit deep learning models using TensorFlow



“I had my own blog for a while, but I decided to go back to just pointless, incessant barking.”