NATURAL LANGUAGE PROCESSING

Fake News Detection

Jasmine Vasandani

Tips to help prevent the spread of rumors & fake news*

1. Understand when a message is forwarded
2. Check photos and media carefully
3. Look out for messages that look different
4. Check your biases
5. Fake news often goes viral
6. Verify with other sources
7. Help stop the spread

Source

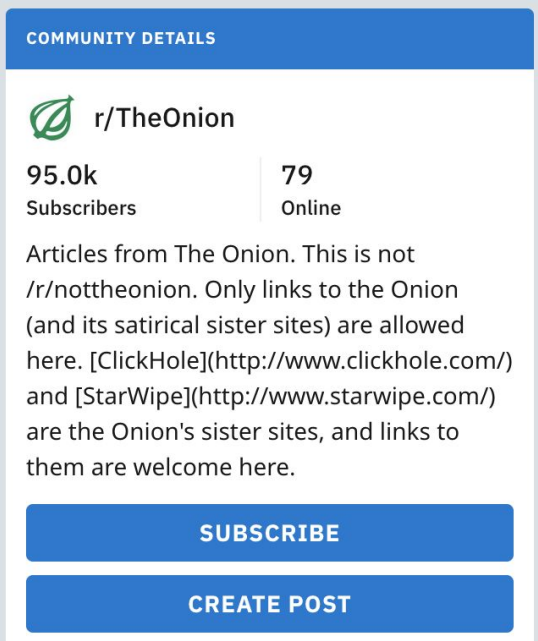
Source

How might Natural Language Processing better inform us about fake news?

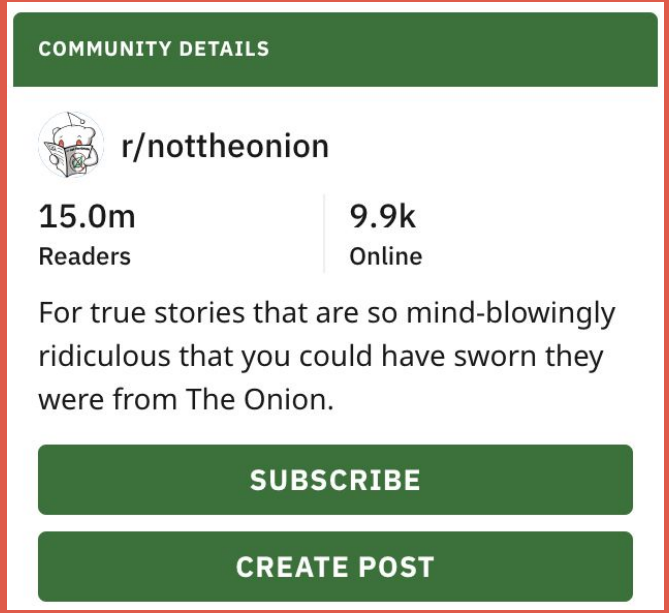
CASE STUDY:

Distinguish between **fake news** and **absurd news** by building a classification model trained on Subreddit posts from r/TheOnion and r/nottheonion.

r/TheOnion



r/nottheonion



Methodology

1

DATA ACQUISITION

Scrape 15k posts from r/TheOnion & r/nottheonion, total of 30k posts. Used pushshift.io API wrapper to acquire data.

2

EXPLORATORY DATA ANALYSIS

Create data visualizations to observe trends about the data. What are distinguishing characteristics about each subreddit?

3

NATURAL LANGUAGE PROCESSING

Prepare data for modeling. Remove punctuation & #s. Drop duplicates. All lowercase.

4

CLASSIFICATION MODELING

Find the best combination of vectorizers and models to achieve the best accuracy when designating a post as either from r/TheOnion or r/nottheonion.

Data Acquisition

- **29,867 posts**, 15k from r/nottheonion, rets r/TheOnion
  - After duplicates, 26,275 posts
  - Features: title (predictor), subreddit (target), comments, author, # sub, score, domain, date

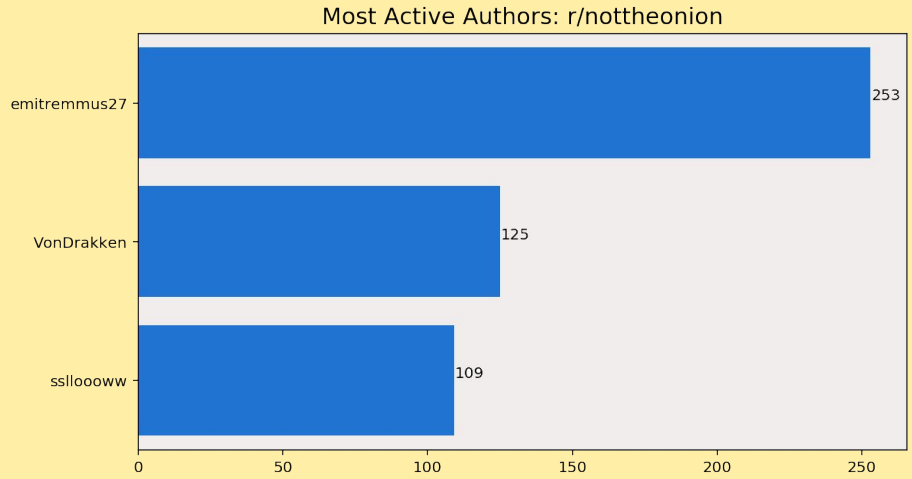
- **Posts Date Range:**
  - r/TheOnion: Oct 4, 2013 - Apr 7, 2019
  - r/nottheonion: Jan 25, 2019 - Apr 7, 2019

- **Data Cleaning:**
  - numbers
  - punctuation
  - case

“

The first biggest mistake in data science is looking [at] every business challenge as [a] predictive approach. Remember 70% of low hanging problems can be solved by just doing an **EDA**.

Sundar Ramamurthy

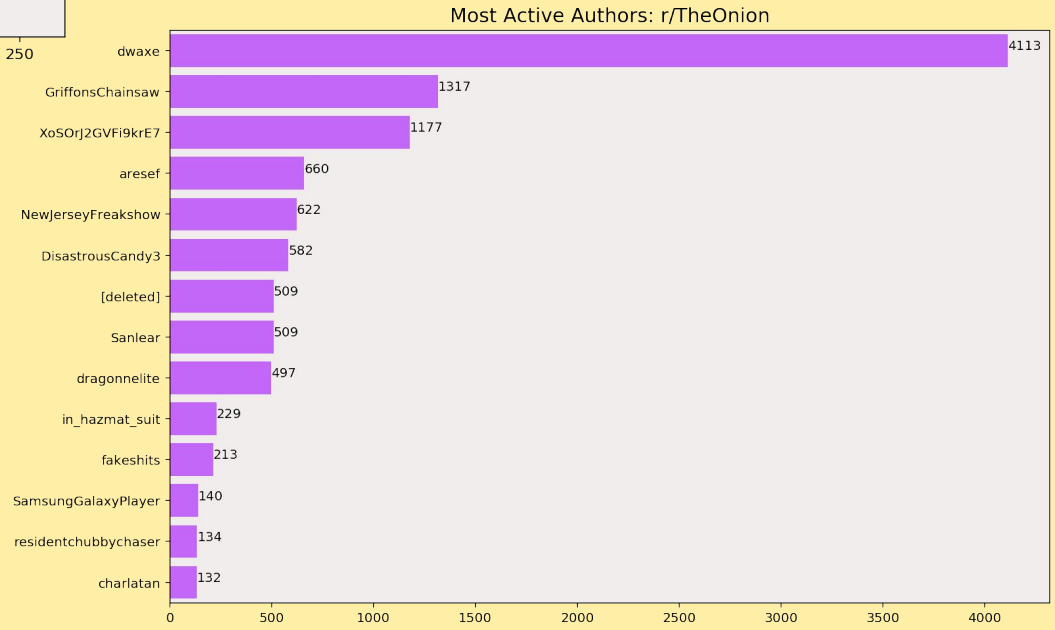


ACTIVE AUTHORS | r/nottheonion

15m subscribers, only 3 authors shared over 100 posts.

ACTIVE AUTHORS | r/TheOnion:

95k subscribers, 14 authors shared over 100 posts, the most being 4,113 posts.

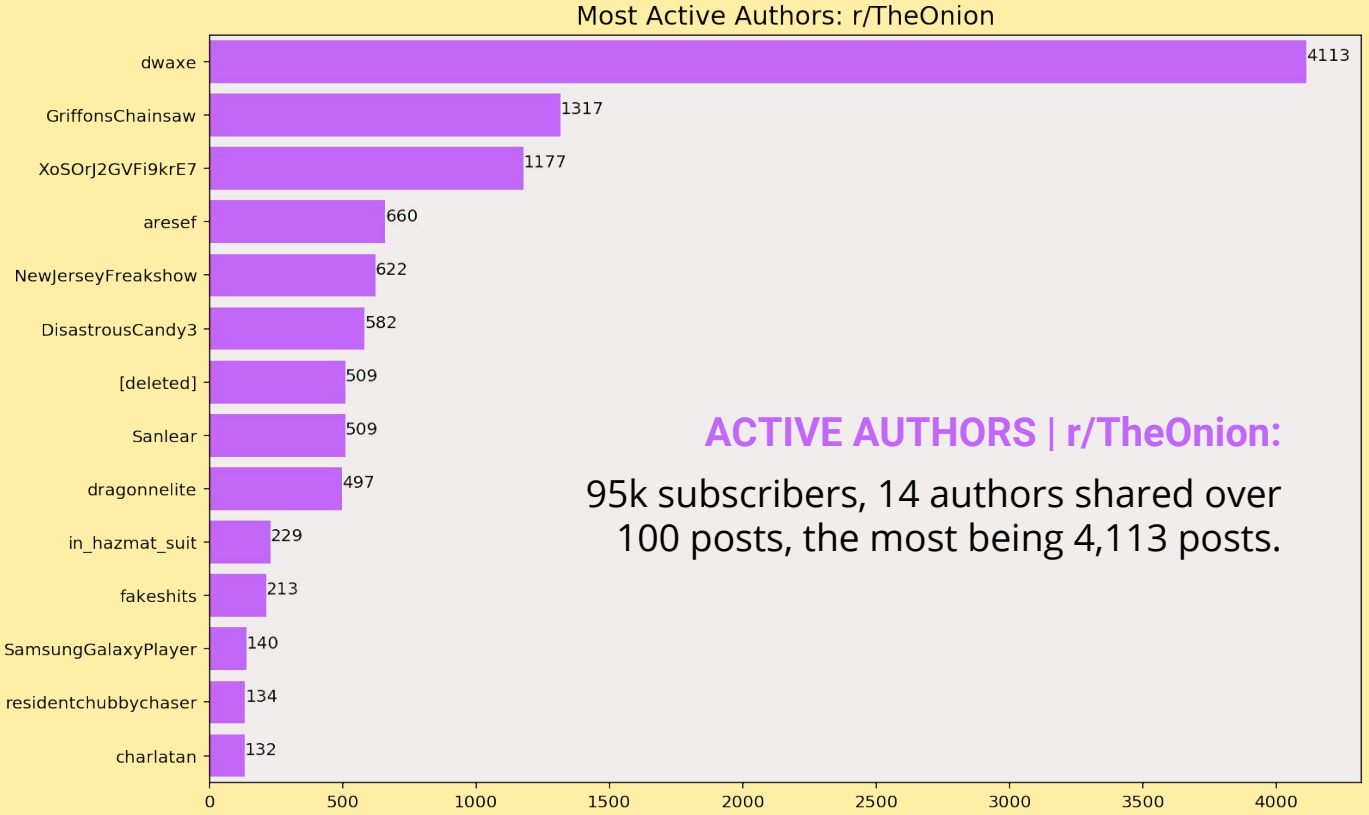


FAKE NEWS PREVENTION VIOLATIONS:

- Understand when a message is forwarded

- Fake news often goes viral

- Help stop the spread



ACTIVE AUTHORS | r/TheOnion:

95k subscribers, 14 authors shared over 100 posts, the most being 4,113 posts.

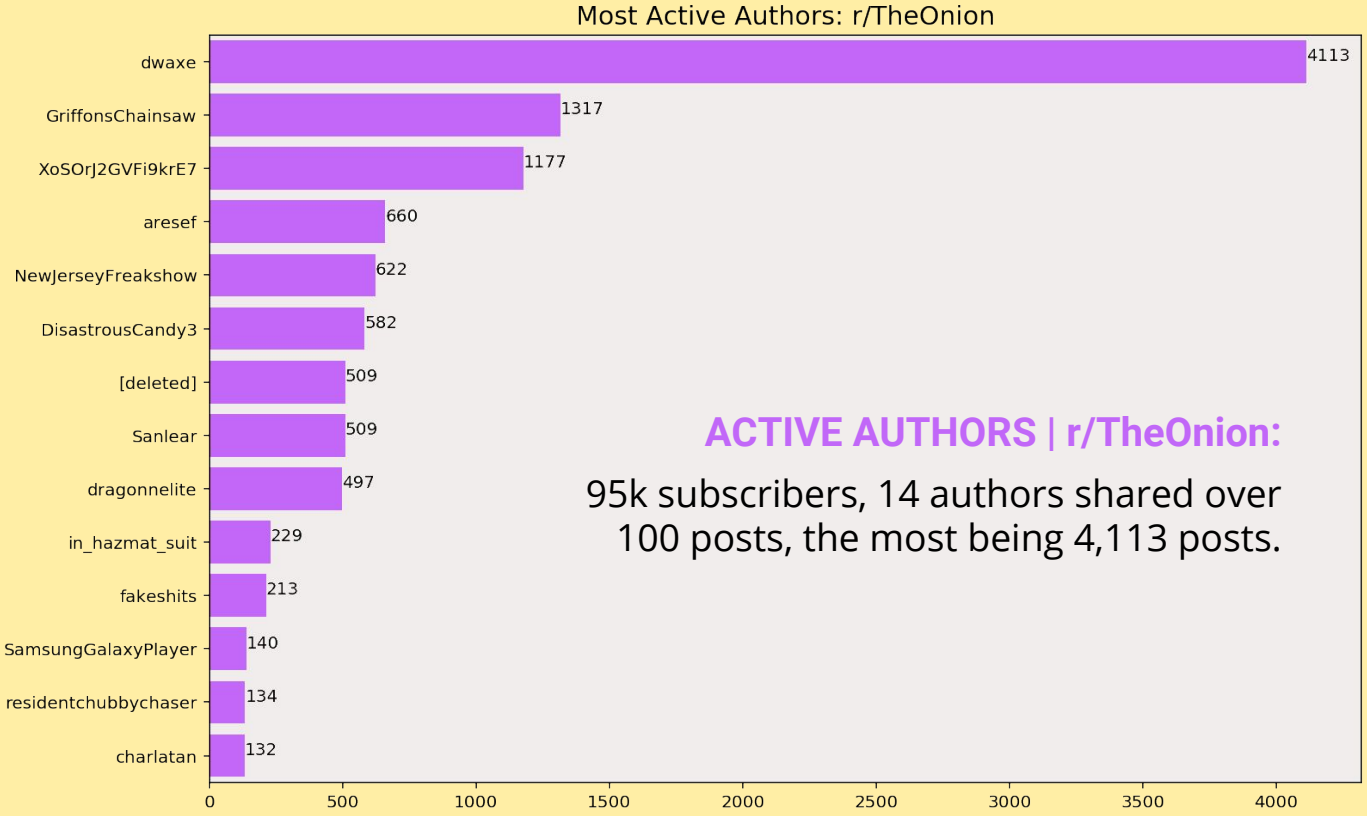
FAKE NEWS PREVENTION VIOLATIONS:

- Understand when a message is forwarded

- Fake news often goes viral

- Help stop the spread

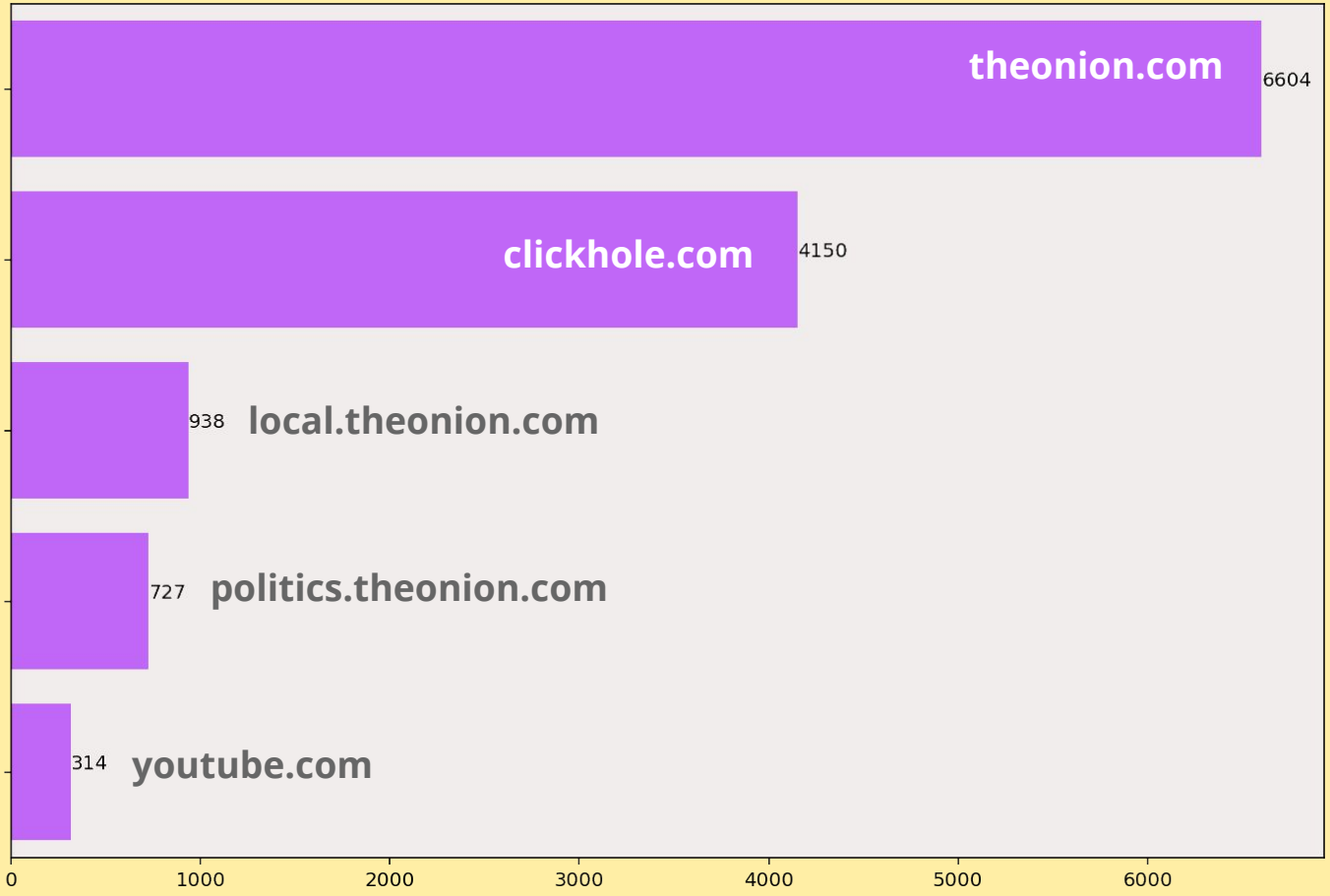
r/TheOnion is a good case study to analyze fake news trends.



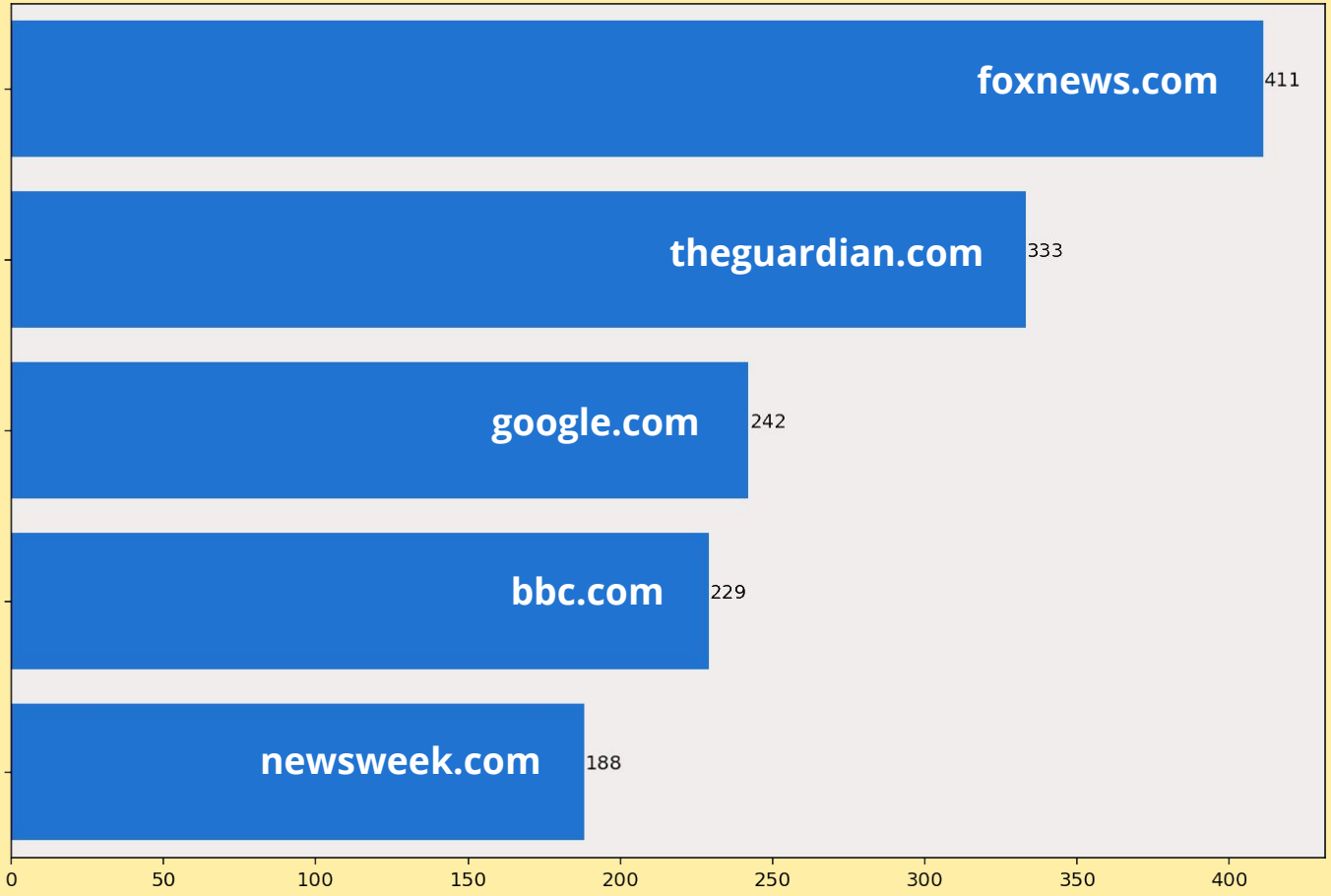
ACTIVE AUTHORS | r/TheOnion:

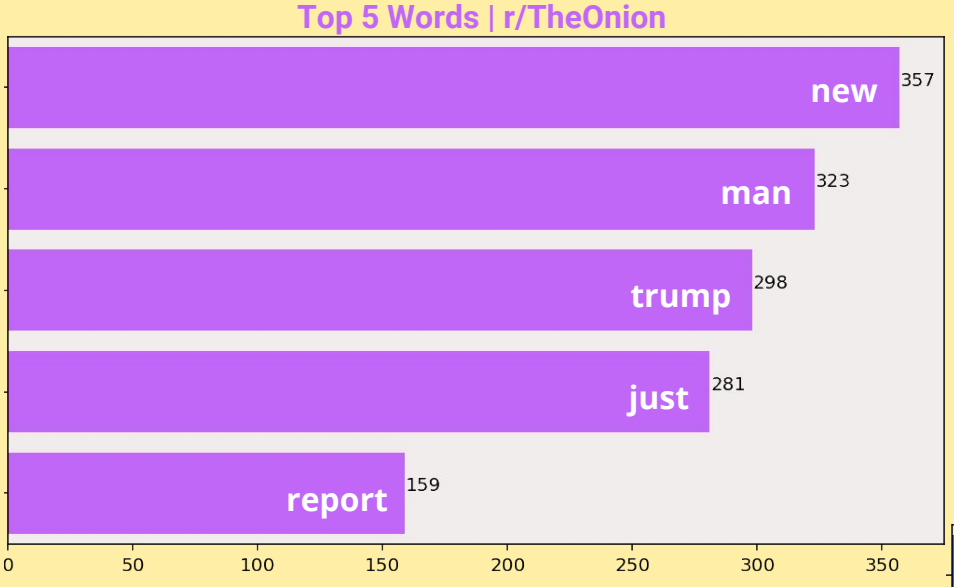
95k subscribers, 14 authors shared over 100 posts, the most being 4,113 posts.

TOP 5 MOST REFERENCED DOMAINS | r/TheOnion



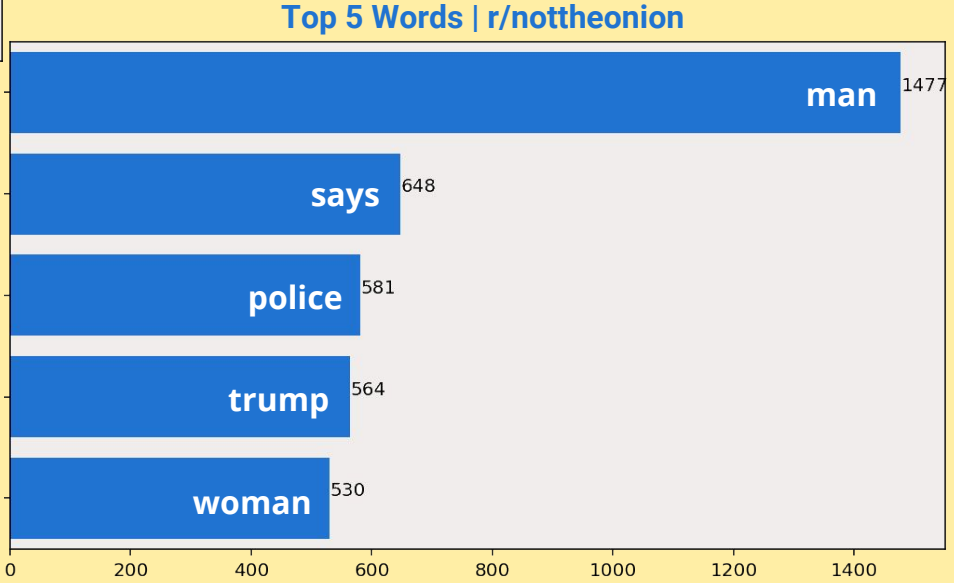
TOP 5 MOST REFERENCED DOMAINS | r/nottheonion

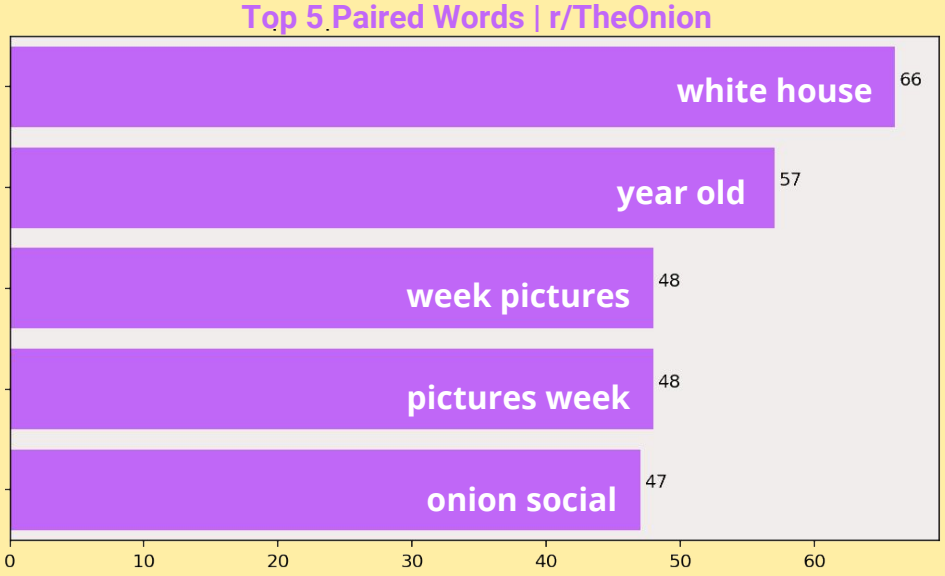



Top 5 Words | r/TheOnion
new
357
man
323
trump
298
just
281
report
159
0
50
100
150
200
250
300
350

MAN

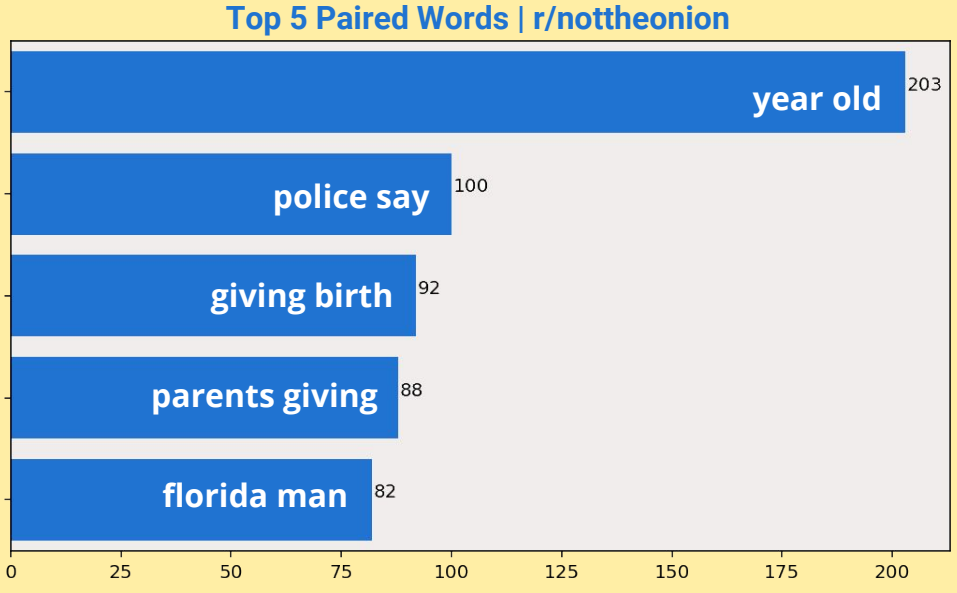
TRUMP


Top 5 Words | r/nottheonion
man
1477
says
648
police
581
trump
564
woman
530
0
200
400
600
800
1000
1200
1400


Top 5 Paired Words | r/TheOnion
white house
66
year old
57
week pictures
48
pictures week
48
onion social
47
0
10
20
30
40
50
60

YEAR

OLD


Top 5 Paired Words | r/nottheonion
year old
203
police say
100
giving birth
92
parents giving
88
florida man
82
0
25
50
75
100
125
150
175
200

MODELING

Pipeline & GridSearchCV

CountVectorizer()
stop_words: None, ‘english’, custom
ngram_range: (1,1), (2,2), (1,3)

LogisticRegression()
C: 0.01, 1

CountVectorizer()
stop_words: None, ‘english’, custom
ngram_range: (1,1), (2,2), (1,3)

MultinomialNB()
C: 0.01, 0.36, 1

TfidfVectorizer()
max_df: .75, .98, 1.0
min_df: 2, 3, 5
ngram_range: (1,1), (1,2), (1,3)

LogisticRegression()
C: 1

TfidfVectorizer()
max_df: .75, .98, 1.0
min_df: 4, 5
ngram_range: (1,1), (1,2), (1,3)

MultinomialNB()
C: 0.1, 1, 100

Best Accuracy Score: 90%

CountVectorizer()
stop_words: None, 'english', custom
ngram_range: (1,1), (2,2), (1,3)

LogisticRegression()
C: 0.01, 1

BEST TEST SCORE: 0.87
Train score: 0.96

CountVectorizer()
stop_words: None
ngram_range: (1,3)

MultinomialNB()
C: 0.36

BEST TEST SCORE: 0.90
Train score: 0.997

TfidfVectorizer()
max_df: .75, .98, 1.0
min_df: 2, 3, 5
ngram_range: (1,1), (1,2), (1,3)

LogisticRegression()
C: 1

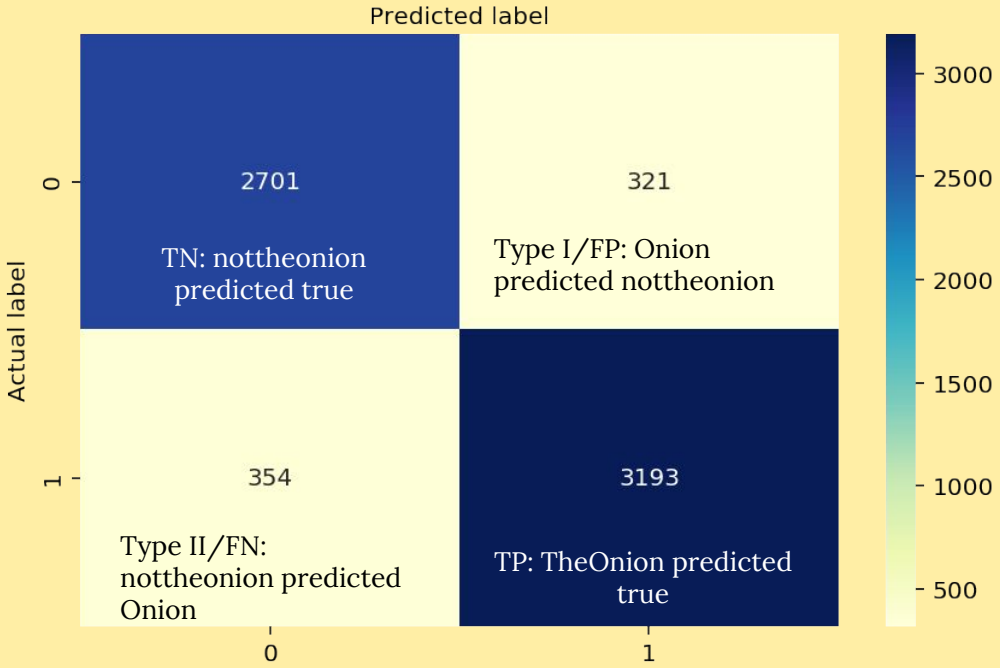
BEST TEST SCORE: 0.86
Train score: 0.90

TfidfVectorizer()
max_df: .75, .98, 1.0
min_df: 4, 5
ngram_range: (1,1), (1,2), (1,3)

MultinomialNB()
C: 0.1, 1, 100

BEST TEST SCORE: 0.86
Train score: 0.92

Model 3: Confusion Matrix



Accuracy
For all predictions, 90% predicted correctly.

Positive Predictive Rate
Among all Onion predictions, 91% predicted correctly.

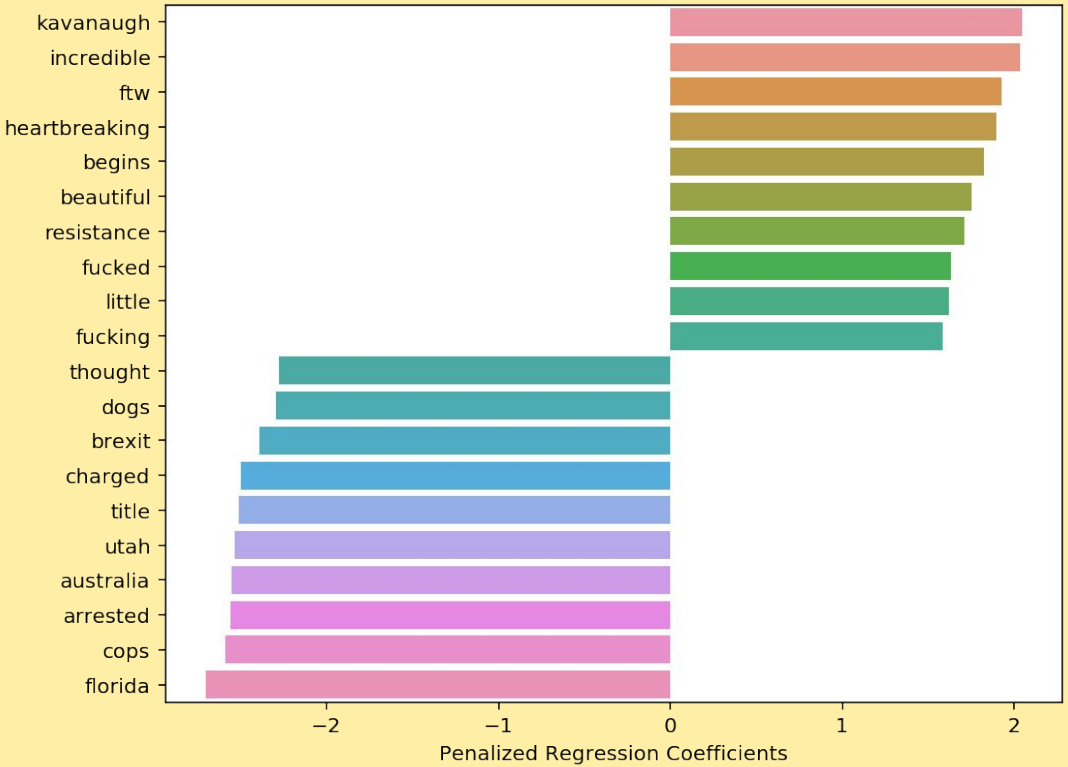
True Positive Rate
Among all Onion titles, 90% predicted correctly.

True Negative Rate
Among all nottheonion, 89% predicted correctly.

Misclassification Rate
For all predictions, 11% predicted incorrectly.

Logistic Regression Coefficients

MODEL 3



r/TheOnion

The word that contributes the most positively to being from r/TheOnion is 'kavanaugh', followed by the 'incredible' and 'ftw'.

- *As occurences of "kavanaugh" increases by 1 in a title, that title is 7.7 times as likely to be classified as r/TheOnion.*

r/nottheonion

The word that contributes the most positively to being from r/nottheonion is 'Florida', followed by 'cops' and 'arrested'.

- *As occurences of "florida" increases by 1 in a title, that title is 14.9 times as likely to be classified as r/nottheonion.*

Reflections

- NLP of text is one way to analyze fake news, but a **major gap** exists:
  - ***image & video analysis***

r/IAmA

- Ask me anything!

Source.