NLP Modeling with reddit

Predicting Scientific Rigor in Blog Posts

Veronica Giannotta
12/21/2018

Reddit is Kind of a Big Deal

Reddit is Kind of a Big Deal

#5 330M+ 138K+ 14B

Reddit is Kind of a Big Deal

#5

Most visited site in US (on Alexa)

330M+

Average monthly active users

138K+

Active communities

14B

Average screenviews per month

Stats courtesy of reddit Nov.12, 2017

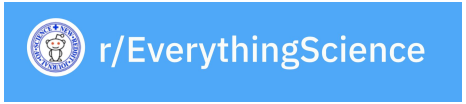
Given the prominence of social media as a communication platform, **can a computer use text from blog posts to distinguish a substantiated claim from an unsubstantiated one?**

Classes

Classes

R/SCIENCE RULES

1. Must be peer-reviewed research
2. No second-hand summaries, reviews, or reposts
3. No editorialized, sensationalized or biased titles
4. Research must be <6 months old
5. No off-topic comments
6. No jokes or memes
7. No abusive or offensive comments
8. No anecdotal comments
9. Not scientific or dismissive of established work
10. No medical advice



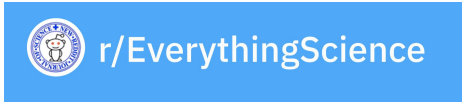
R/EVERYTHINGSCIENCE RULES

1. Be civil
2. Maintain scientific integrity
3. Up to date content
4. No link dumping
5. No reposts
6. No misleading, inaccurate or clickbait titles
7. No rehosted content
8. No spam
9. No promotional material
10. No audio visual material

Classes

R/SCIENCE RULES

1. Must be peer-reviewed research
2. No second-hand summaries, reviews, or reposts
3. No editorialized, sensationalized or biased titles
4. Research must be <6 months old
5. No off-topic comments
6. No jokes or memes
7. No abusive or offensive comments
8. No anecdotal comments
9. Not scientific or dismissive of established work
10. No medical advice



R/EVERYTHINGSCIENCE RULES

1. Be civil
2. Maintain scientific integrity
3. Up to date content
4. No link dumping
5. No reposts
6. No misleading, inaccurate or clickbait titles
7. No rehosted content
8. No spam
9. No promotional material
10. No audio visual material

Classes

The NEW REDDIT JOURNAL of SCIENCE

r/EverythingScience

98

Posted by u/drewiepoodle 9 hours ago

Astronomy Scientists have yet to directly detect dark matter, but dark matter's presence and influence is reflected in the patterns and movement of light. Using data collected by NASA/ESA Hubble Space Telescope, astronomers have developed a new way to "see" dark matter via distant starlight.

spacetelescope.org/news/h...

9 Comments Share Save ...

5

Posted by u/Ajaatshatru34 9 hours ago

Psychology Humans are wired for negativity, for good or ill – Jacob Burak | Aeon Essays

aeon.co/essays...

1 Comment Share Save ...

Methodology

Methodology

- Query post data from the Reddit API
- Clean and transform the text data into a suitable format for modeling in SciKit Learn
- Build and fit multiple NLP models, using only the text from each Reddit post as the predictive variable
- Assess whether the model was successful in distinguishing substantiated claims from unsubstantiated ones
- Examine the strongest text influencers on model performance, and establish next steps for a second iteration of this experiment

Modeling the Text Data

What was in the data set?

Modeling the Text Data

What was in the data set?

- 9 columns and 1916 rows of Reddit post data
- Collected through API queries of each subreddit url
- Text strings and numeric values
- Converted from JSON strings

Cleaning the Data

Cleaning the Data

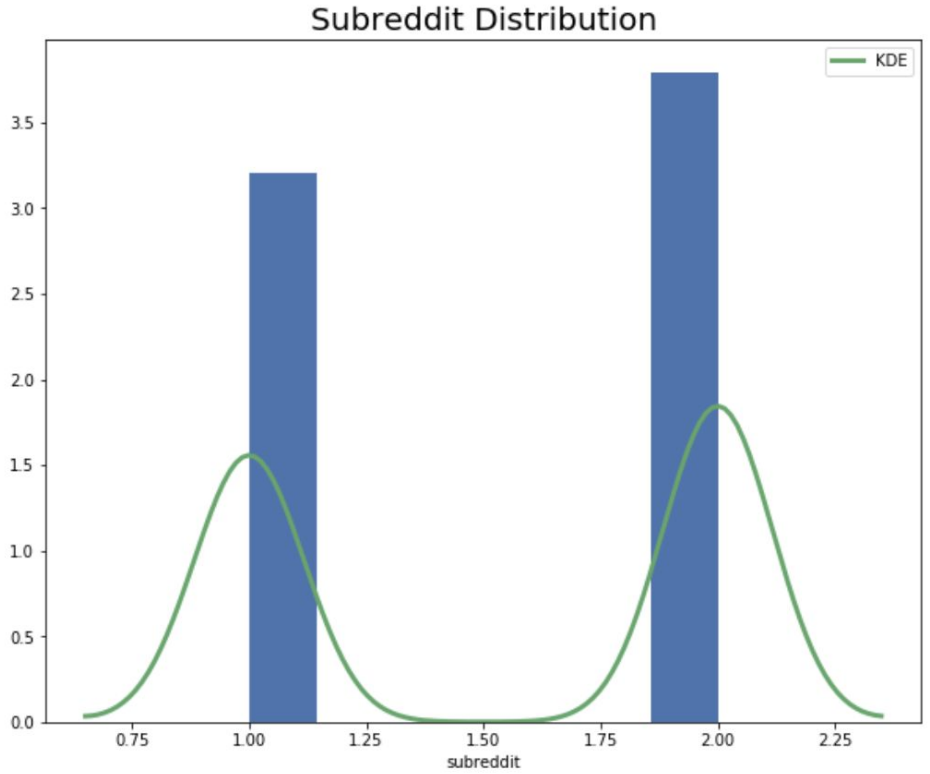
Considerations

- Large number of potentially irrelevant data points
- Punctuation, capitalization, and formatting of text
- Few descriptive features
- Strategy for feature set

Cleaning the Data

Visualizing the Class Distribution

EverythingScience	0.542276
science	0.457724



Cleaning the Data

Steps Taken

- Dropped all columns with null values
- Selected a small subset of the remaining columns
- Created a small number of transformation columns
- Left the ‘title’ column as is

	0
author	neutronfish
domain	worldofweirdthings.com
num_comments	7
title	Scientists and bureaucrats are really, really ...
subreddit	EverythingScience
url	https://worldofweirdthings.com/2018/10/03/the-...
subreddit_class	0
word_count	26
num_stopwords	7

Preprocessing & Model Selection

Preprocessing & Model Selection

Preparing the Data for Modeling

- Grid Searched four models using TF-IDF and Count Vectorizors
  - Logistic Regression
  - Random Forest Classifier
  - Multinomial Naive Bayes
  - Bagging Classifier
- Compared train and test scores
- Chose the two highest performing models: Logistic Regression and Random Forest

Modeling

Modeling

Results

Baseline Accuracy:
.542276

Modeling

Results

Baseline Accuracy:
.542276

Logistic Regression
with TF-IDF Vectorizer

```
LogReg Train Score: 0.9227557411273486
LogReg Test Score: 0.7265135699373695
```

Random Forest
with TF-IDF Vectorizer

```
RF Train Score: 0.9930410577592206
RF Test Score: 0.7202505219206681
```

Modeling

Results

Baseline Accuracy:
.542276

Logistic Regression
with TF-IDF Vectorizer

```
LogReg Train Score: 0.9227557411273486
LogReg Test Score: 0.7265135699373695
```

	pred_neg	pred_pos
actual_neg	195	65
actual_pos	66	153

Random Forest
with TF-IDF Vectorizer

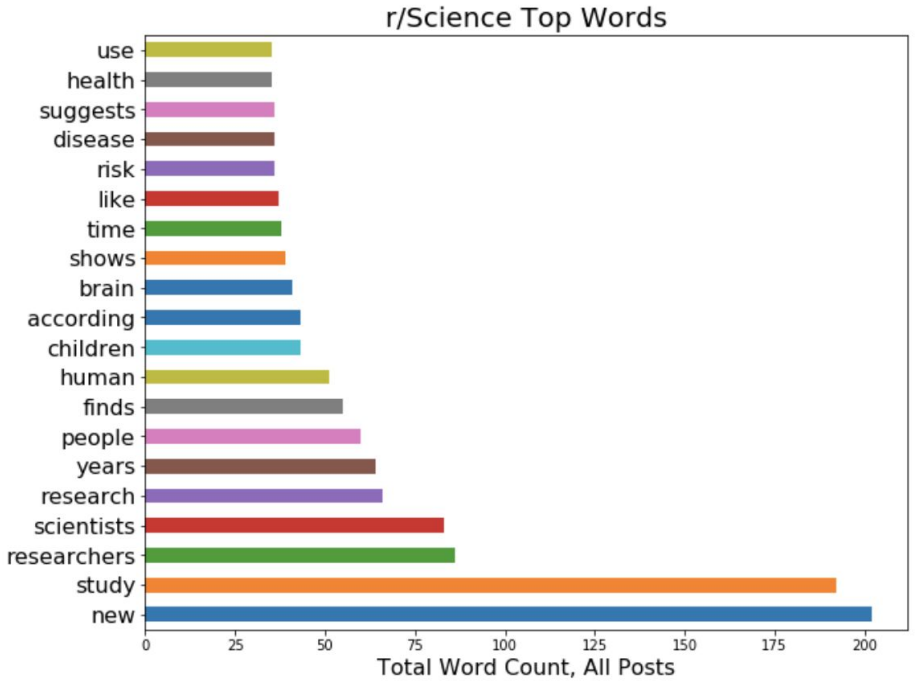
```
RF Train Score: 0.9930410577592206
RF Test Score: 0.7202505219206681
```

	pred_neg	pred_pos
actual_neg	207	53
actual_pos	81	138

Modeling

Best Predictors:

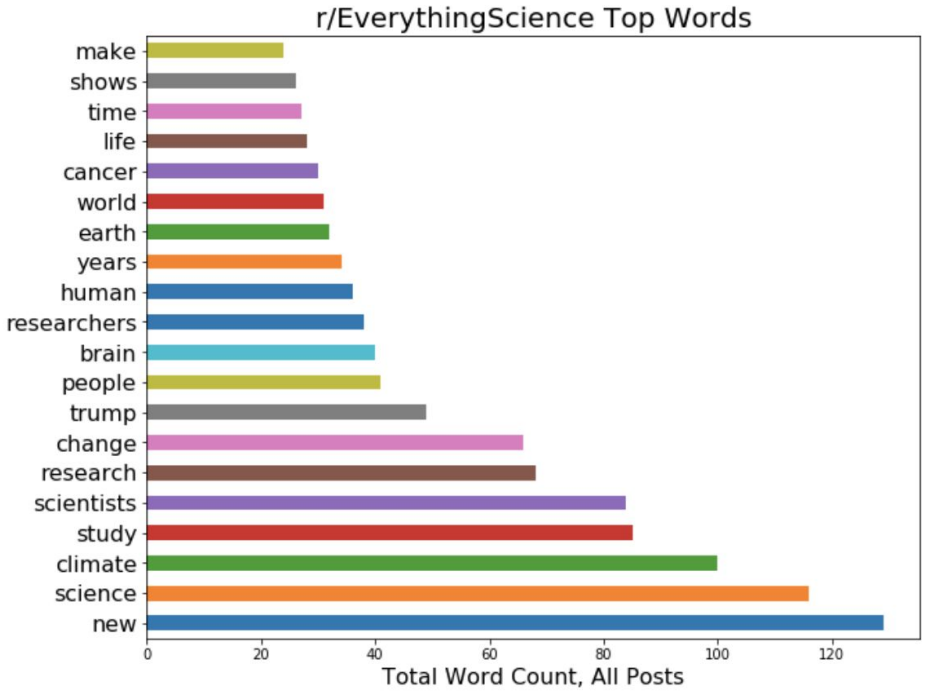
in	2.961455
science	2.413186
and	1.991791
of	1.761679
trump	1.537756
climate	1.481126
study	1.404956
finds	1.367595
by	1.367103
is	1.362622



Modeling

Best Predictors:

in	2.961455
science	2.413186
and	1.991791
of	1.761679
trump	1.537756
climate	1.481126
study	1.404956
finds	1.367595
by	1.367103
is	1.362622



Insights

Suggestions for Future Iterations:

- Text alone is not the best indicator of class
- Test the models on many kinds of subreddits, not just like ones
- Incorporate additional features into the X variable
- Format / transform documents prior to model fitting

Thank You.