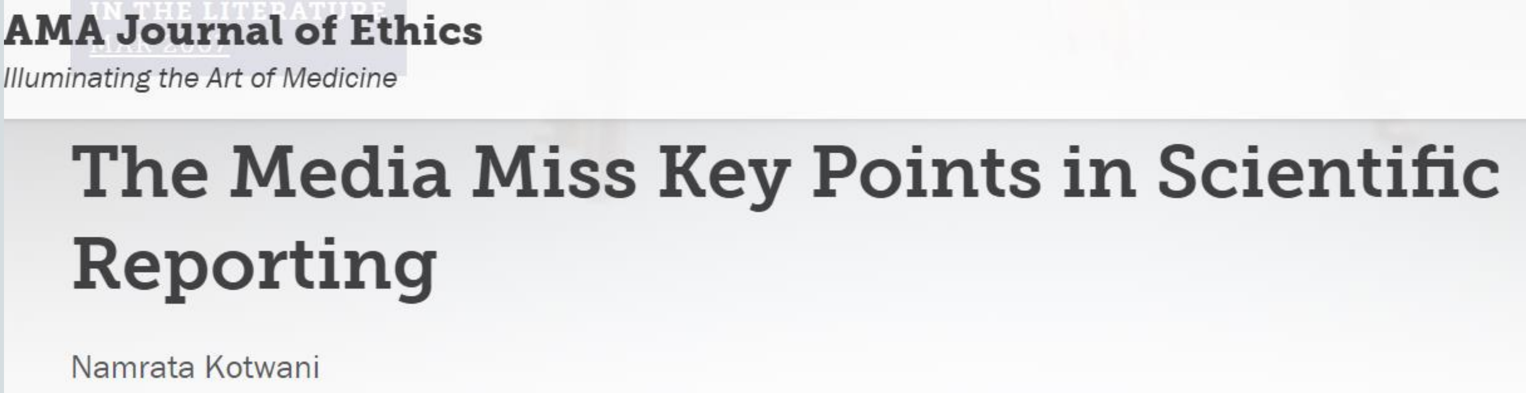
Subreddit Classification Model

Kirsty Hawke



IN THE LITERATURE

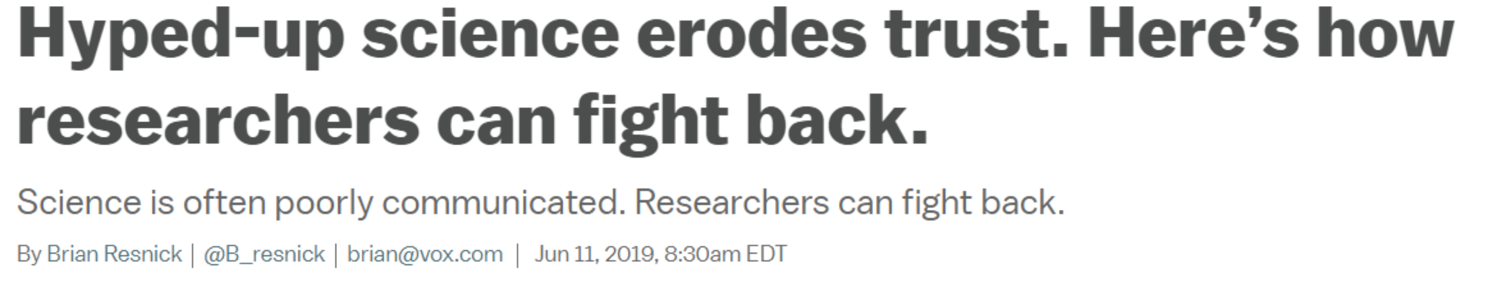
MAR 2007

AMA Journal of Ethics

Illuminating the Art of Medicine

The Media Miss Key Points in Scientific Reporting

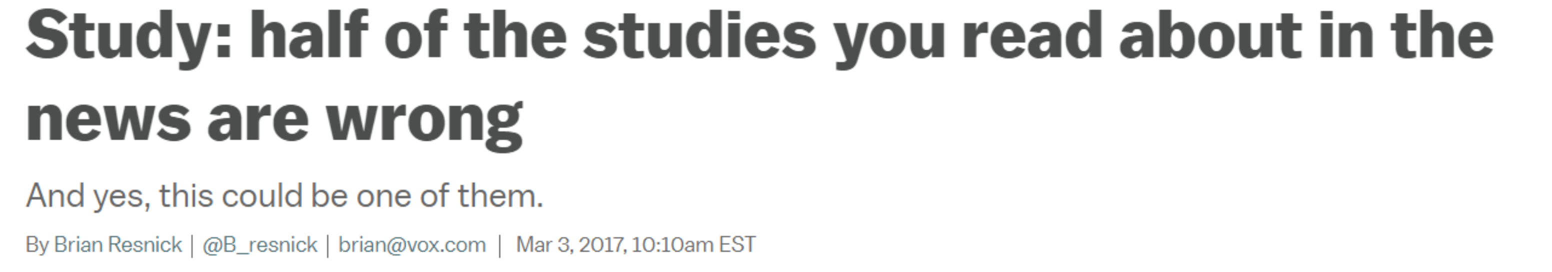
Namrata Kotwani



Hyped-up science erodes trust. Here’s how researchers can fight back.

Science is often poorly communicated. Researchers can fight back.

By Brian Resnick | @B_resnick | brian@vox.com | Jun 11, 2019, 8:30am EDT



Study: half of the studies you read about in the news are wrong

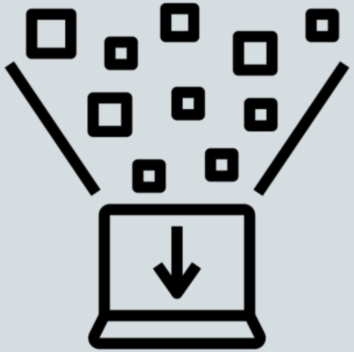
And yes, this could be one of them.

By Brian Resnick | @B_resnick | brian@vox.com | Mar 3, 2017, 10:10am EST

Problem Statement: Analyze data from r/science and r/worldnews in order to train a binary classifier based on comments, posts and a combination with a test score of over 0.85.

Aim: to illuminate the difference in general news reporting and scientific reporting. The assumption is that the language used by these communities will be different and telling of underlying values.

Data Collection



- Using pmaw - PushShift API wrapper
- Collected top 10,000 comments & submissions by score
- Binary class:
    - 0 = world news
    - 1 = science
- 50/50 split of training data

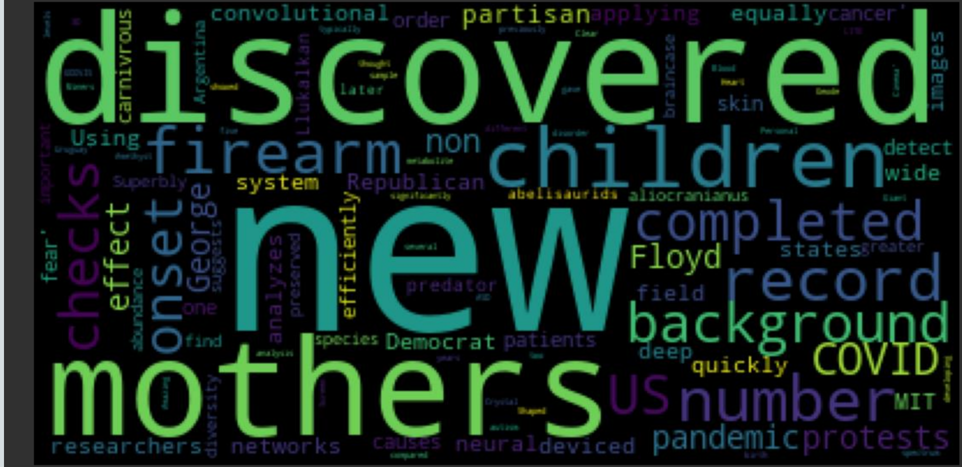
Data Cleaning

Created by Annette Spithoven
From the Noun Project

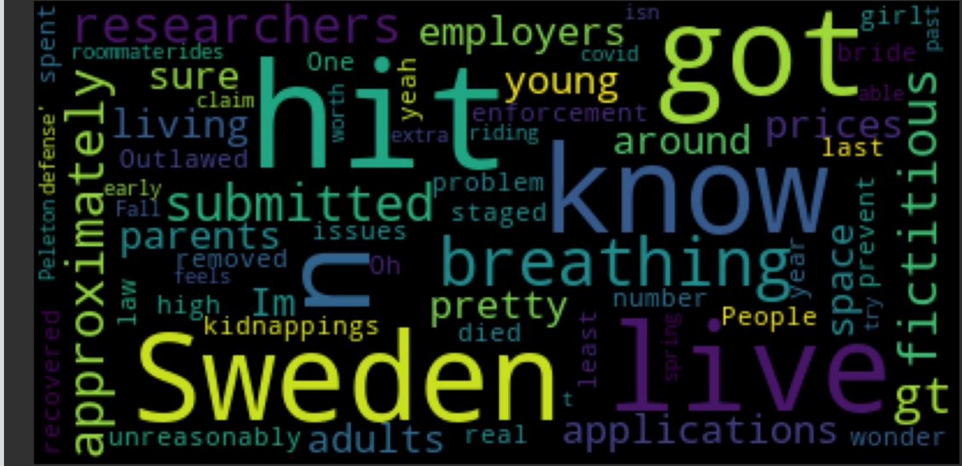
- Lowercased everything
- Using Regex Removed:
  - Emojis
  - Special characters (“”?)
  - Links
  - Html

Exploratory Data Analysis

r/science subreddit



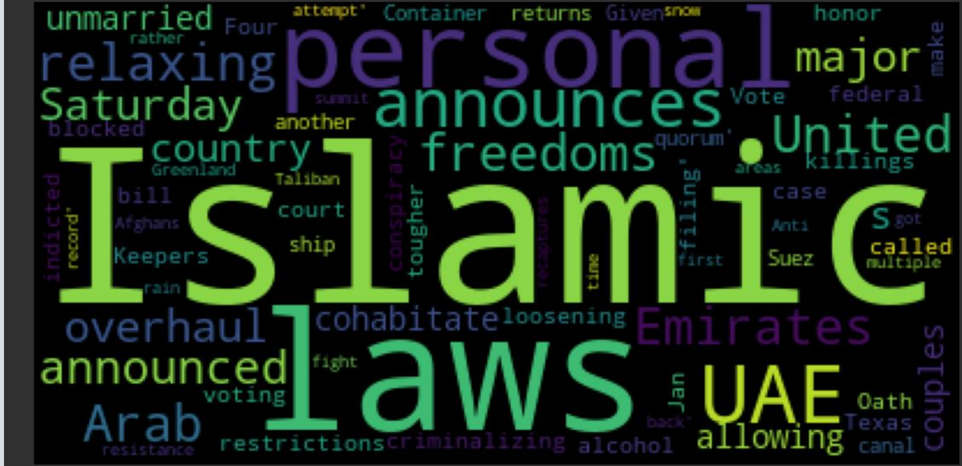
Posts



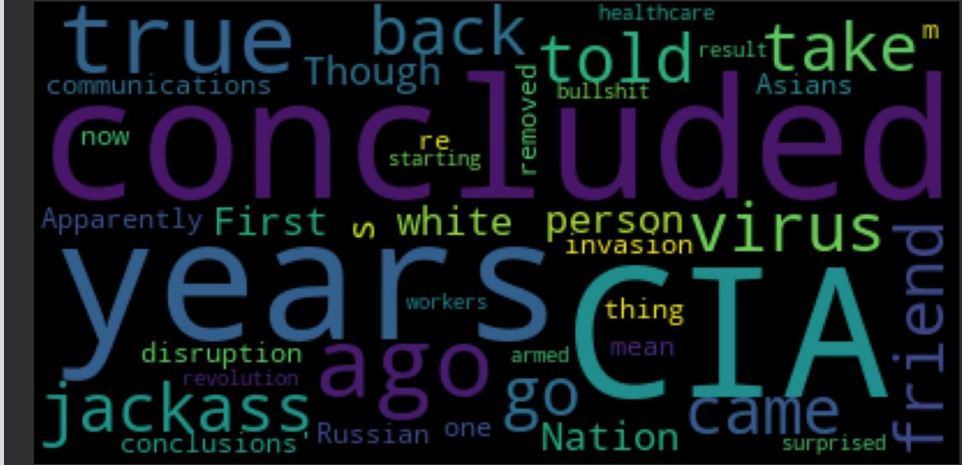
Comments

Exploratory Data Analysis

r/worldnews subreddit



Posts



Comments

Preprocessing

- CountVectorizer
- TFIDFVectorizer
- Stemming:
  - Porter
  - Snowball
  - Lancaster
- Lemmatizing

Models

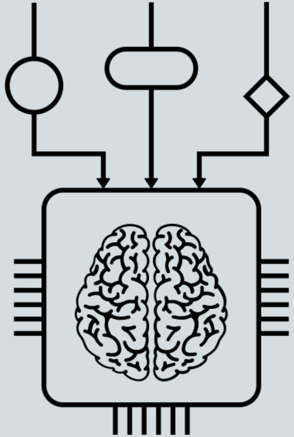
- Logistic Regression
- K Nearest Neighbours (KNN)
- Multinomial Naïve Bayes
- Random Forest
- Extra Trees
- AdaBoost
- Support-vector Classification

Model	Vectorizer	Post Score	Comments Score	Both Score
Logistic Regression	Count	0.9104	0.8071	0.8570
	TFIDF	0.8961	0.7859	0.8451
Multinomial Naïve Bayes	Count	**0.9141**	0.8177	0.8616
	TFIDF	0.9128	0.8219	0.8628
KNN	Count	0.6225	0.5048	0.5981
	TFIDF	0.7585	0.7404	0.5893
Random Forest	Count	0.8871	0.7893	0.8459
	TFIDF	0.8881	0.7845	0.8403
Extra Trees	Count	0.9027	0.8064	0.8547
	TFIDF	0.9033	0.8011	0.8562
AdaBoost	Count	0.7769	0.7041	0.7230
	TFIDF	0.7869	0.6967	0.7269
SVC	Count	0.8857	0.7419	0.8231
	TFIDF	**0.9151**	0.8104	0.8659

Max Score	0.9151	0.8219	0.8659

Table 1.0: Heatmap of R^2 Scores

Models Chosen



- Based on preliminary 5-fold cross validation scores
- Multinomial Naïve Bayes
  - CountVectorizer
  - Posts only dataset
- Support Vector Machine
  - TFIDFVectorizer
  - Posts only dataset

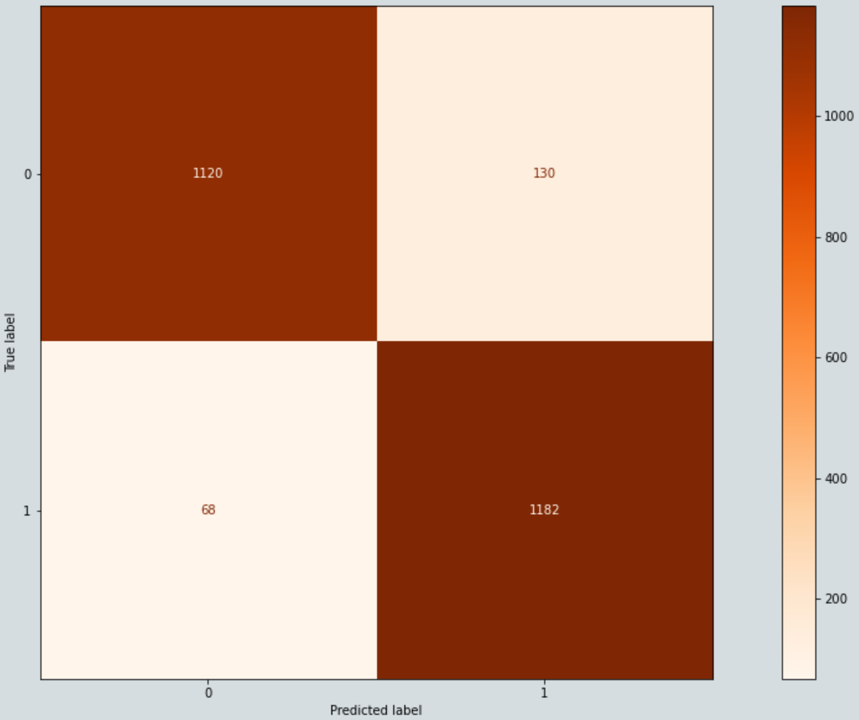
Hyperparameter Tuning – Grid Search

Parameter	Options				
Max Features	None	5000	10000	20000	
Min DF	Default	25%	50%	75%	10
Max DF	Default	25%	50%	75%	95%
Ngram	(1,1)	(1,2)	(1,3)		
Analyzer	Default ‘words’	Lemmatizer	Porter Stem	Snowball Stem	Lancaster Stem
Stop Words	None	‘english’	Nlptk stopwords		
MNB: Alpha	None	10	1	0.1	0.01
SVC: C	0.1	1	10	0.9	0.5
SVC: Gamma	1	0.1	0.01	2	5

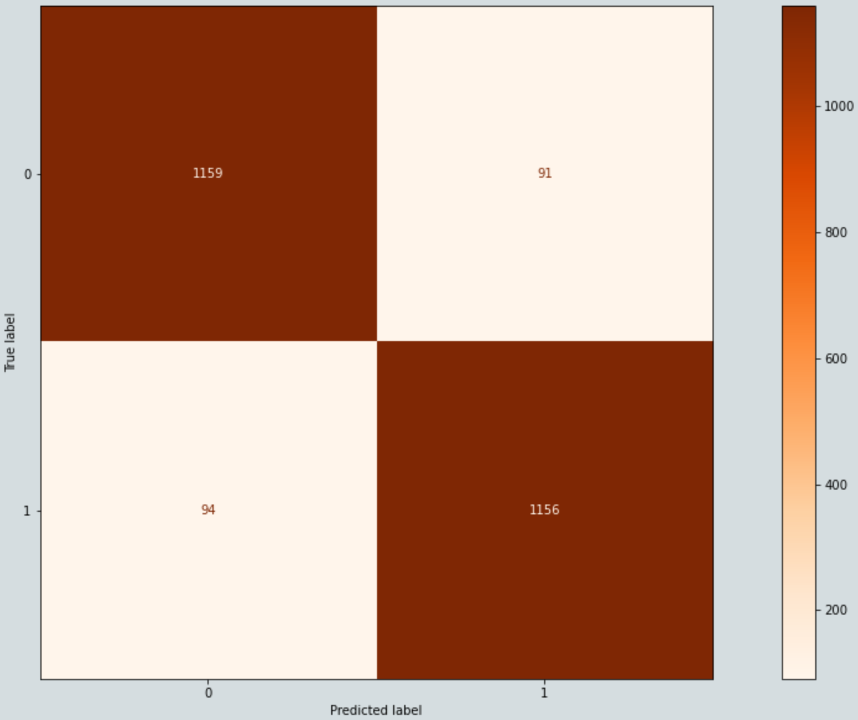
Results

Model	Vectorizor	Parameters	CV Score	Train Score	Test Score	Accuracy
MNB	Count	analyzer = lemmed, max_df: 0.95, max_features: 20000, stop_words: None, alpha: 1	0.9141	0.952	0.9208	92.1%
SVC	TFIDF	C = 0.9, gamma = 2,analyzer = stemmed_snow, max_df = 0.95, max_features = 10000, min_df = 10	0.9151	0.9898	0.926	92.6%

Evaluation Metrics



MNB Confusion Matrix



SVC Confusion Matrix

Conclusions

- SVC had the best test score and most even split of false positive and false negative cases
- It had lower bias than MNB by 0.5% but MNB had lower variance by ~2%
- Both CountVectorizor and TFIDFVectorizor worked well
- Using only the posts was the best data set
- Stop words were not helpful
- Future study
    - Using other subreddits
    - Using neural networks

Thank you!

Citations

- https://www.redditinc.com/brand
- https://www.vox.com/science-and-health/2019/6/11/18652225/hype-science-press-releases
- https://www.vox.com/science-and-health/2017/3/3/14792174/half-scientific-studies-news-are-wrong
- https://journalofethics.ama-assn.org/article/media-miss-key-points-scientific-reporting/2007-03
- The Noun Project symbols:
  - Data Collection by Kamin Ginkaew from NounProject.com
  - Cleaned Data by Annette Spithoven from NounProject.com
  - Process by Gregor Cresnar from NounProject.com
  - Machine Learning by Product Pencil from NounProject.com