Fake Article Detector

Team: Fake It Till We Make It

Michael Brashaw, Abhik Tambe, Shusruto Rishik & Sven Klumpe



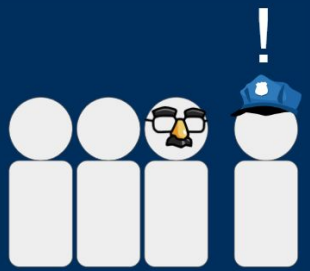

Scientific fraud - a real problem

+2 million articles are published per year

1.97% of authors admit to data fabrication (Fanelli, 2009)

Current fraud prevention methods:

1. Peer-review
2. On-site clinical trial monitoring
3. Dedicated Tweeters like Elisabeth Bik @MicrobiomDigest



Scientific fraud - a real problem

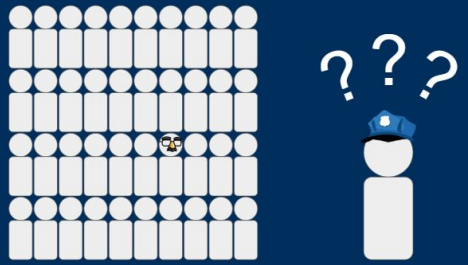
+2 million

1.97% of

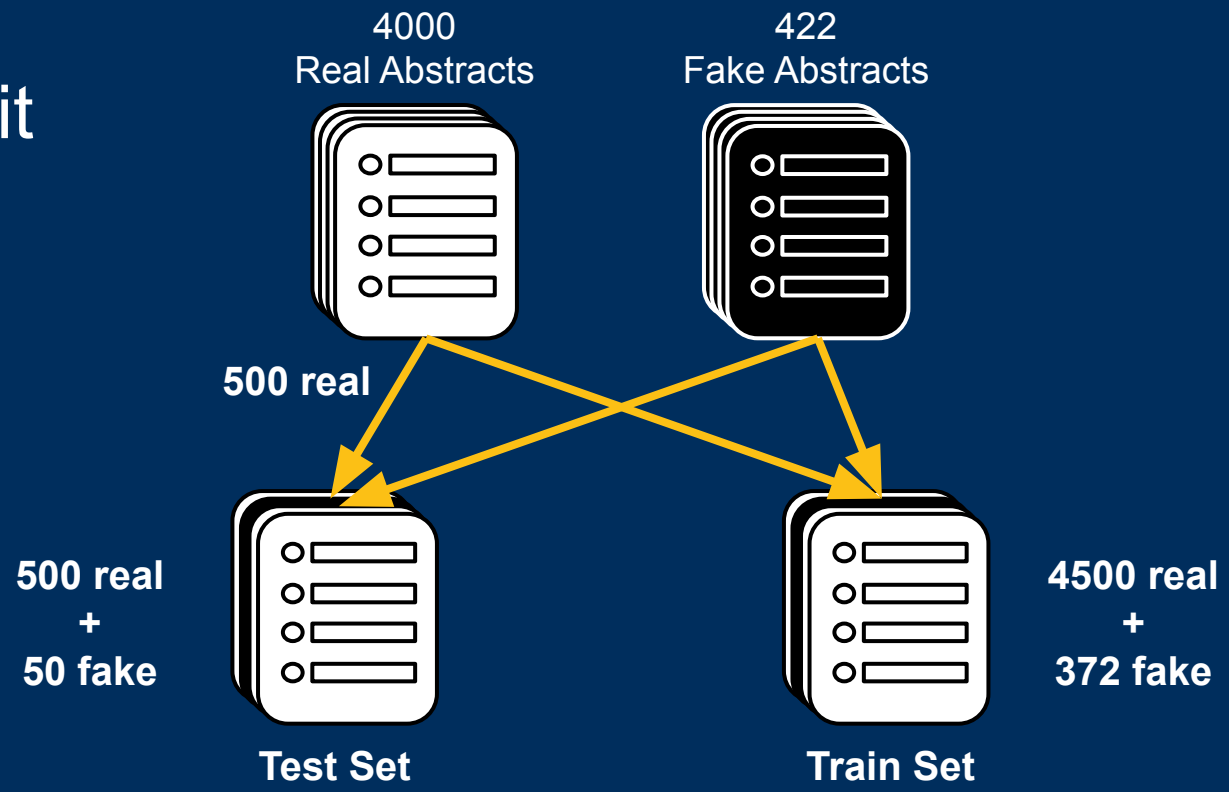
Current fr

1. Peer
2. On-s
3. Dedi

How can we systematically screen thousands of articles?



Train Test Split



n-Grams

“Quick brown fox”

1-grams:
“Quick”, “brown”, “fox”

2-grams:
“Quick brown”, “brown fox”

3-grams:
“Quick brown fox”

	“Quick brown”	“Brown fox”
Abstract 1	0	1
Abstract 1	2	3

SciKit-Learn support vector machine

⅔ train ⅓ 2-Gram validation accuracy: 0.949

Training time: 4 hours+

2-Grams - PCA

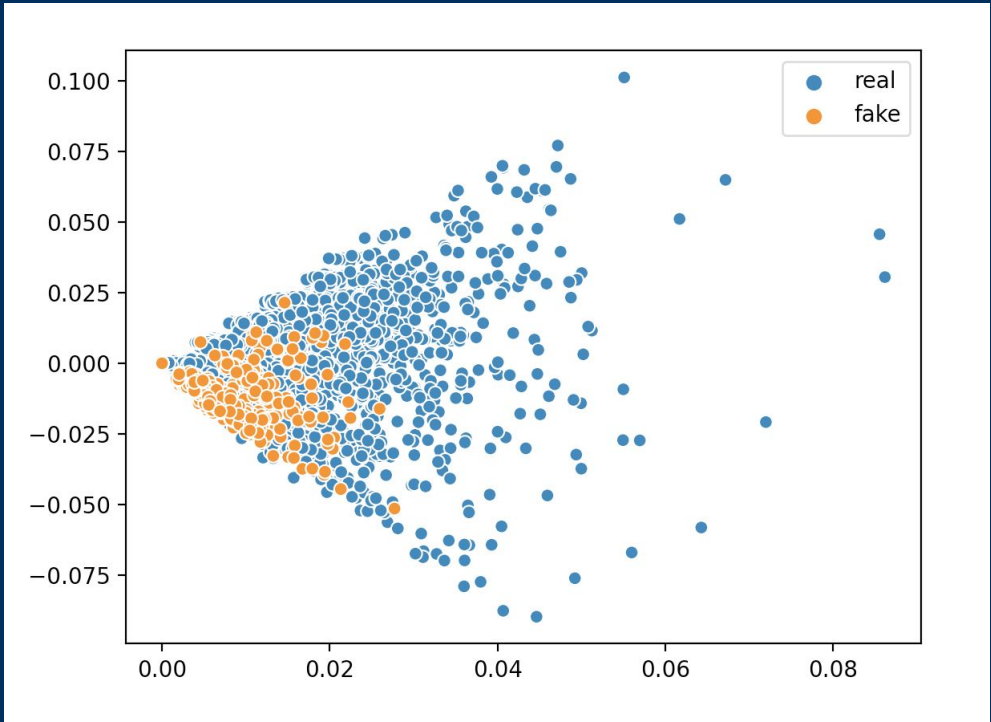
SciKit-Learn support vector machine

PCA Features:
Cross validations: 0.91, 0.91, 0.91, 0.91, 0.91

Predicted 100% of articles as real.

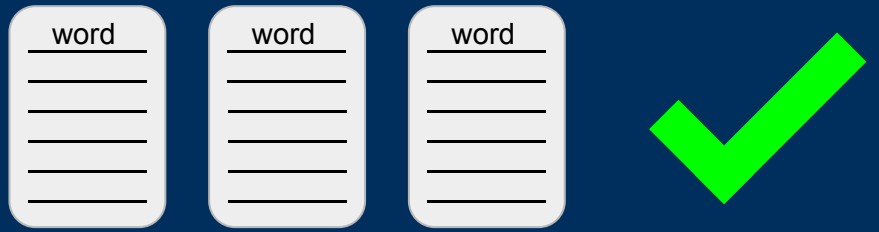
Training Data Composition:
Fake: 372
Real: 3720

90% of data is real



TF-IDF: term frequency–inverse document frequency

Terms receives for weight if it appears in many documents (implies common theme)



Terms loses weight if it appears in the same document multiple times (is an unimportant word like “the”)

TF-IDF Results

Model	Testing Accuracy	Optimized Cross-validation Accuracy	Cross-validation Accuracy	
Support Vector M.	0.988	0.977	0.978	
Random Forest	0.988	0.976	0.969	
Passive Aggressive	0.987	0.976	0.978	
Stochastic Gradient	0.988	0.976	0.976	

Predicting on 2,236 other abstracts

2,236 abstracts from “Chinese” authors in 916 journals

Predicted fake: ¾ optimized models called it fake

38

Potentially
Fake Abstracts

What Next?

Aside from TF-IDF vectors what do the 38 share with the 422?

- Images?
- Data?
- Format?
- Institution?

Sources

Github: https://github.com/MSBradshaw/BioHackathon2020

How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data; Fanelli, Daniele; PLOS ONE; May 29, 2009; https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005738

The End