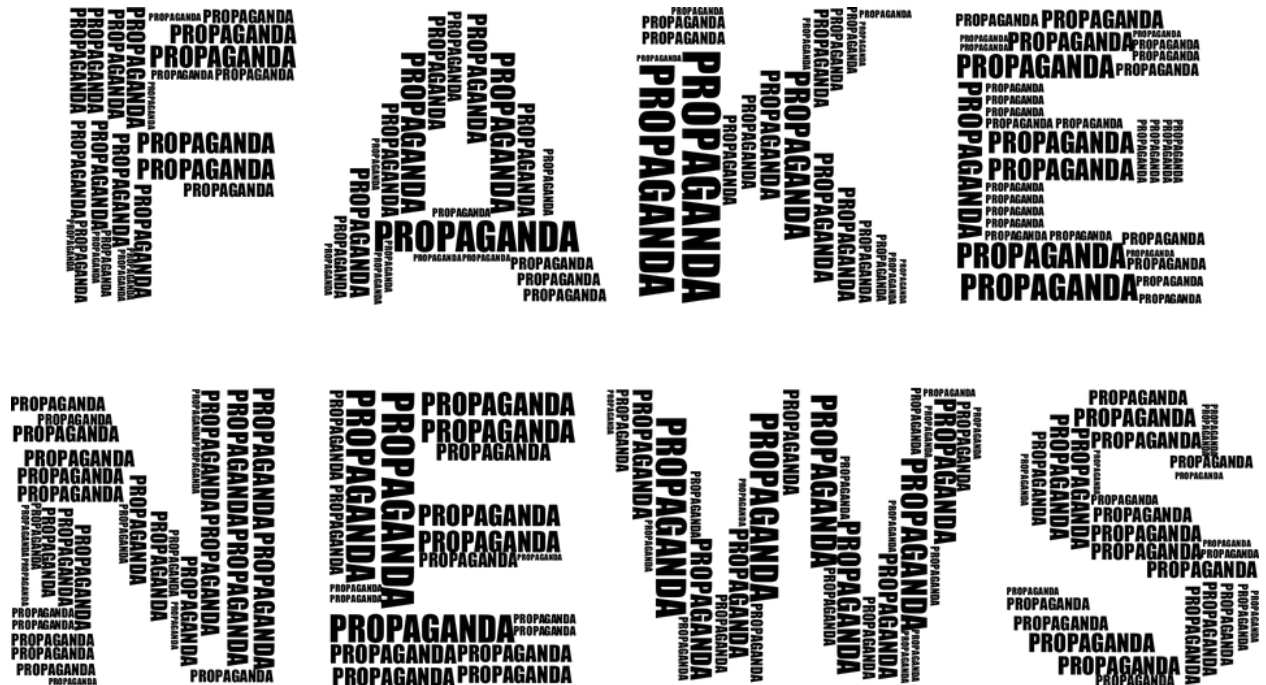


Classifying Fake News



Introduction	3
Data Mining Task	5
Technical Approach	6
Evaluation Methodology	7
Results and Discussion	8
Lessons Learned	9
Bibliography	10
Acknowledgements	10

Introduction

Modern society has been facing an issue consisting of hearsay, rumors and agenda led motivation. Fake news is the hot topic being discussed all over the media. It's not just making individuals fall prey to false assumptions but makes them less likely to believe information being spread. Just a few years ago the phrase was almost never known. Despite being relatively new, the Pew Research Center has claimed "Americans rate it as a larger problem than racism, climate change, or terrorism."

We wanted to explore the solution computer science could give to viewers worried that what they're reading falls under the category of "fake news".



Figure 1: Donald Trump is believed to have made the phrase “Fake News” popular

The questions we were curious to answer were; Could data mining/machine learning predict whether a piece of news was real or fake? Which of the classifiers that we’ve learned in class would best suit this issue? The second question was relatively interesting because as computer scientists we wanted to reaffirm the “No Free Lunch Principle” and associate the best utility under our belt with the solution.

Growing up as young Americans in a very interesting time is plenty motivation. The turmoil of mainstream media being doubted constantly is always being debated. Engineering a solution would quell our own doubts of whether or not the pieces of news can be binary in regards to real or fake.

Luckily for us we are provided a data set and data mining & machine learning knowledge from our courses taken at WSU. However every situation arises certain challenges. We were dealing with a delicate issue where certain news sources are close to home for certain individuals. Thus it was important to build a classifier that had a high accuracy rate (90%+).

Our approach was to generate a machine learning model that could report back its accuracy in predicting whether the article was real or fake. We’re happy to announce that we managed to get a classifier with our confidence measure reached.

Data Mining Task

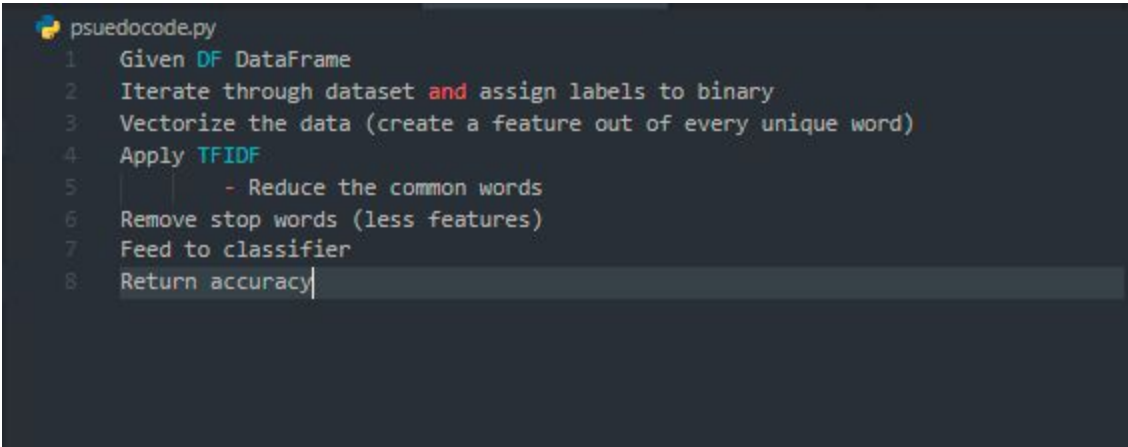
The input data is a CSV file of thousands of news articles, journals and reports where they are either fake or real. Dealing with data of “fake” could be of the following nature; propaganda, biased, pseudo science, unprofessional, satire or factually inaccurate.

8221	The Conspiracy ‘Theory’ Conspiracy [Video Documentary]	Hundreds	FAKE
7071	Clinton's Policies Look Like a Death Sentence for Americans	22 Shares	FAKE
4025	State Dept. IDs 2 Americans killed in Nepal quake; 2 others reportedly dead	The State	REAL
2067	Time to press the presidential candidates on Flint’s water crisis	In every	REAL
1371	This is why Trump was smart to avoid her: Megyn Kelly just crushed the GOP debate	Despite	REAL
10344	Life: Touching: After Her Brother Passed Away, This Woman Took Over His Facebook Page To	Email	FAKE
1926	Clinton Foundation will continue to accept foreign money during Hillary’s run	Clinton,	REAL
7657	Boy wearing a ‘My dad is an ATM’ T-shirt chased by mob; father frisked, robbed	Boy	FAKE
9800	Why Isn’t NSA Surveillance an Election Issue?	Behind	FAKE
6599	Genius Kid Trolled White House Halloween Party, Idiot Obama Didn’t Notice	Genius	FAKE
8588	Police Turn In Badges Rather Than Incite Violence Against Standing Rock Protesters	At least	FAKE
2535	The judge immigration foes wanted	One	REAL
9972	Madman Merkel Demands the Internet Publicly Release All Closed-Source Code	Madman	FAKE
2466	Here's How Obamacare Is Going To Affect Your Taxes	When	REAL
5158	George Will: Trump's judge comments prompted exit from GOP	"After	REAL
8557	Vote as if your life depended upon it, because it does.	Eric Zuess	FAKE
2282	The new argument against gay equality: Same-sex marriage kills	As the	REAL
5053	Dem convention speeches Day 4: 's Reality Check Team vets the claims	(CNN)	REAL
1515	Now Ted Cruz is the enemy: Rupert Murdoch and WSJ open fire in new GOP civil war	Social	REAL
4445	The last days of Washington, D.C.: America can no longer mask its steep decline	Since	REAL
2315	16 Times The Obama Administration Lied About The President's Position On Same-Sex Marriage	WASHIN	REAL
6454	Alternative Cancer Treatments With Positive Results and Generic-Drug Probe to Be Filed by Y	Alternati	FAKE
4213	In Bronx, Sanders voters find more common ground with Trump than Clinton	Sanders	REAL
372	The debate moderators missed the opportunity to ask about a real Democratic divide	So far,	REAL
826	Four big takeaways from Trump's 'Acela Primary' triumph	Whoever	REAL

Figure 2: Portion of data

As discussed previously our output is an accuracy rate. We want to be able to generate a classifier that is able to predict real or fake with a confidence measure of 90%. The data mining questions we want to answer is does there exist a classifier that reaches our confidence measure? The challenge is to ensure that all our previously discussed “fake news” is encompassed correctly since there are so many versions.

Technical Approach



```
psuedocode.py
1  Given DF DataFrame
2  Iterate through dataset and assign labels to binary
3  Vectorize the data (create a feature out of every unique word)
4  Apply TFIDF
5      - Reduce the common words
6  Remove stop words (less features)
7  Feed to classifier
8  Return accuracy
```

Figure 3: Pseudocode for the given program

We'll be testing all the classifiers we've learned previously in our courses until we find one that can reach our confidence measure.

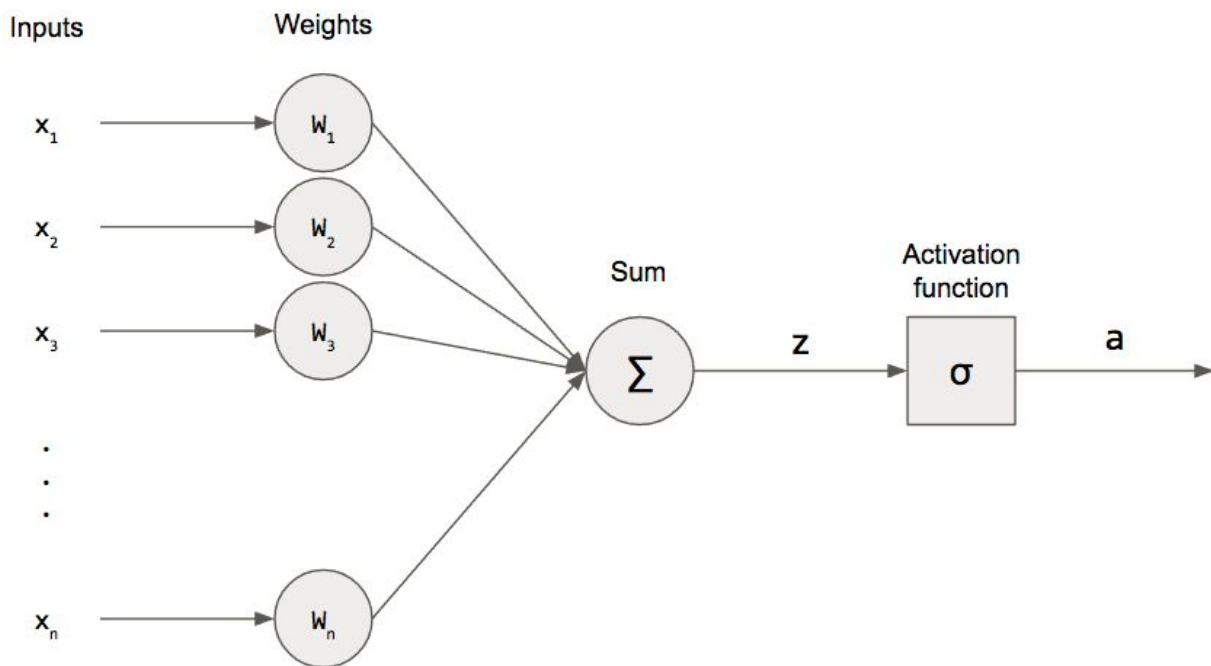
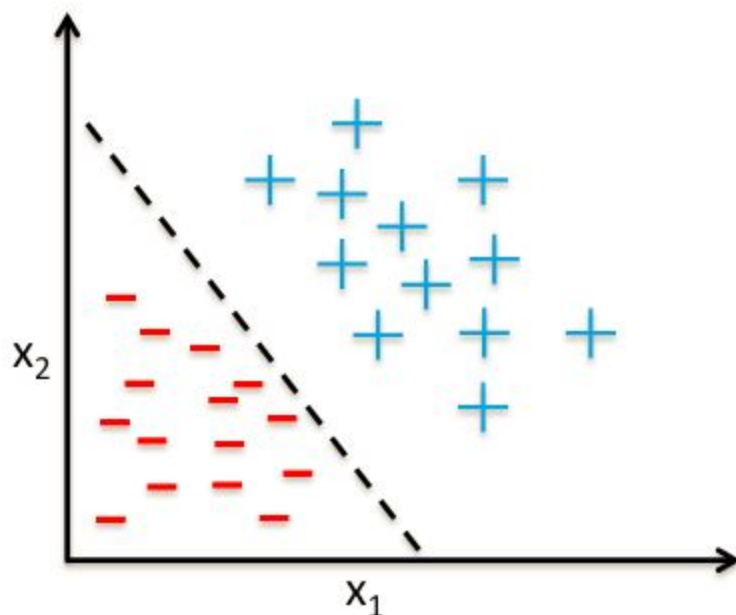


Figure 4: Classifier we chose (ended up having best accuracy)

Evaluation Methodology

Our dataset was provided by a fake news org. The goal is to use machine learning to classify news. Our only manipulation involved placing the dataset into frames and labeling them as binary.

Since confidence is the probability of the input to fall into a certain class and accuracy is the skill of the learning algorithm to predict accurately. We both understand how delicate labeling news sources as real or fake. We needed to ensure that the classifier had a high confidence rate so that without any bias it would be quite hard to doubt in regards to whether the classifier was right or wrong.



**Example of a linear decision boundary
for binary classification.**

Figure 5: Perceptron works great for linearly separable problems

Results and Discussion

We chose a perceptron as our main classifier. This model managed to reach our 90% capacity as acceptable.

Accuracy: 91.71%

Figure 6: Accuracy of the Perceptron

The outcome of the other classifiers we tried are the following:

KNN	87%
Decision Tree	77%
Perceptron	92%
Log Regression	90%

What worked best was our Perceptron classifier. This is perhaps because the problem itself was binary and linearly separable. However we're interested in seeing how the classifiers would've done given onensemble methods such as boosting or bagging. Creating a further question of which classifier can be best improved upon?

Lessons Learned

Machine learning allows a programmer to have countless methods of approaching a problem. Using the limited knowledge that we have on techniques we've learned in our courses is technically enough, but not adequate enough. Researching online there are countless innovative papers involving new methods of complex programming to solve the issue of classifying news as real or fake.

We definitely overestimated data manipulation. Using pandas isn't as simple as it seems. In order to overcome the obstacle we watched plenty of youtube videos and attended Taha's lectures on the subject. SK learn is a beast in itself. The module is almost impossible to understand if not thanks to the great work they do with documentation.

The project definitely gave us an amazing perspective into solving real world applications. We all know that industries aren't just creating redundant code. They're all very interested in being a part of the cutting edge future we all have planned.

Zeid Al-Ameedi

David Henshaw

CPTS 315 Report

Having this on our resume will definitely make us much more competitive candidates.

Bibliography

1. "Debunking False Stories Archives." *FactCheck.org*, www.factcheck.org/fake-news/.
2. "Fake News Challenge Stage 1 (FNC-I): Stance Detection." *Fake News Challenge*, www.fakenewschallenge.org/.
3. Graham, David A. "Some Real News About Fake News." *The Atlantic*, Atlantic Media Company, 12 June 2019, www.theatlantic.com/ideas/archive/2019/06/fake-news-republicans-democrats/591211/.
4. "Learn." *Scikit*, scikit-learn.org/stable/.

Acknowledgements

We'd like to thank Dr. Jana for teaching us the fundamentals of data mining. Dr. Diane Cook for giving us a strong foundation of beginner machine learning.