MACHINE LEARNING PROFESSIONALS HATE THEM!!! GENIUS PRINCETON STUDENTS DISCOVER **MIRACLE** METHOD TO IDENTIFY FAKE NEWS

The Real Deal with Fake News

By Ilene E, Ze-Xin Koh, Christine Kwon, and Eileen Wang

ABSTRACT

The widespread phenomenon of fake news contributes to misinformation, especially in politics.
Motivation: Understanding which phrases or linguistic structures best distinguish between real and fake news headlines.
Many of the individual phrases found in real news are found in fake news as well. The reason we, as human users are capable of distinguishing between the two is because we are able to draw links between the different words in a statement to determine its plausibility.
Beyond observing word frequencies, a successful model would have to detect semantic difference between fake and real news.

RELATED WORK

Many fake news detection algorithms that have been developed take into account factors besides the actual text, such as the website it's from, and statistics such as the like, comment, or repost counts.
Some even check how long the website has been around for, as most fake news websites are young.

CLASSIFICATION METHODS

We tested classifiers using the following algorithms:

i. Bernoulli Naive Bayes (NB)
ii. Linear Support Vector Machine (SVM)
iii. Random Forest (RF)
iv. Multi-Layer Perceptron (MLP)
v. Decision Trees (DT)

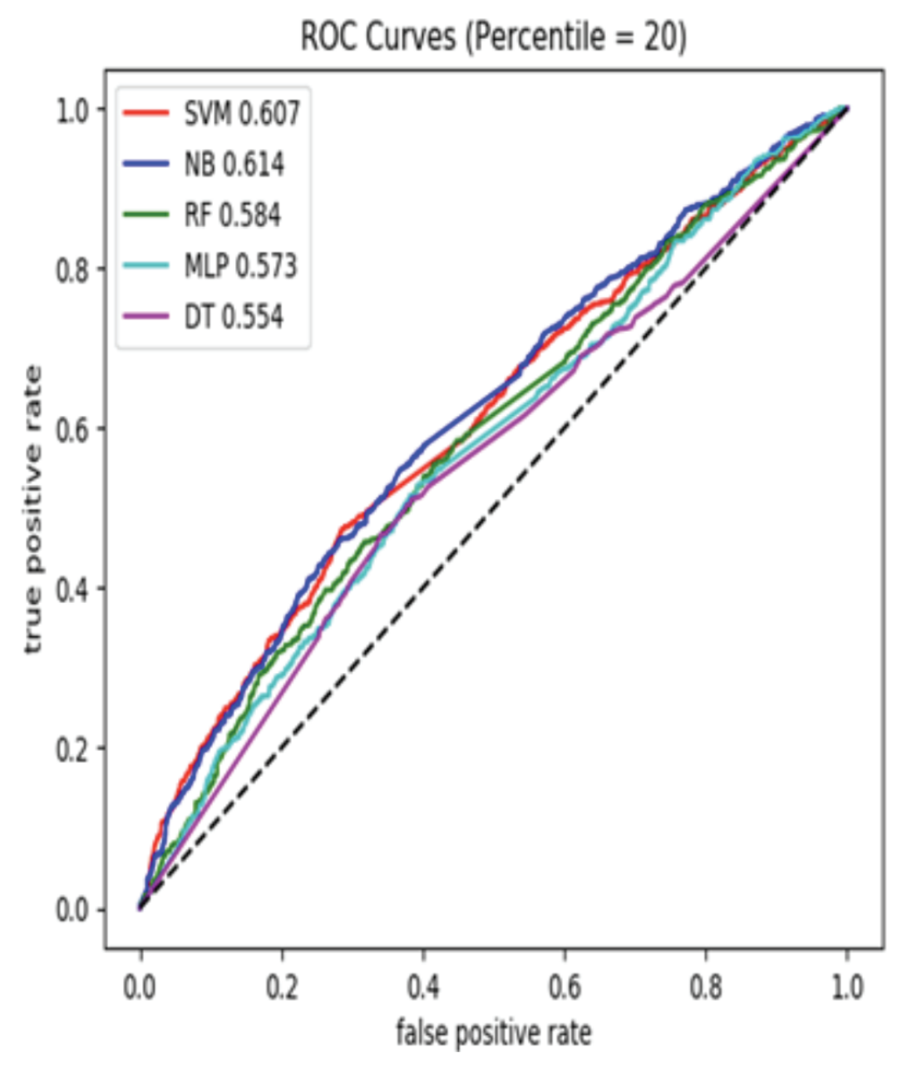
THE DATASET

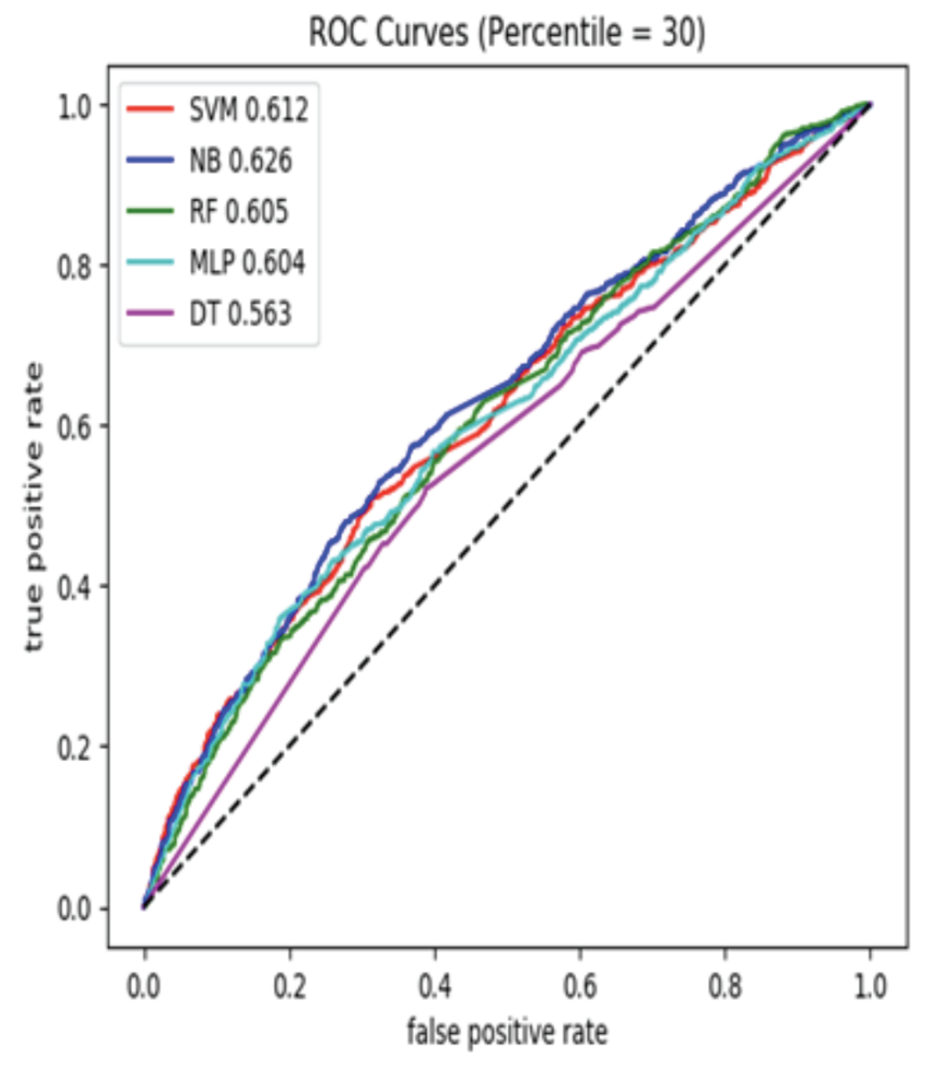
i. The dataset we used was obtained from Kaggle [1], consisting of 10223 news statements scraped from PolitiFact.com [2].
ii. 56% of the statements in the dataset are real and the rest are false.
iii. We used a 80/20 ratio to split the dataset into a training and a testing set.
iv. Each sentence was tokenized, stemmed, and lemmatized to produce a bag-of-words representation featuring single words, bigrams, and trigrams.
v. The vocabulary produced consisted of 12522 unigrams, bigrams and trigrams.

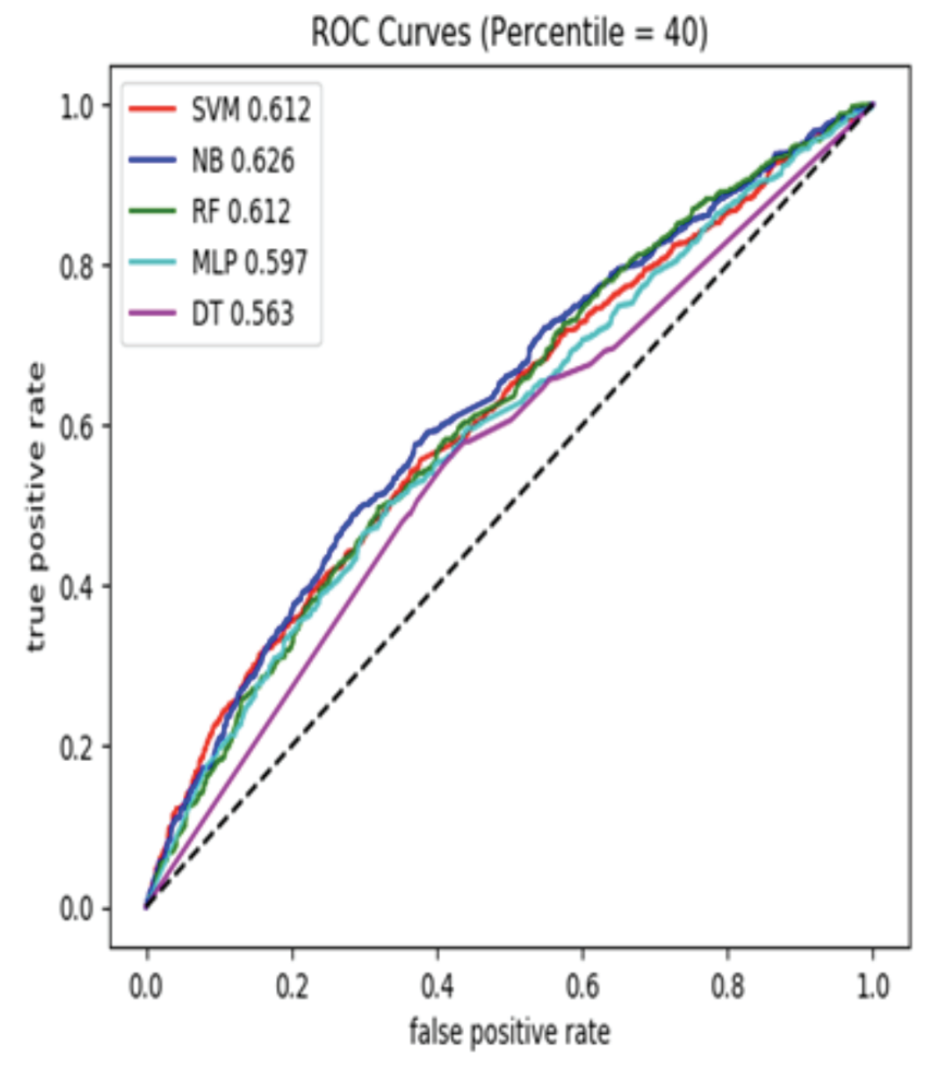
RESULTS

	Precision	Recall	Specificity
Bernoulli NB	0.615	0.716	0.431
Linear SVM	0.601	0.817	0.312
Random Forests	0.593	0.593	0.408
Multi-Layer Perceptron	0.588	0.588	0.431
Decision Tree	0.589	0.616	0.456

ROC





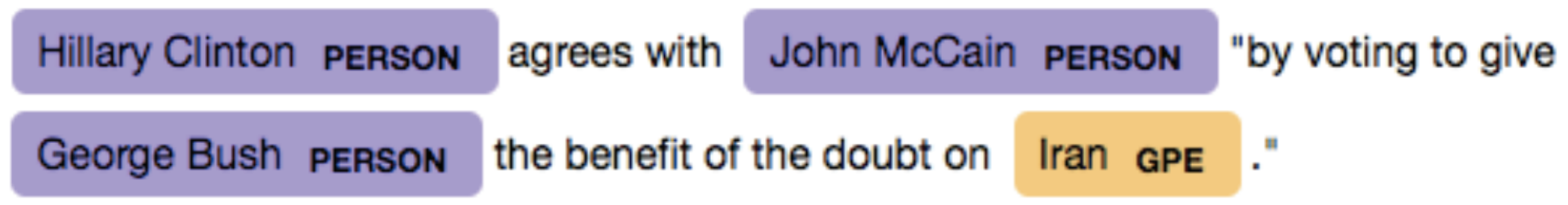


LDA

	topic 0	topic 1	topic 2	topic 3	topic 4
fake	obama says president barack 000 percent year years obamacare security	people health percent says pay care voted tax taxes like	says illegal dollars tax million did city america county bush	says state wisconsin care scott health jobs 000 clinton billion	says states texas united school dont rick know public state
real	000 jobs state new year million public office oil clinton	percent states tax people taxes united rate says years 10	obama says president said barack city 30 country number illegal	health year care billion budget spending 000 tax federal plan	says texas state voted republican times day senate romney trump

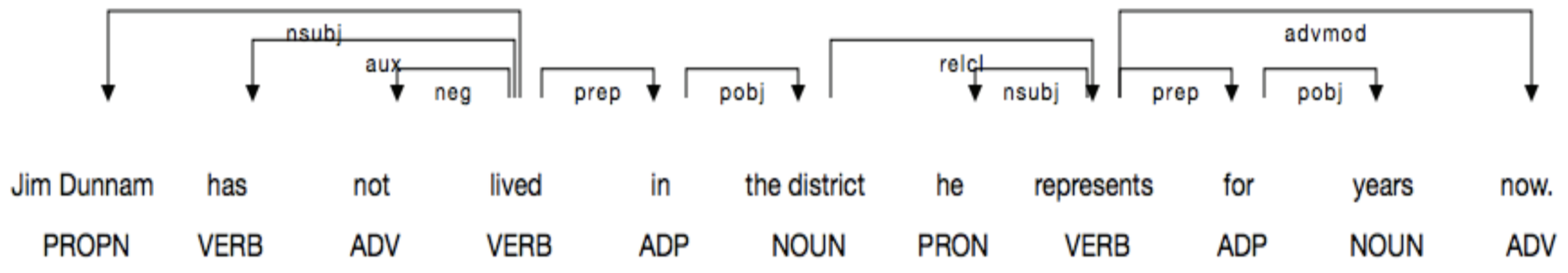
PARTS-OF-SPEECH TAGGING (POS)

Words in news headlines are annotated using Parts-of-Speech (POS) tagging, which identifies linguistic components such as nouns, verbs, and adjectives. Named entities are also identified.



Dependency labels are used to model semantic relationships between words in the sentence.

A neural network model is used in the classification.



REFERENCES

[1] Patro, S. (2019, January 21). Fake News Detection Dataset. Retrieved from https://www.kaggle.com/ksaivenketpatro/-fake-news-detection-dataset/activity
[2] Fact-checking U.S. politics. (n.d.). Retrieved from https://www.politifact.com/
[3] SpaCy · Industrial-strength Natural Language Processing in Python. (n.d.). Retrieved from https://spacy.io/