NLP ALGORITHMS FOR FAKE NEWS DETECTION

Angad, Danny, Emma, Noah, Prisha
June 30, 2023

BACKGROUND

The Problem:

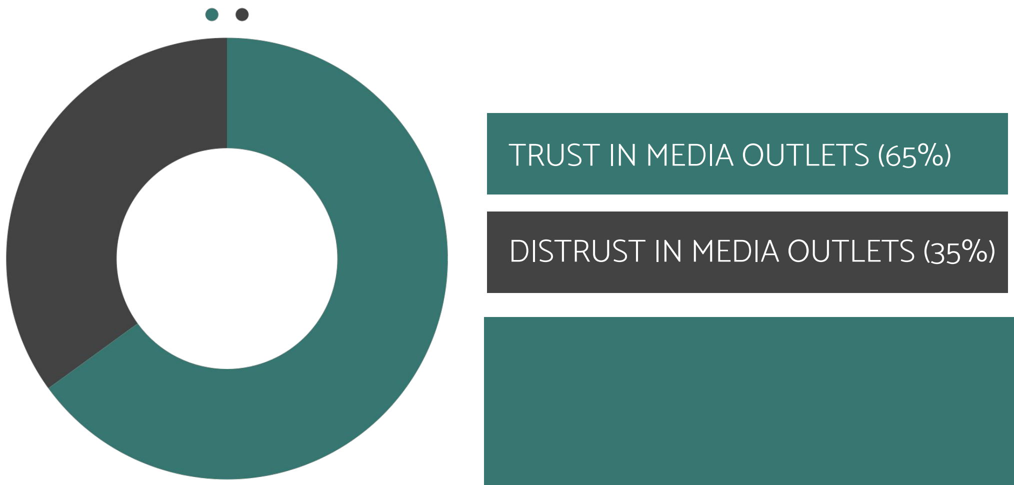
- Fake news has detrimentally influenced society and manipulated people's opinions as well as perceptions on a distinct topic, leading to bias

Impact:

- Amplified the spread of misinformation, disinformation, and malformation
- Results in making truth difficult to find because of biased sources

Effect in Numbers:

- 80% of adults in the United States have consumed fake news (Statista)
- 67% of adults in the United States have read false information on social media outlets (Statista)
- 64% of adults in the United States believe fake news causes considerable confusion about current issues (PEW Research)
- 38.2% of Americans have accidentally shared fake news (Techjury)

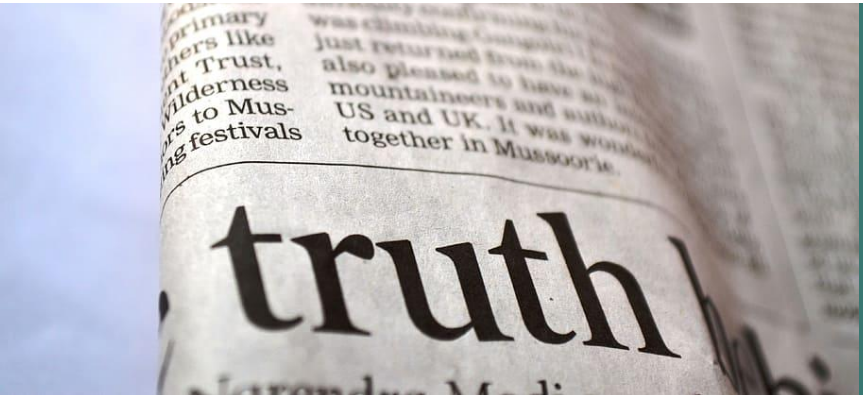


DETECTING FAKE NEWS WITH ML + NLP

The Charlotte Observer
How McCaffrey handles the NFL's biggest workload
A widow shares her experience after husband's sudden death

ML

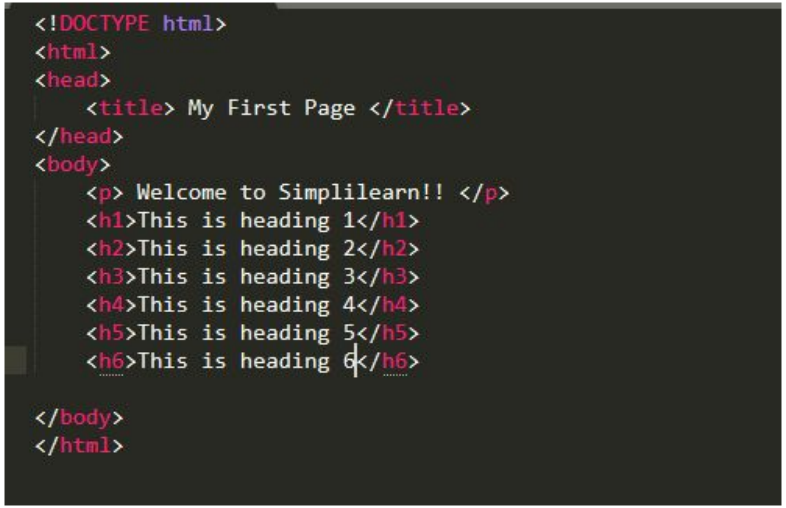
With ML, our fake news detection model has the opportunity to automate processes and deliver accurate responses.



NLP

The application of NLP algorithms in text classification has proffered the opportunities to determine if the news is real or fake.

Data

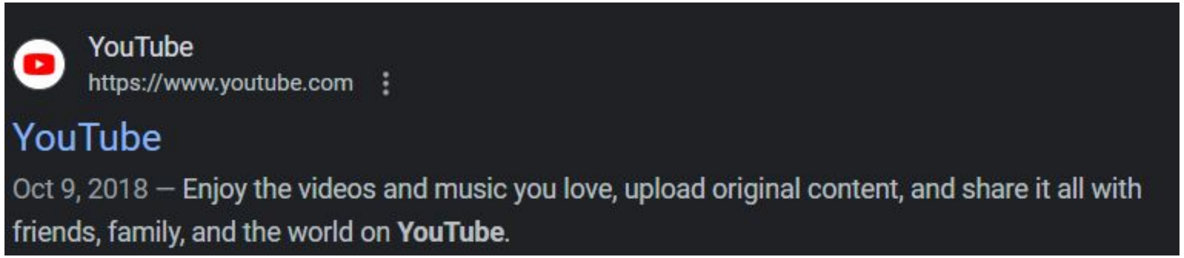
- Collection of news websites from all over the Internet
- Features
  - URL: the link to the website ex. ‘https://google.com’
  - HTML: content of the website
- Label: 0 or 1 for real or fake
- HTML is form of content
  - 

```
<!DOCTYPE html>
<html>
<head>
    <title> My First Page </title>
</head>
<body>
    <p> Welcome to Simplilearn!! </p>
    <h1>This is heading 1</h1>
    <h2>This is heading 2</h2>
    <h3>This is heading 3</h3>
    <h4>This is heading 4</h4>
    <h5>This is heading 5</h5>
    <h6>This is heading 6</h6>

</body>
</html>
```

HTML

- Preprocessing: convert words to lowercase
- Beautiful Soup to parse HTML content for ‘meta’
  - Gets metadata
- Metadata is like this:



YouTube
https://www.youtube.com

YouTube

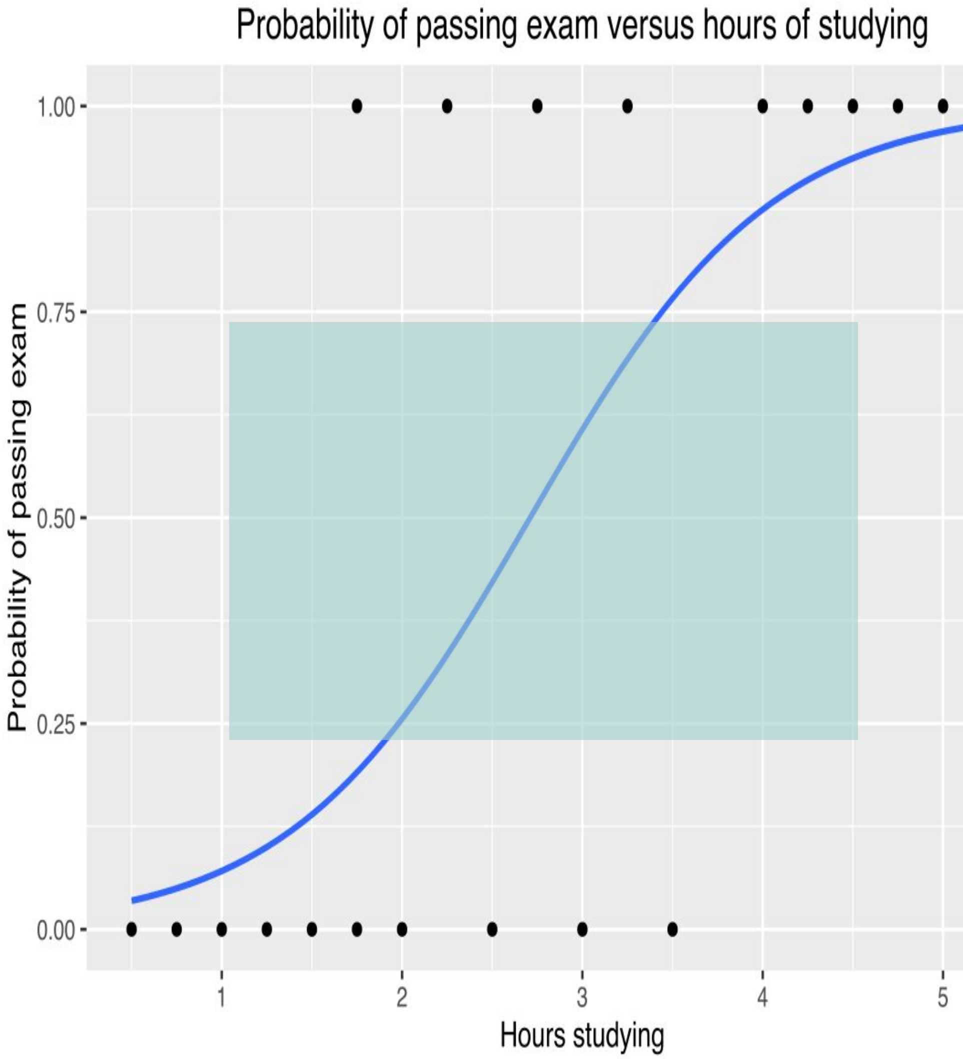
Oct 9, 2018 — Enjoy the videos and music you love, upload original content, and share it all with friends, family, and the world on **YouTube**.

Logistic Regression Model

- Logistic regression is a classification model:
- **Returns probability rather than numerical values**

- (for basic models) *Line of best fit* most commonly looks like an s-shaped curve
-
- After training, new data can be inputted and model can predict based on X value and plotted curve.

- We used the URL, metadata (using NLP), and different keywords, as parameters, like **hours studying** on the right
-
- Each of these was assigned a weight during training, the value of which is used to compute the final probability.



Training the Model

BOW

Val Accuracy: 66.3%

Train Accuracy: 87.5%

extract counts from the description for particular keywords and use these as features automatically

Glove

Val Accuracy: 77%

Train Accuracy: 86.5%

Through this model words are put into word vectors to identify similar words

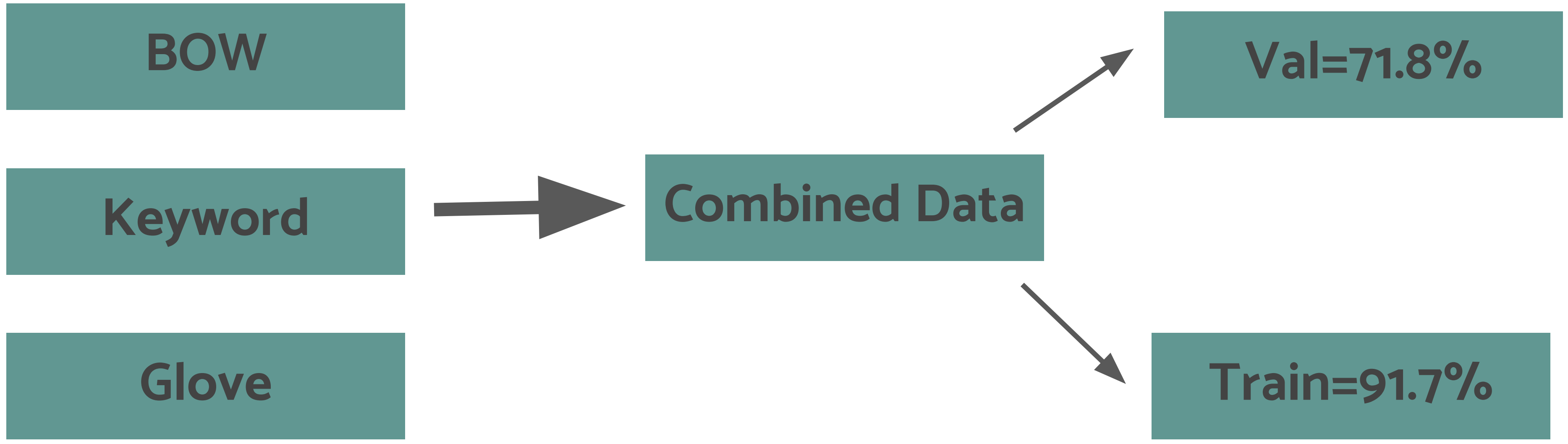
Keyword

Val Accuracy: 73.5%

Train Accuracy: 79.2%

A very similar process to BOW, but requires manual input of keywords. It is not an Automated process

Results



Examples of Results & Confusion Matrix

Actual \ Predicted		
	TN	FP
	FN	TP

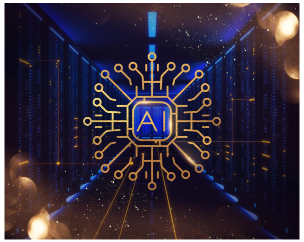
TN: True Negative
TP: True Positive
FP: False Positive
FN: False Negative

↳ MOST DAMAGING

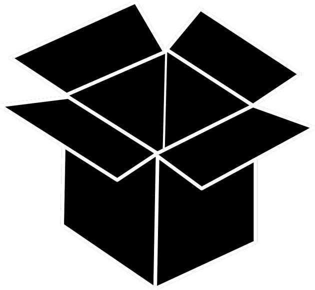
www.pintrest.com → **Real**

www.amazon.com → **Real**

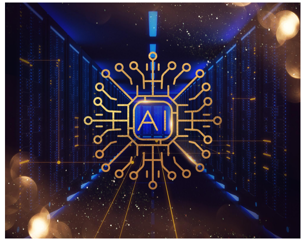
FAKE NEWS


AI

≠



FAKE NEWS


AI

=

Thank You!

Does anyone have any questions?

The Tribune
Partition Of Assets And Liabilities
Complete Agreement Reached Between India & Pakistan
Kashmir Issue Not Discussed
GANDHIJI'S TALKS WITH THE VICEROY
The Hindustan Times
SPEEDY EVACUATION BOTH WAYS ESSENTIAL
PATEL ON GRAVITY OF PROBLEM
UNDISTURBED PASSAGE TO BE ENSURED
PUNJAB STATES OFFER TO ABSORB REFUGEES
RESTORE REAL PEACE IN DELHI
RAPID RETURN OF PEACE IN DELHI
SARDAR PATEL VISITS