News Headline Classification

Presented by: Madoria Thomas

Topics We'll Cover

Outline

Overview
Business Problem
Data
Methods
Results
Recommendations
Next Steps
Thank You

Overview

There is much speculation as to when exactly newspapers became a source of news, but there is evidence that proves they were circulated as early as 202 BC in ancient China as “dibaos” which were basically like bulletins or posters. It would take until about 1440 when Johannes Gutenberg invented the moving printing press that the first newspapers as we know them would appear in Europe where they could more quickly be mass produced. From there, the first true newspaper in America would eventually pop up in 1690 in Boston. However the publisher, Benjamin Harris, was arrested for including political criticisms and his newspaper was suppressed and all copies were destroyed. Of course now in the digital age, the need for paper newspapers has waned in favor of online news sources. In addition, certain niches and industries might prefer being online rather than in print. There are more news sources online than ever which is where I step in.

Source: https://www.psprint.com/resources/newspaper-history/

Business Problem

I have been tasked by the New York Times (NYT) to provide a quick and easily accessible news solution for users who may already be overwhelmed. My goal is to build a news classifier to accurately predict what the topic is from the headline

The dataset has been provided by the News API team and can be found on their GitHub. There are:

- Total Rows: 108774
- Target: 'topic'
- Predictor: 'title'
- 8 categories: TECHNOLOGY, HEALTH, WORLD, ENTERTAINMENT, SPORTS, BUSINESS, NATION, and SCIENCE.
- August 2020

DATA

</newscatcher>

METHODS

LOWERCASED

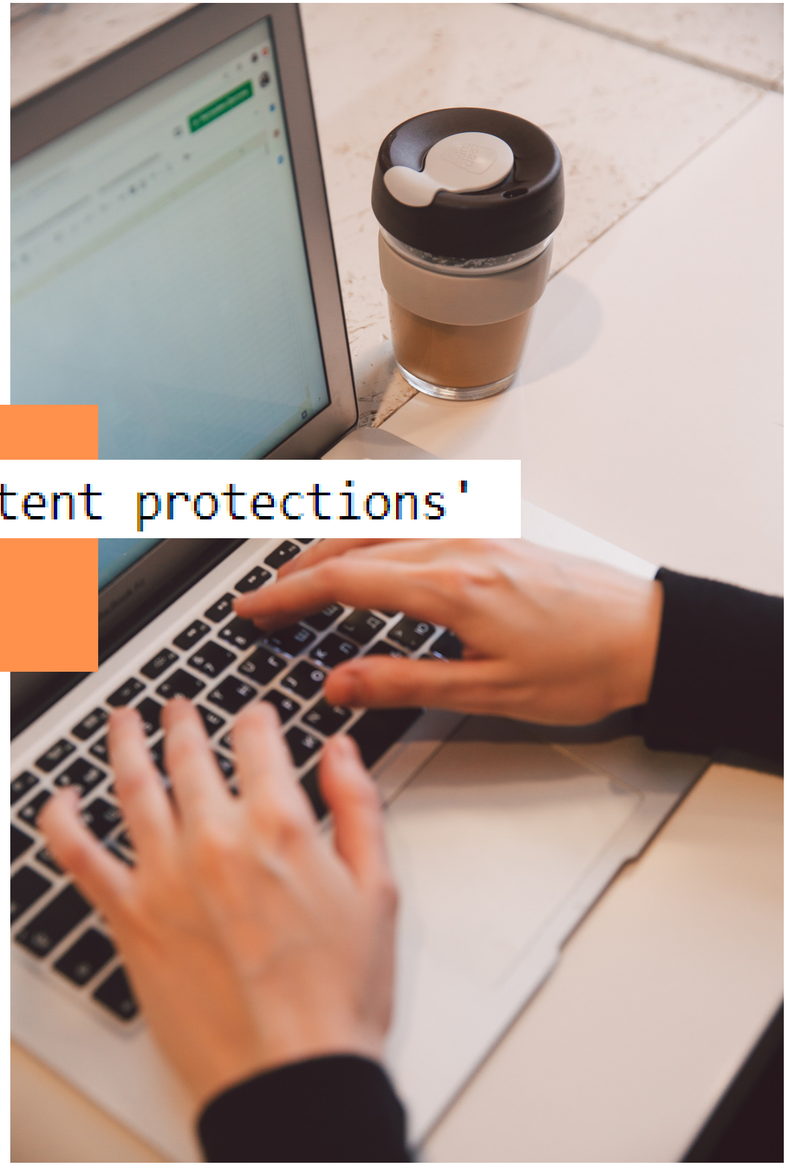
TOKENIZED

LEMMATIZED

REMOVED STOP WORDS

VECTORIZED

'Moderna eases up on vaccine patent protections'



METHODS

LOWERCASED ✓

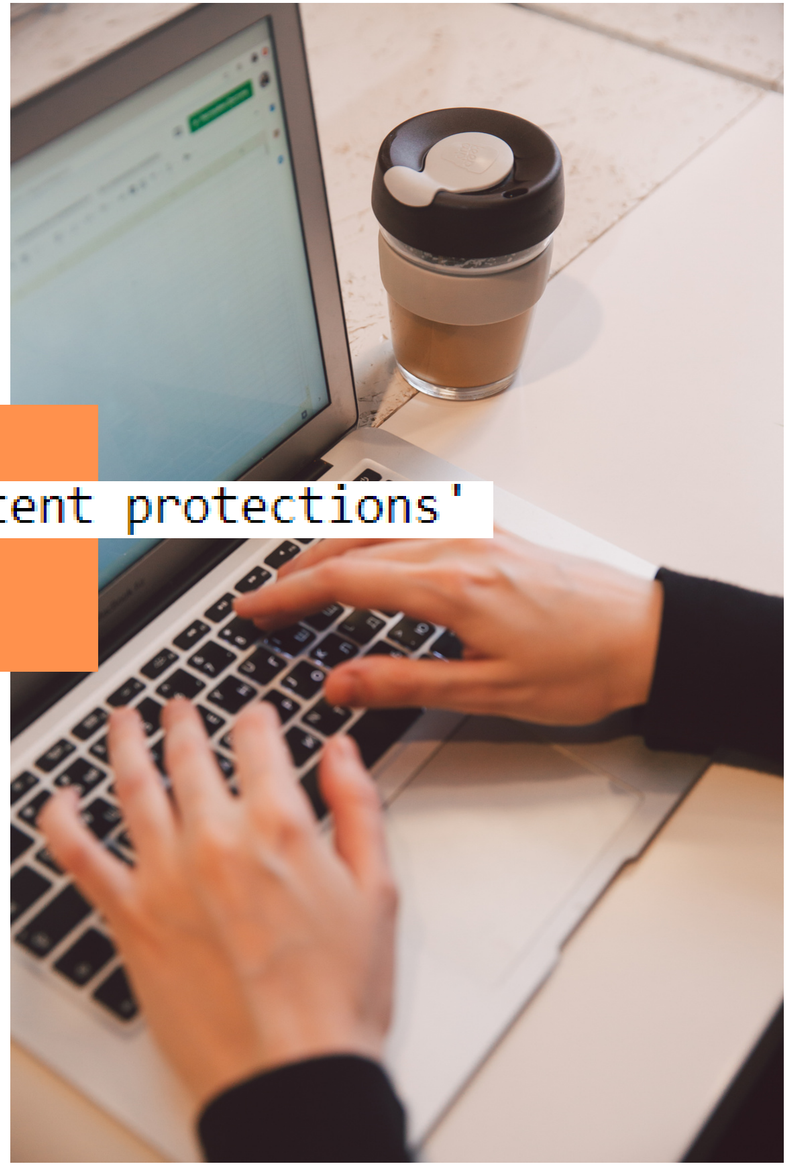
TOKENIZED

LEMMATIZED

REMOVED STOP WORDS

VECTORIZED

```
'moderna eases up on vaccine patent protections'
```



METHODS

LOWERCASED ✓

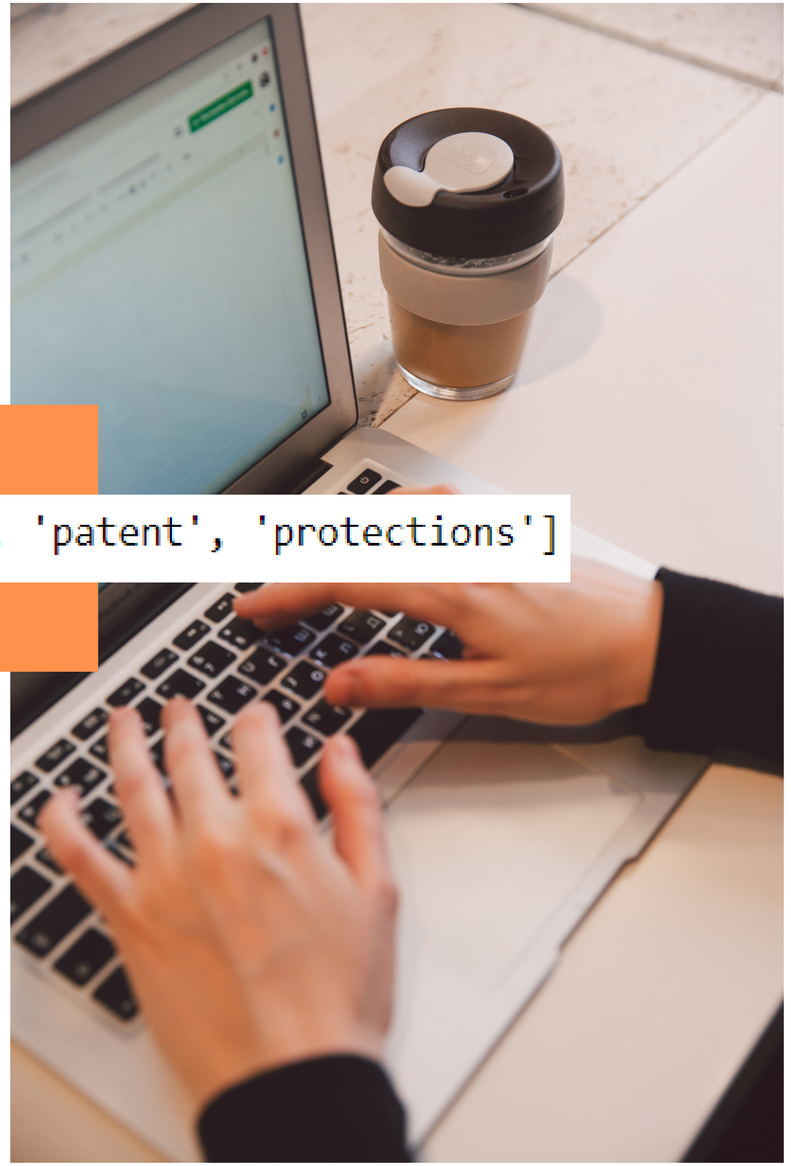
TOKENIZED ✓

LEMMATIZED

['moderna', 'eases', 'up', 'on', 'vaccine', 'patent', 'protections']

REMOVED STOP WORDS

VECTORIZED



METHODS

LOWERCASED ✓

TOKENIZED ✓

LEMMATIZED ✓

```
['moderna', 'eas', 'up', 'on', 'vaccine', 'patent', 'protection']
```

REMOVED STOP WORDS

VECTORIZED

METHODS

LOWERCASED ✓

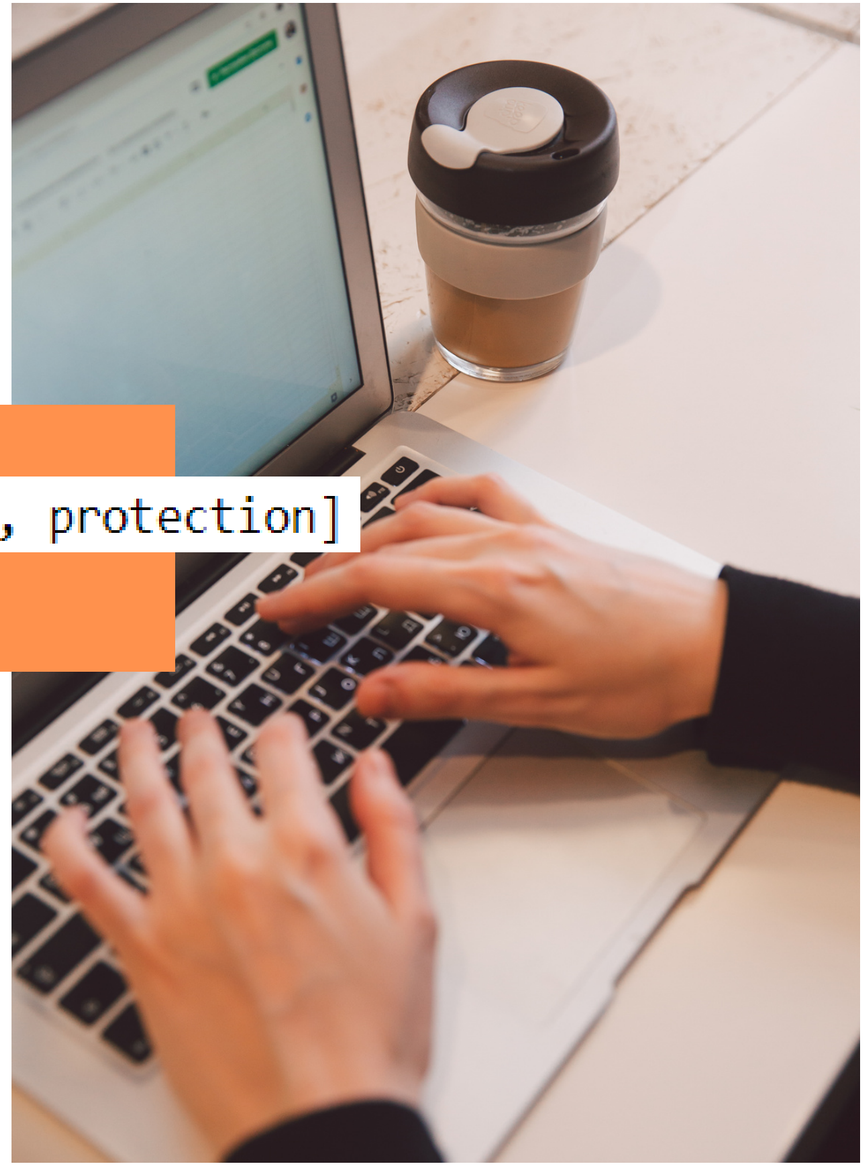
TOKENIZED ✓

LEMMATIZED ✓

REMOVED STOP WORDS ✓

VECTORIZED

[moderna, eas, vaccine, patent, protection]



METHODS

LOWERCASED ✓

TOKENIZED ✓

LEMMATIZED ✓

REMOVED STOP WORDS ✓

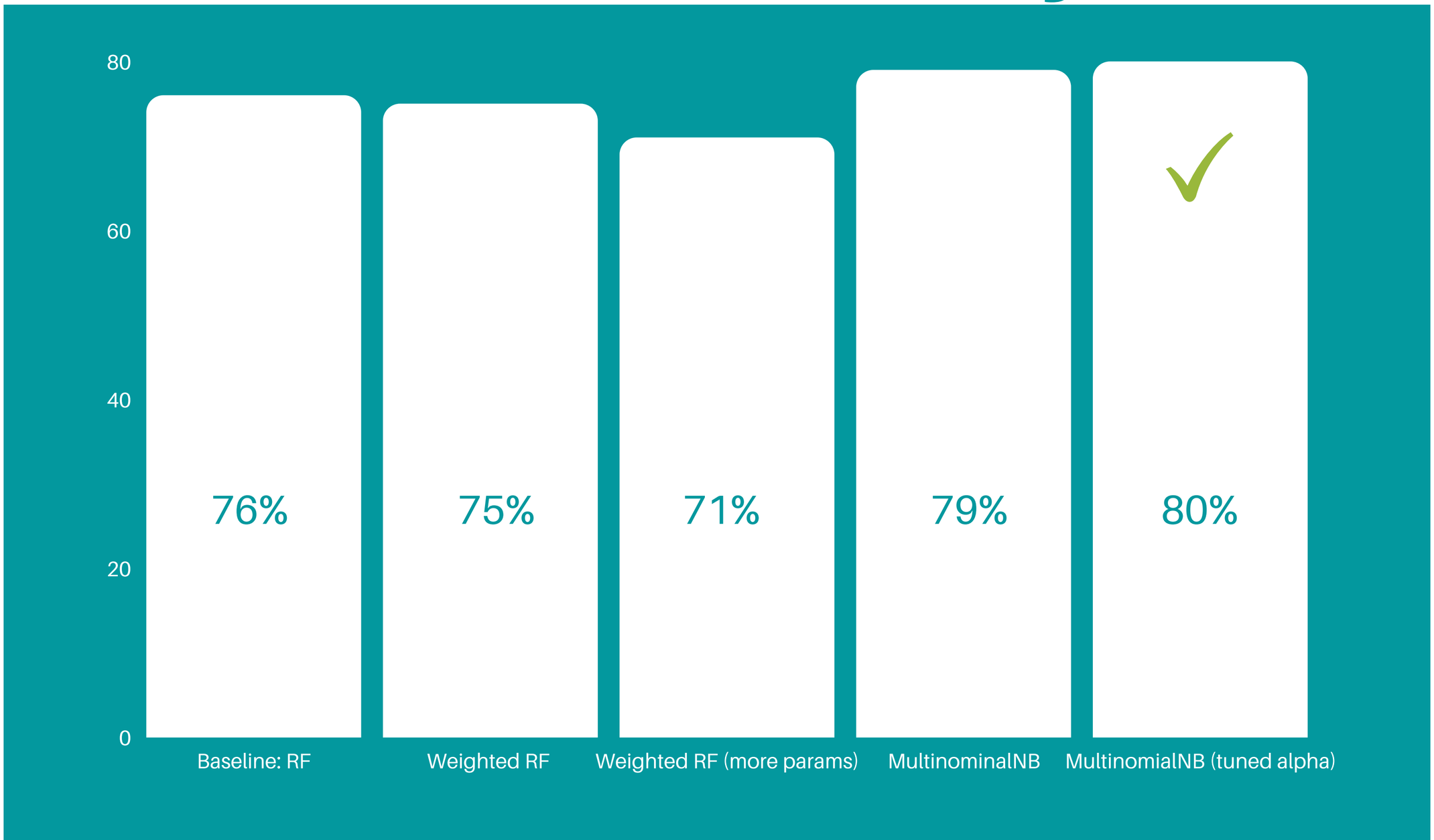
VECTORIZED ✓

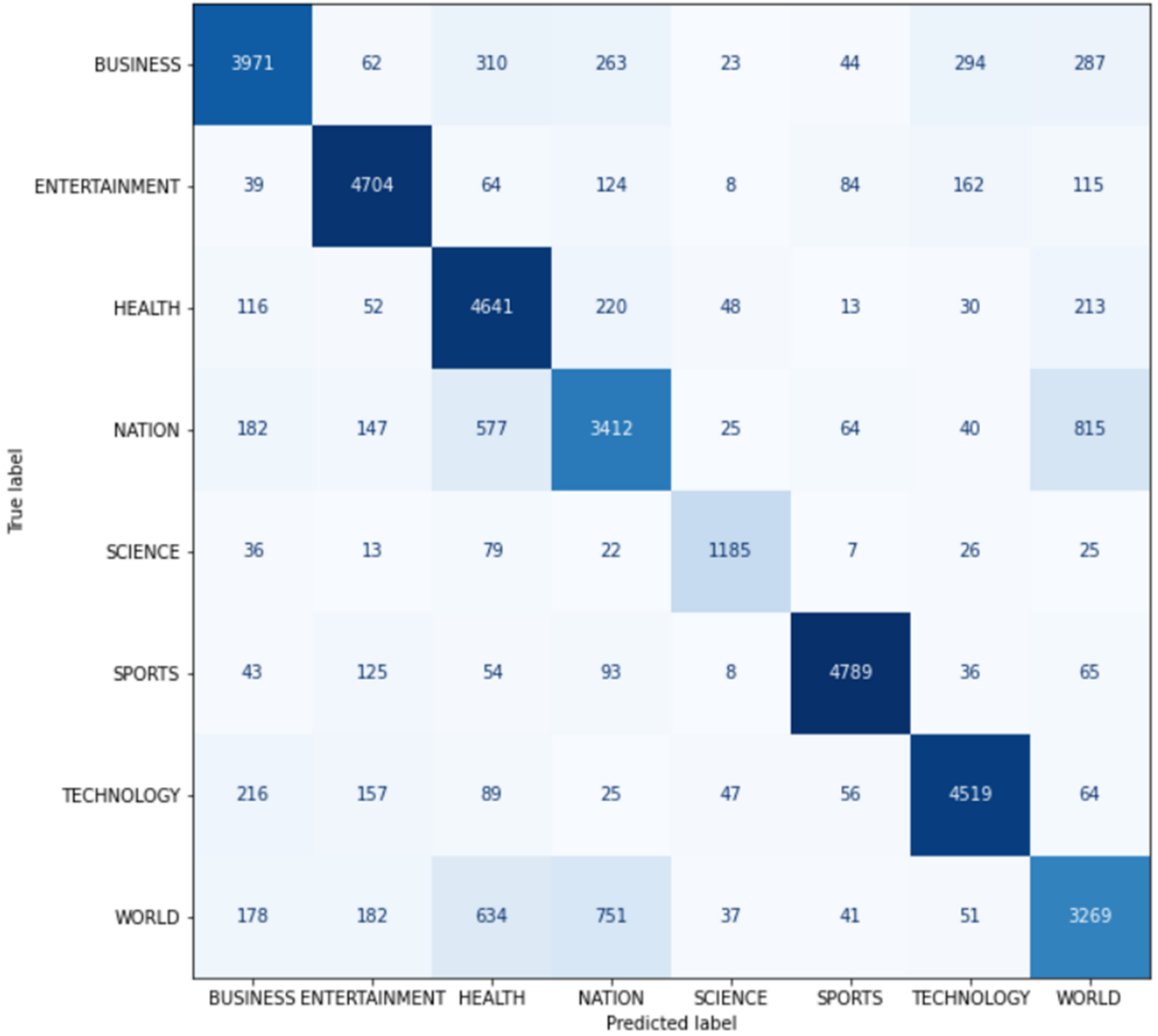
	tfidf
eas	0.447214
moderna	0.447214
patent	0.447214
protections	0.447214
vaccine	0.447214

original:

```
'Moderna eases up on vaccine patent protections'
```

Model Accuracy





Accuracy: 80%

Recall: 81%

Precision: 81%

common ground

WORLD

NATION

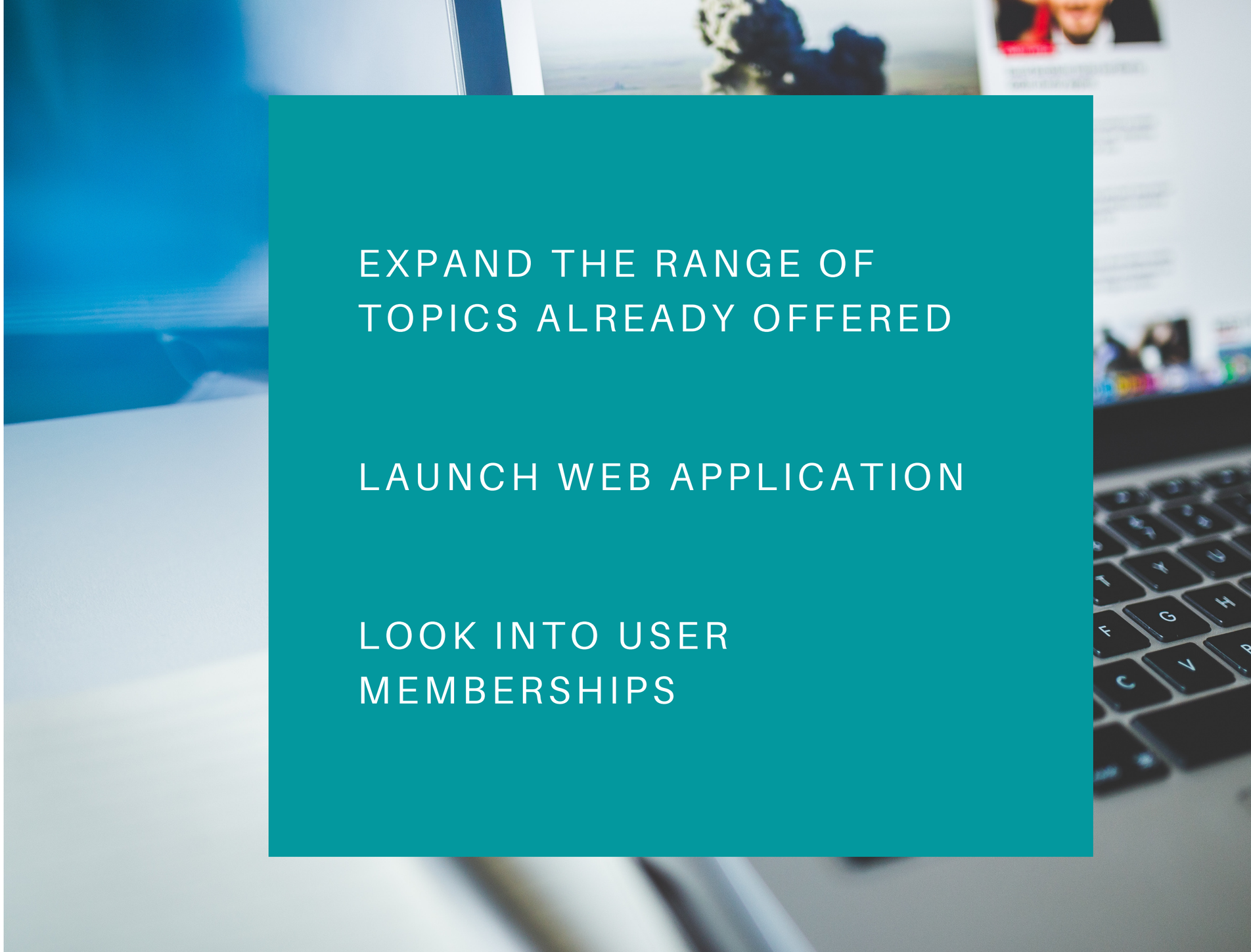
HEALTH

Recommendations

USE MODEL TO PREDICT NEWS TOPICS BY THEIR HEADLINE

USE MODEL TO ONLY SHOW TOPICS/HEADLINES USER IS INTERESTED IN AT THAT TIME

Next Steps



EXPAND THE RANGE OF TOPICS ALREADY OFFERED

LAUNCH WEB APPLICATION

LOOK INTO USER MEMBERSHIPS

Thank You!

FEEL FREE TO REACH ME AT:

DEAUDREY011@GMAIL.COM
GITHUB.COM/@MADORIATHOMAS