
CNN
NEWS
CATEGORY
CLASSIFICATION
HENRY HALIM OKTAKUSUMA

OUTLINE

1
BACKGROUND

2
DATA
CLEANING

3
TEXT
PROCESSING

4
EDA

5
MODELING

6
CONCLUSION

7
RECOMENDATION

BACKGROUND

CNN News is an international online news portal that provides news from all areas of information. The data used is news published from 2012-2022 by CNN News.

Published news has categories that are classified **manually** by the author, so there may be categories that are not relevant to the news description.

I want to create machine learning that automatically classifies news categories based on descriptions

DATA CLEANING

The data consists of 4076 rows and there are no missing values and no duplicated data

3 Features used:
Date Published,
Category, Description

9 Feature:
Author, Date Published,
Category, Section, Url,
Headline, Description, Keywords,
Second Headline, Article Text

The text in the Description will be classified into Categories

DATA CLEANING

Date Published is transformed into feature month published and year published

	Category	Description	Month published	Year published
0	news	The e-commerce boom has exacerbated a global t...	July	2021
1	news	Working in a factory can mean doing the same t...	May	2021
2	news	In a Hong Kong warehouse, a swarm of autonomou...	June	2021
3	business	For many years, the world's most popular emerg...	March	2022
4	business	The European Union formally approved on Tuesda...	March	2022

TEXT PROCESSING

01 **Change all letters to lowercase**

So that the same words are not detected differently just because they differ in capital letters

02 **Delete special character and numbers**

Special characters and numbers have no meaning in text analysis, so they need to be deleted

03 **Remove Stop Words**

Most of the text data contains common words that have no meaning and of course this will affect the accuracy of the analysis results

04 **Group words that have the same meaning**

It is necessary to get valid word from variations of a word

05 **Take root words**

taking the root of a valid word which may have an inflection in the word

06 **Tokenizing to separate words in a sentence**

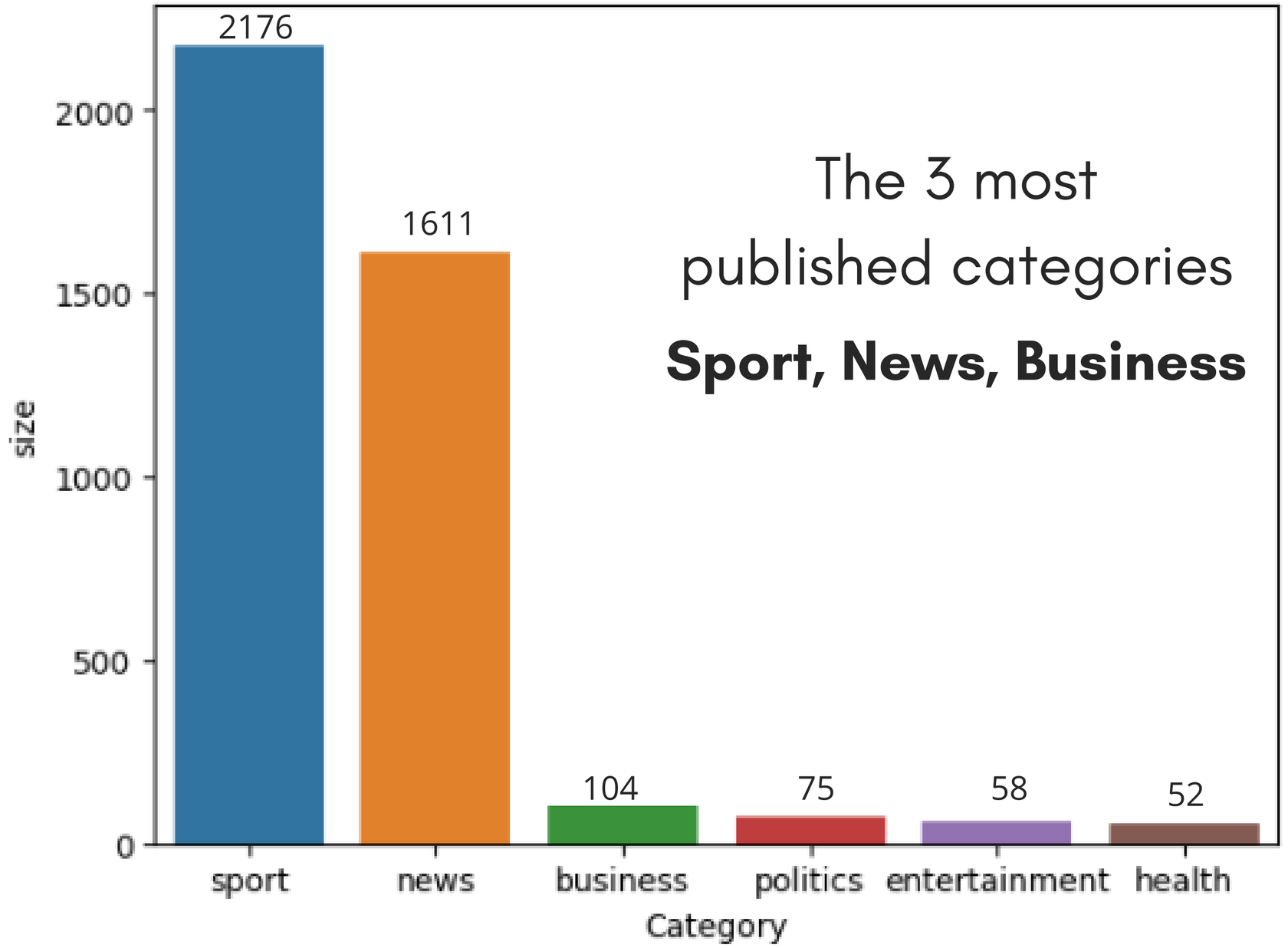
Tokenizing is used to obtain pieces of words that are mutually exclusive and become a value

EXPLORATORY DATA ANALYSIS

There are 6 categories in CNN News data for the 2012-2022 period

- SPORT
- NEWS
- BUSINESS
- POLITICS
- ENTERTAIN MENT
- HEALTH

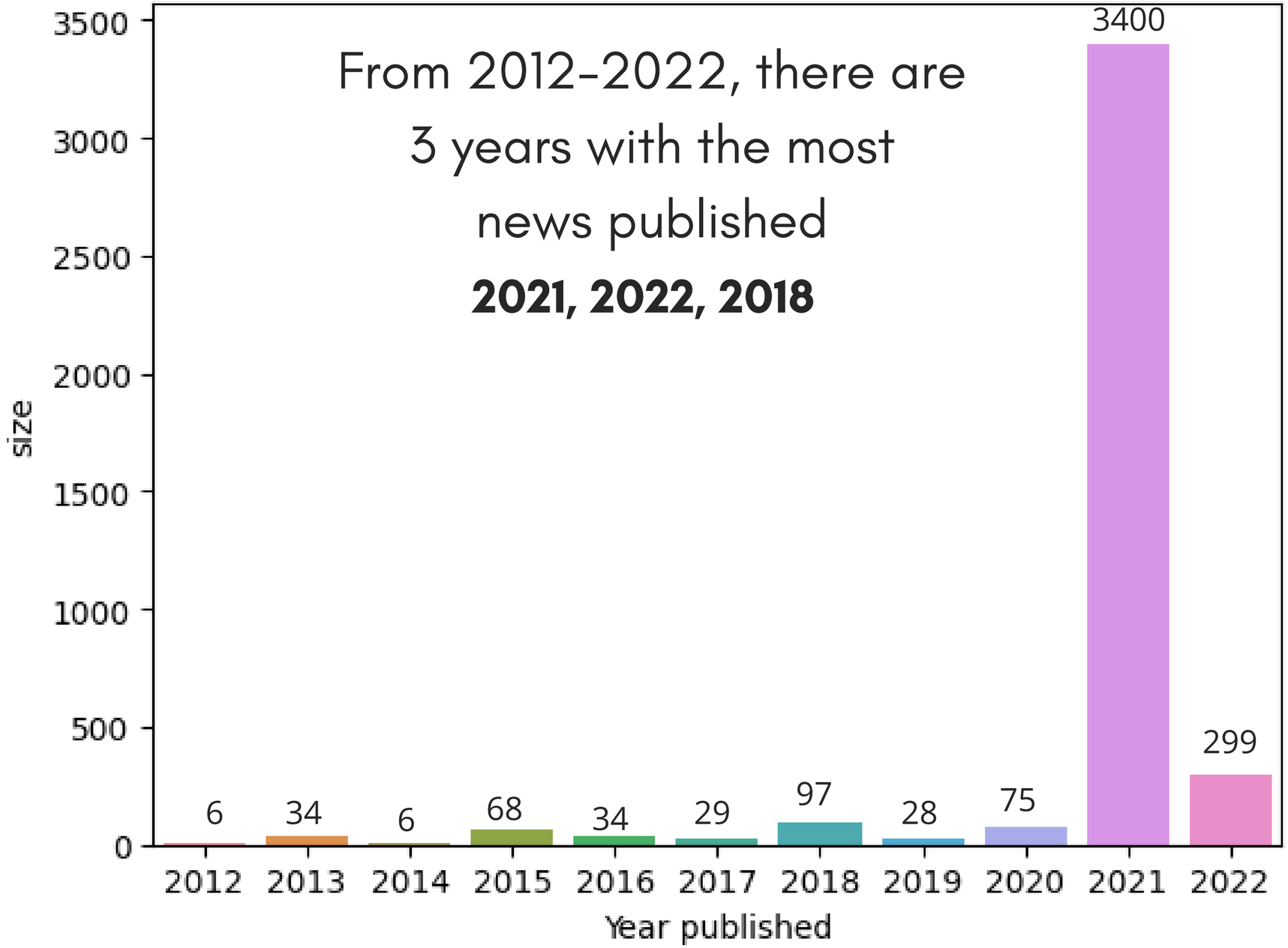
EXPLORATORY DATA ANALYSIS



The 3 most published categories

Sport, News, Business

EXPLORATORY DATA ANALYSIS

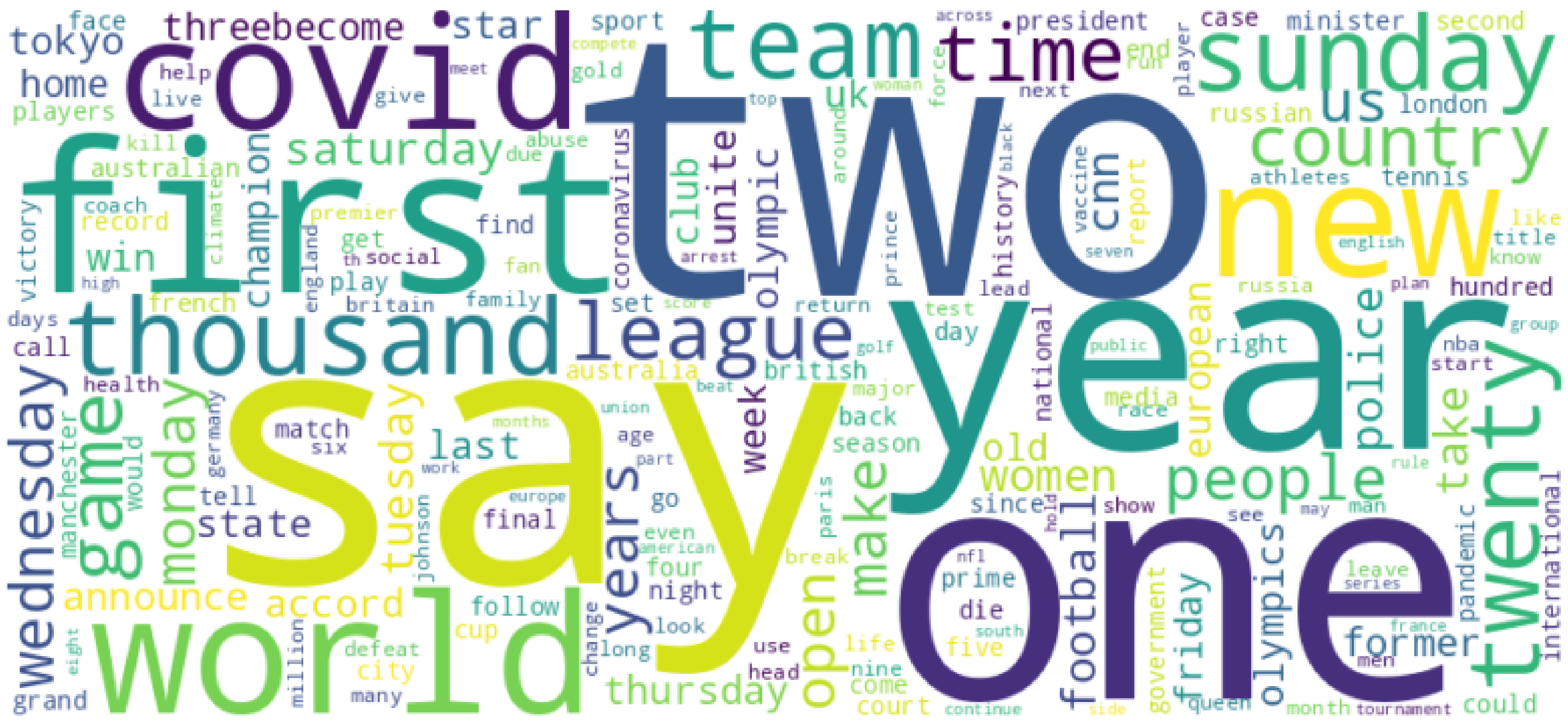


EXPLORATORY DATA ANALYSIS

WORD CLOUD

The most published year is 2021, so the word Covid appears a lot.

The Sport category has the most publications, so words related to sports and competitions appear the most



MODELING

BASELINE

MODEL 1

LOGISTIC REGRESION ACCURACY 85%

MODEL 2

DECISION TREE ACCURACY 75%

MODEL 3

RANDOM FOREST ACCURACY 83%

Best model is Logistic Regresion

The Decision Tree and Random Forest models experienced a significant decrease in accuracy after tuning, while Logistic Regression had relatively the same level of accuracy

TUNING

MODEL 1

LOGISTIC REGRESION ACCURACY 85%

MODEL 2

DECISION TREE ACCURACY 67%

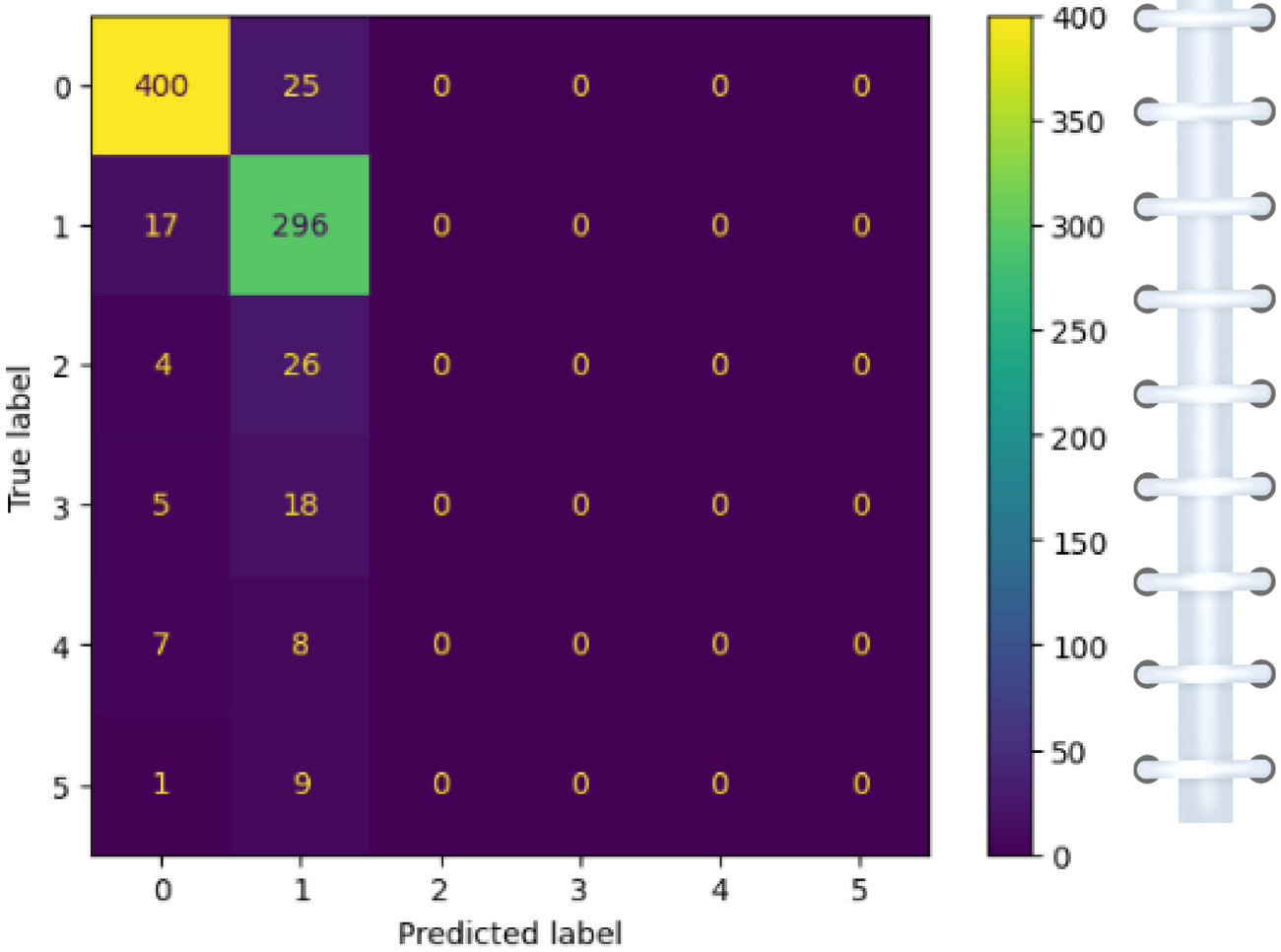
MODEL 3

RANDOM FOREST ACCURACY 57%

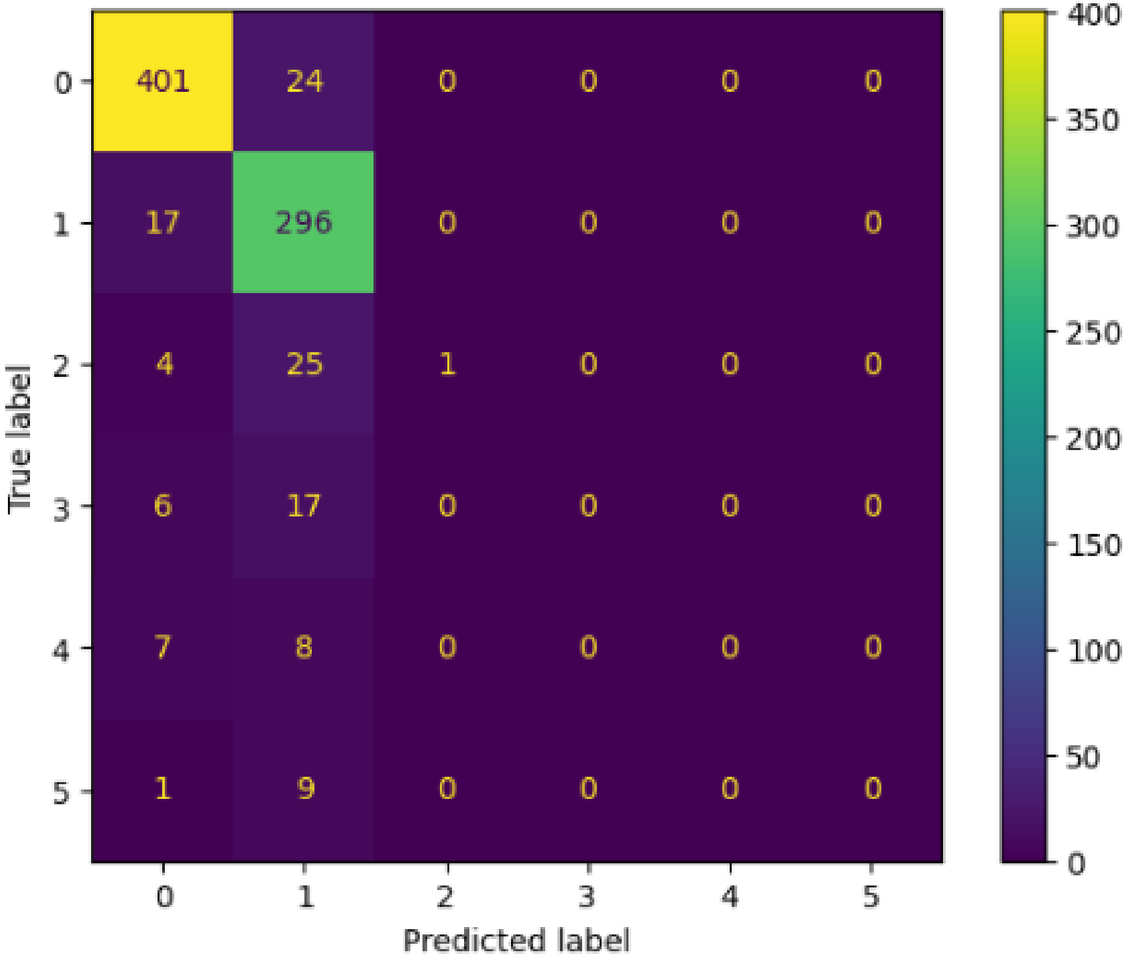
CONFUSION MATRIX

LOGISTIC REGRESION

BEFORE TUNING



AFTER TUNING



CONCLUSION

1. Dominate the Sports and News categories
CNN focuses more on reporting on sports and general news

2. 2021 will be the year with the most news publications, this is due to the high level of Covid, so many people have become Authors for CNN and many readers want to know sports news during the Covid era.

RECOMENDATION

Analyze Reader Trends

to be able to optimize news and improve news in other categories

Diversification of News Categories

I made suggestions to increase the diversity of news topics. Increase news publications with political, business, health and entertainment categories

VIRAL NEWS

News publications with categories that are currently viral


Thank You