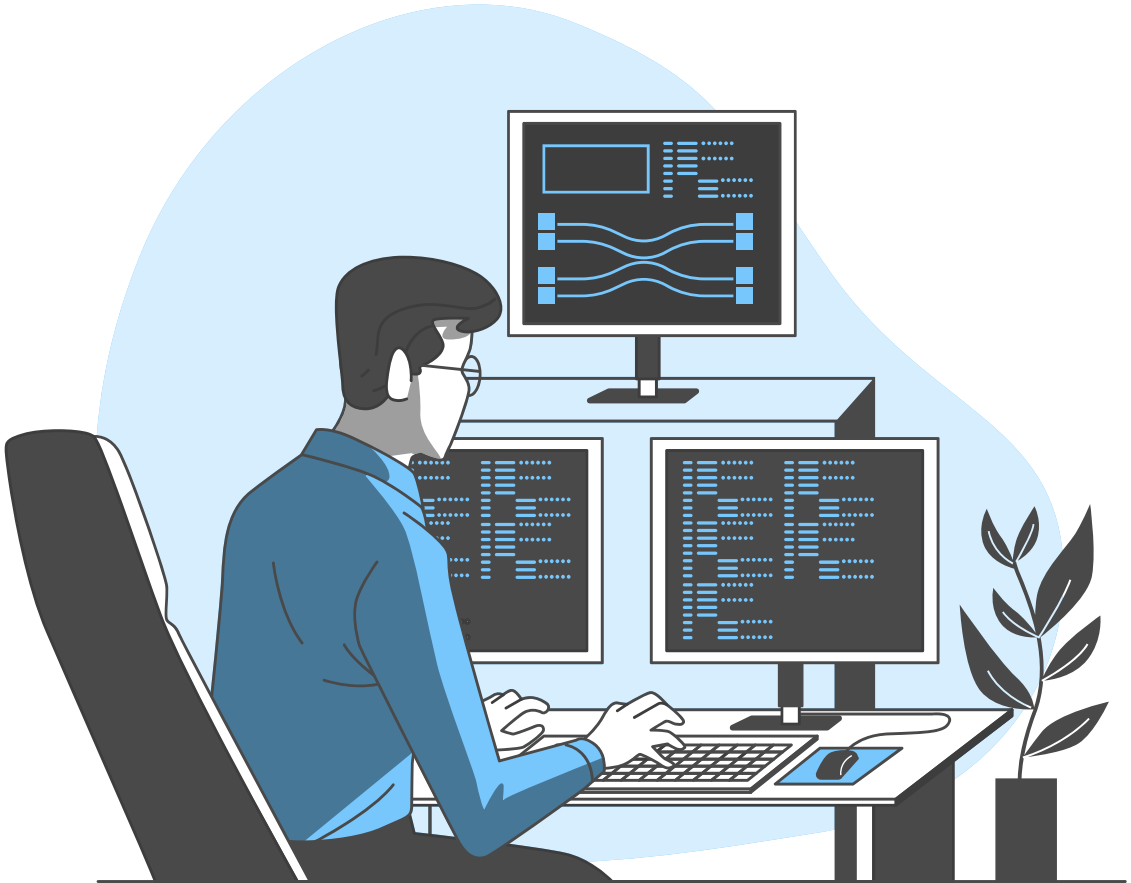


Publication Prediction

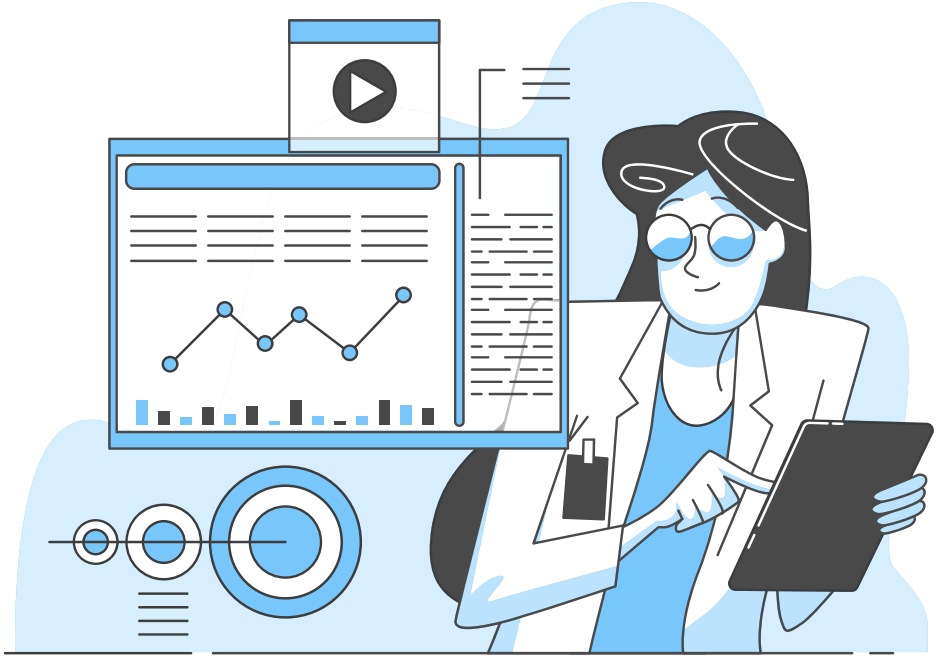
By Kennedy Bagnol

Purpose

01 NLP

02 News and Politics

03 Relevancy



Problem Statement(s)

Can we build a model that determines the political bias of a news article?

Can we build a model that can determine what publication an article comes from?

Data

1. Kaggle Dataset
2. 140,000+ articles
3. 2015-2017

Publications:

- Buzzfeed News
- CNN
- Vox
- Atlantic
- Guardian
- Washington Post
- New York Times
- NPR
- Reuters
- New York Post
- Fox News
- National Review
- Breitbart

Data Cleaning

10,000

Caveats

Data is from a relatively small time frame

01

Due to technical limitations I could not make my model as robust as I would like

02

Bias is inherently subjective

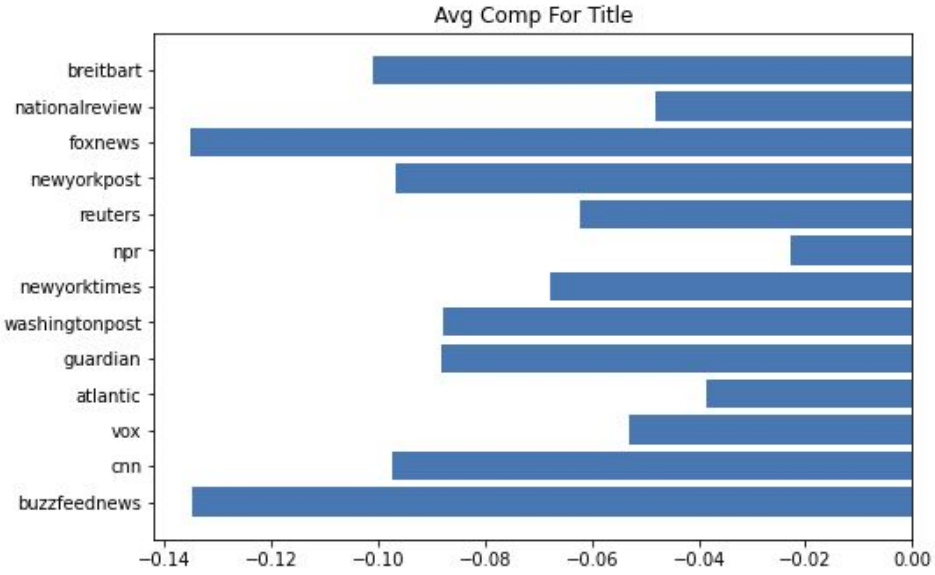
03

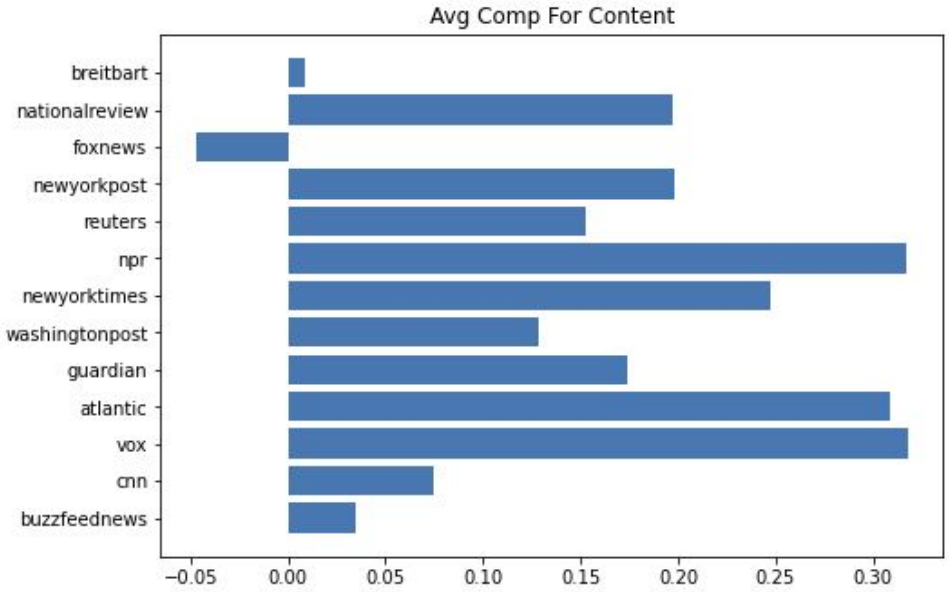
EDA

Sentiment

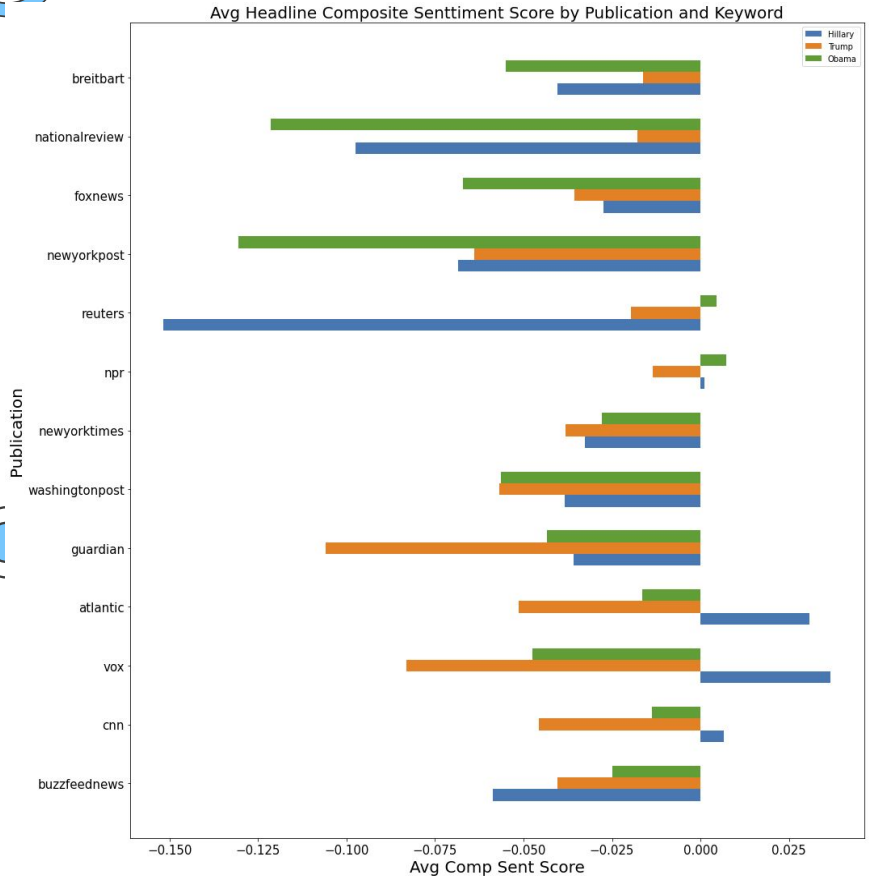
Topics

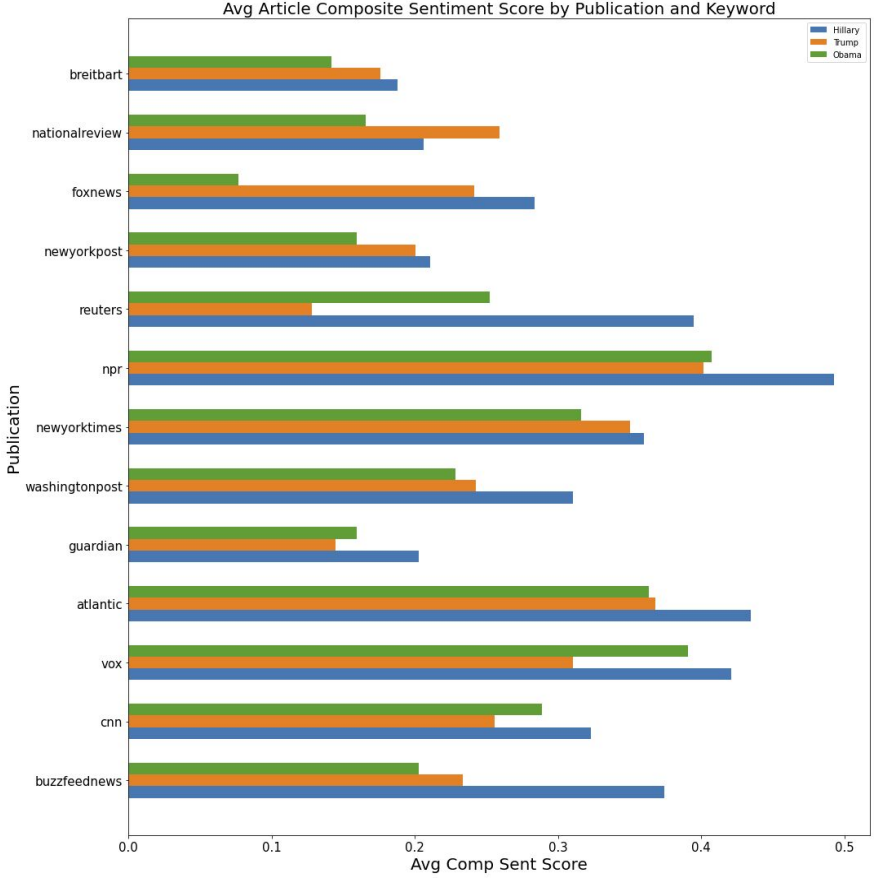
EDA (Sentiment)





EDA (Sentiment)





EDA (Topics)

Topic analysis is used to determine which abstract ‘topics’ occur together in a collection of documents

Latent Dirichlet Allocation (LDA) is what I used to do topic modeling, many other types though

It is a type of unsupervised learning, somewhat similar to K-means clustering

EDA (Topics)

CNN Titles:

Trump and Russia
Trump and North Korea
Trump and Russia
America and China
Hillary, ISIS, Emails

1. 'trump', 'says', 'vote', 'gop', 'things', 'family', 'people', 'senate', 'paul', 'russia'
2. 'new', 'trump', 'police', 'donald', 'dies', 'killed', 'north', 'dead', 'korea', 'gop'
3. 'trump', 'obama', 'house', 'white', 'day', 'man', 'russia', '2016', 'campaign', 'president'
4. 'china', 'obama', 'state', 'bush', 'democrats', 'know', 'crash', 'american', 'life', 'don'
5. 'clinton', 'trump', 'isis', 'hillary', 'opinion', 'star', 'ban', 'fbi', 'women', 'case'

Breitbart Titles:

Election
Anti GOP Rhetoric
Obama and Gun Laws
Immigrants
ISIS

1. 'breitbart', 'trump', 'donald', 'poll', 'report', 'sanders', 'election', 'illegal', 'twitter', 'obama'
2. 'breitbart', 'trump', 'new', 'donald', 'milo', 'people', 'watch', 'anti', 'campaign', 'paul'
3. 'breitbart', 'trump', 'obama', 'gop', 'gun', 'white', 'rubio', 'news', 'attack', 'house'
4. 'clinton', 'breitbart', 'hillary', 'police', 'black', 'texas', 'border', 'video', 'migrants', 'debate'
5. 'breitbart', 'cruz', 'state', 'islamic', 'ted', 'man', 'muslim', 'year', 'day', 'women'

EDA (Topics)

CNN Articles:

Law
Police and People
Trump and Russia
Foreign Fears
Trump vs Clinton

1. 'said', 'court', 'law', 'state', 'federal', 'case', 'department', 'attorney', 'justice', 'states'
2. 'said', 'people', 'police', 'told', 'just', 'like', 'time', 'says', 'new', 'family'
3. 'trump', 'said', 'president', 'house', 'white', 'administration', 'obama', 'russia', 'campaign', 'intelligence'
4. 'said', 'isis', 'military', 'north', 'united', 'china', 'government', 'korea', 'country', 'syria'
5. 'trump', 'clinton', 'said', 'campaign', 'president', 'republican', 'obama', 'new', 'party', 'people'

Breitbart Articles:

What people are saying
Obama is Muslim
Breitbart on Twitter, Gun March
Political Campaigns
Police Cases, Texas

1. 'said', 'people', 'like', 'think', 'just', 'going', 'don', 'know', 'want', 've'
2. 'said', 'obama', 'president', 'state', 'states', 'united', 'islamic', 'government', 'muslim', 'security'
3. '2016', 'twitter', 'says', 'news', 'pic', 'breitbart', 'https', 'gun', 'march', 'july'
4. 'trump', 'clinton', 'hillary', 'donald', 'campaign', 'president', 'cruz', 'republican', 'said', 'presidential'
5. 'police', 'said', 'breitbart', 'reported', 'texas', 'according', 'year', 'new', '000', 'told'

EDA (Topics)

Buzzfeed:

Trump and Social Media

??????

Company and Law

Social/Gender Dynamics

Presidency and Law

1. 'trump', 'said', 'like', 'people', 'news', 'facebook', 'just', 'twitter', 'new', 'think'
2. 'la', 'le', 'et', 'en', 'les', 'des', 'que', 'est', 'dee', 'il'
3. 'said', 'company', 'news', 'buzzfeed', 'new', 'court', 'told', 'year', 'according', 'case'
4. 'said', 'people', 'women', 'like', 'just', 'time', 'years', 'told', 'family', 'men'
5. 'said', 'trump', 'police', 'president', 'told', 'people', 'news', 'state', 'officials', 'clinton'

Modeling

01 Data Prep

02 BERT Transformer

03 Neural Network

04 Fine Tuning

BERT Transformer?

BERT Transformer

Developed by Google and released in 2018

Trained on millions of articles from Wikipedia

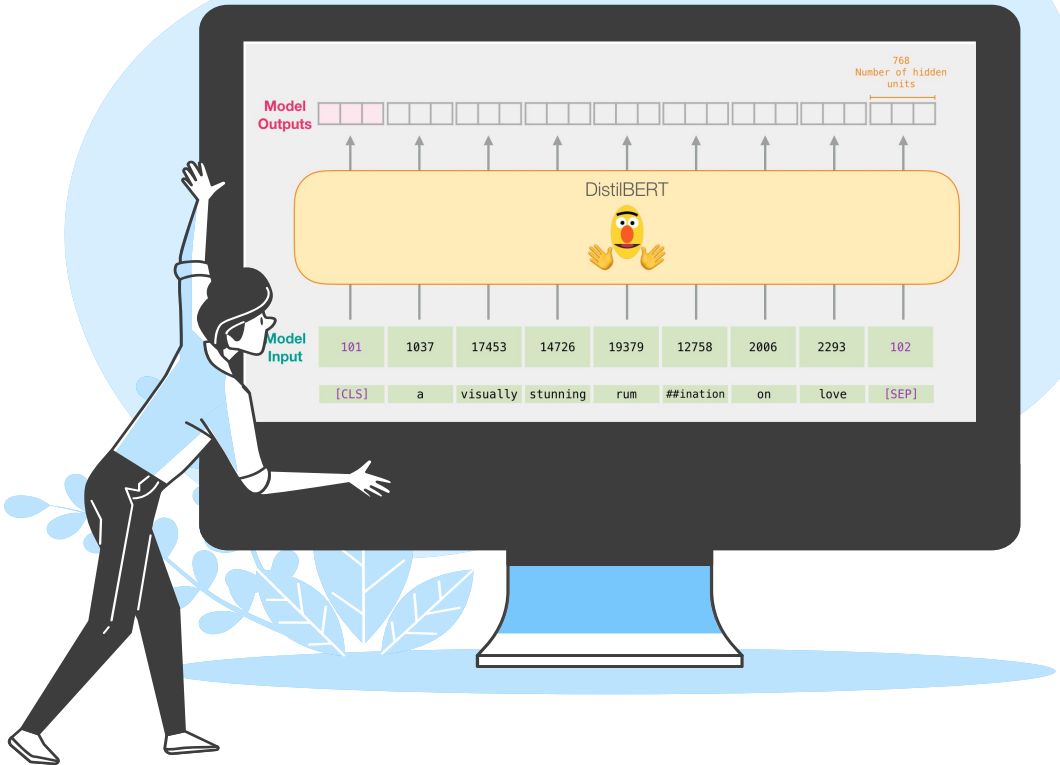
Contextual Embeddings

"They did not give him a fair chance"

"Did you go to the fair today?"

The BERT_base model has 110 Million Parameters

BERT_large has over 345 Million!



A Tale of Two Models

Bias

39,000 Rows
13 Classes

Publication

12,500 Rows
5 Classes

BERT Transformer
Standard NN
8192 Nodes
.5 Dropout

Bias Model Methodology

01

Each publication was put into a separate category of bias taken from allsides.com

02

Filtered them to only pull political articles

03

From this subsection, I sampled 2500 posts from each category (12,500) to train my model

Publication Model Methodology

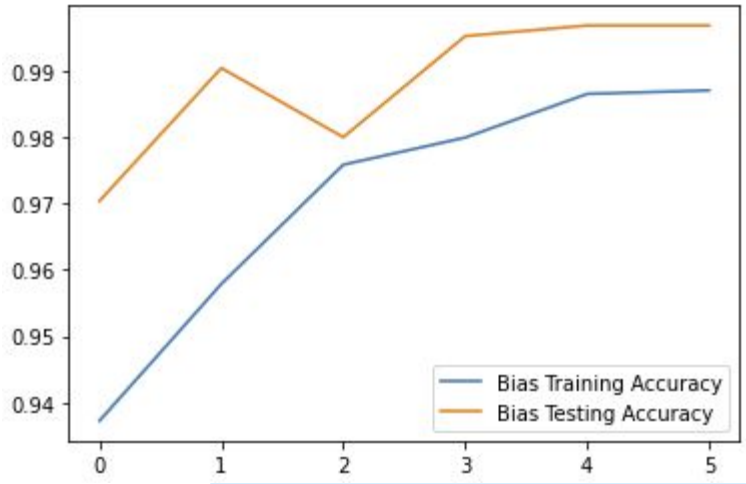
01

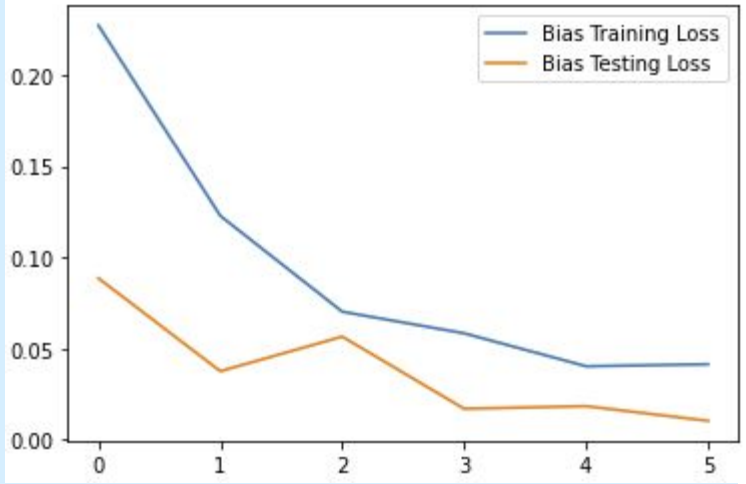
Created a separate dataframe for each publication

02

Sampled 3,000 articles from each publication
146,000 → 39,000

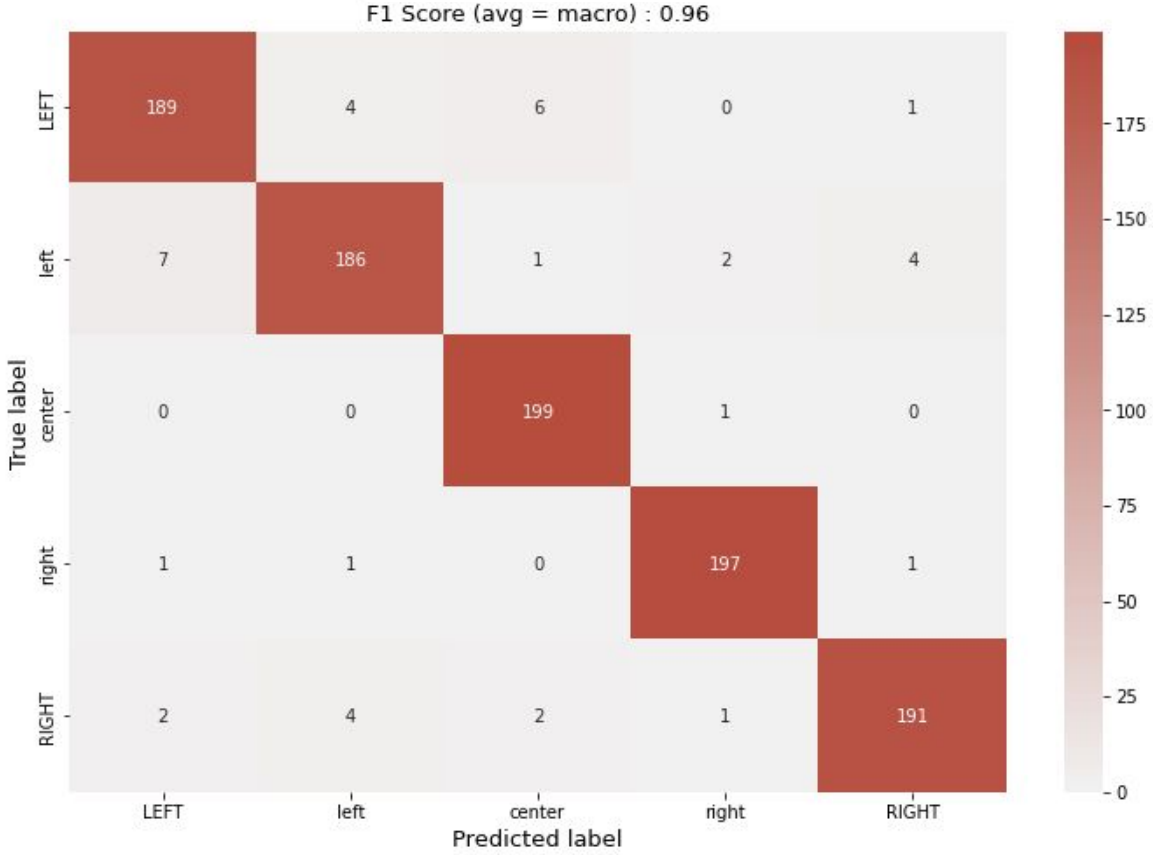
Results!

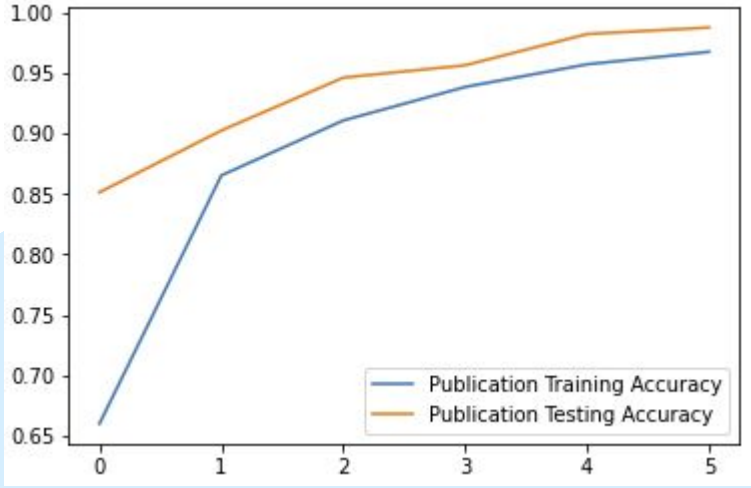

0.99
0.98
0.97
0.96
0.95
0.94
0
1
2
3
4
5
Bias Training Accuracy
Bias Testing Accuracy

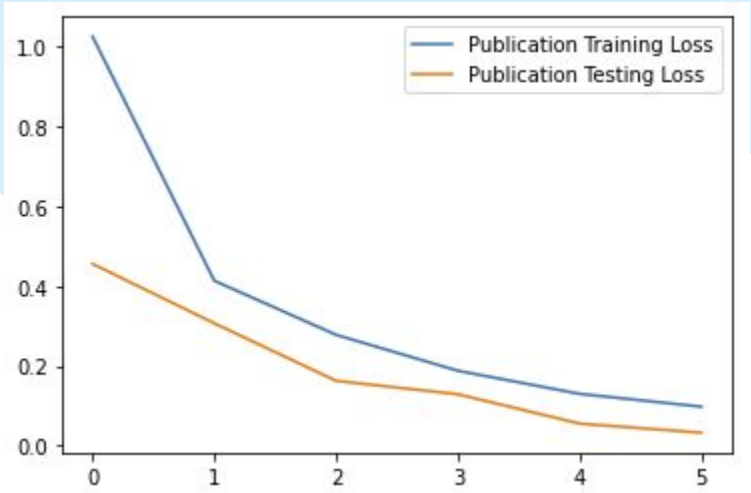

0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
Bias Training Loss
Bias Testing Loss

accuracy			0.96	1000
macro avg	0.96	0.96	0.96	1000
weighted avg	0.96	0.96	0.96	1000

Matthews Correlation Coefficient: 0.9525595368316362
hamming loss: 0.038

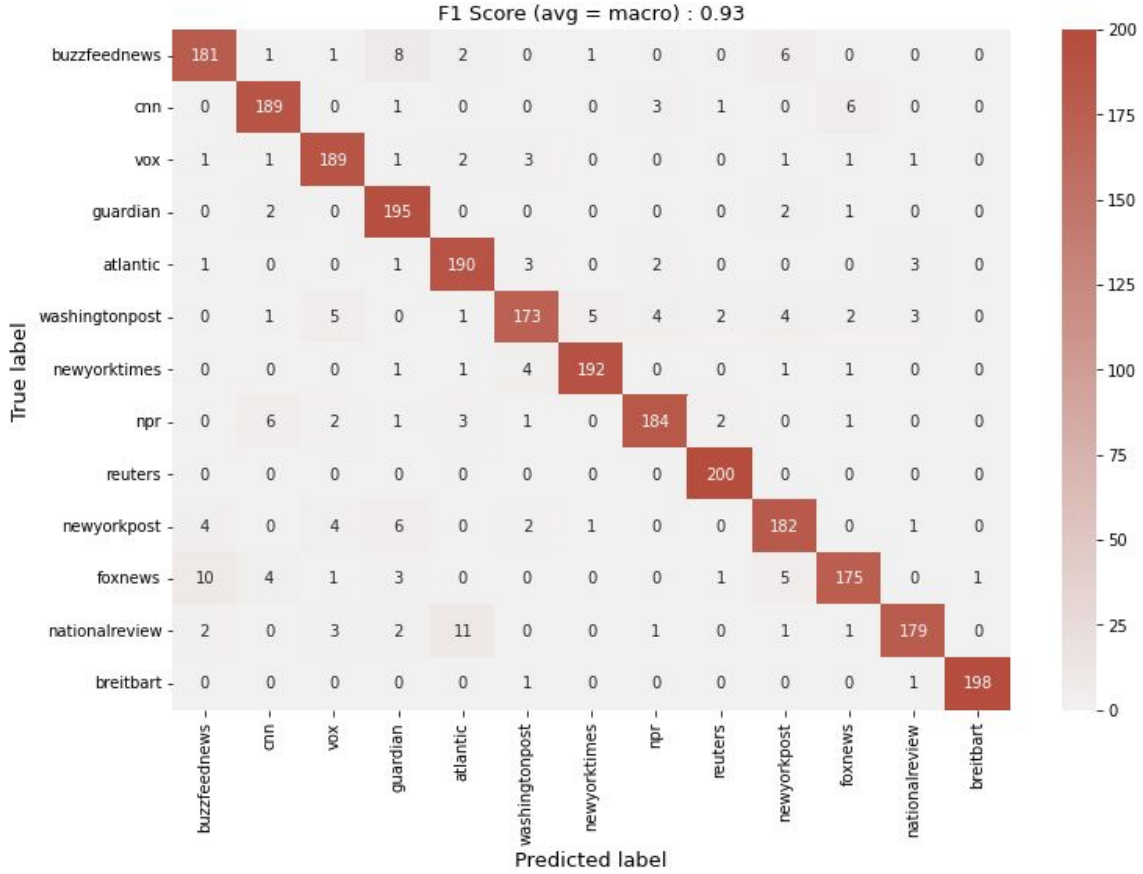



1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0
1
2
3
4
5
Publication Training Accuracy
Publication Testing Accuracy


1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
Publication Training Loss
Publication Testing Loss

accuracy			0.93	2600
macro avg	0.93	0.93	0.93	2600
weighted avg	0.93	0.93	0.93	2600

Matthews Correlation Coefficient: 0.9279938539809546
hamming loss: 0.06653846153846153



Applications

Plagarism

Response Organization

Case Search

Future

Explore the Neural Network

Visualize Structure

More Power

END

Do you have any questions?

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, infographics & images by Freepik and illustrations by Stories