Satirical Headlines

Recognizing Truth and Lies

Brian Nguyen
at Metis

Headlines

Year 2020

1,081 — **/r/TheOnion (Fake)**

'Iron Man' Trailer To Be Made Into Feature Film

36,262 — **/r/NotTheOnion (Real)**

Wild Turkey that Shut Down an Oakland Park Remains on the Loose

Process and Technologies

DATA

Pushshift API
PRAW (Reddit API)

PRE-PROCESSING

Pandas
Numpy

EDA

Pandas
Numpy
Matplotlib/Seaborn

TOPIC MODEL

Scikit-Learn
Matplotlib/Seaborn
t-SNE
Gensim

CLASSIFICATION

XGBoost
VADER Sentiment
Scikit-Learn

FUTURE WORK

Word2Vec
Neural Networks

Python and Jupyter Notebook were used throughout this entire project

Iterative Workflow

Pre-Processing

Text Cleaning

Punctuation
Word (3) Minimum

Vectorize

TF-IDF
bigrams

Topic Model or Cluster

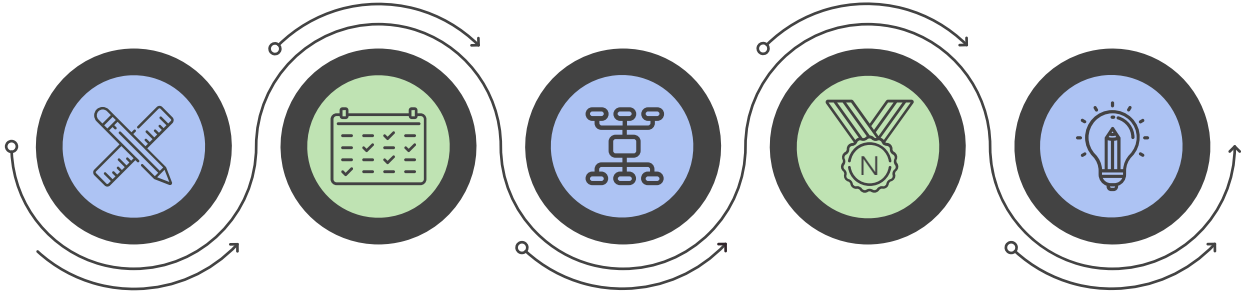
LSA (SVD)
Doc-Topic Matrix

Qty Topics/Clusters

Six Topics
Explained Variance

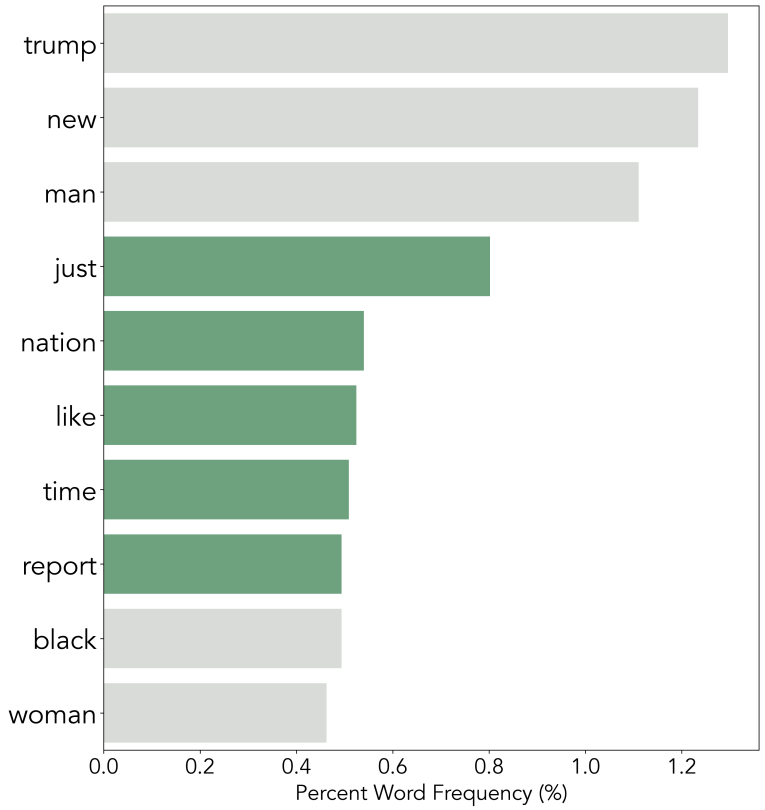
Interpret

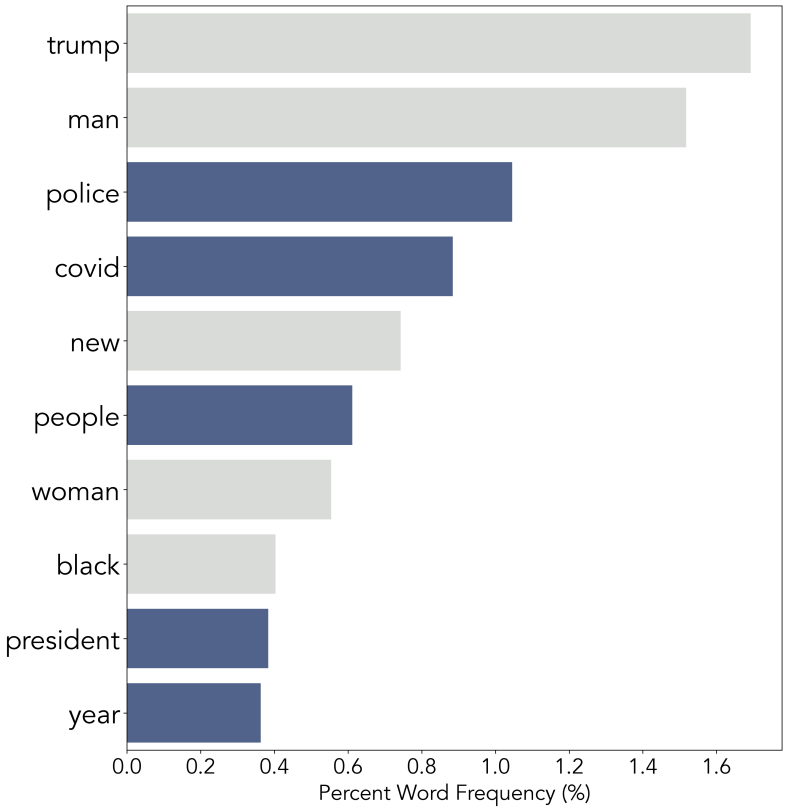
Sensible Topics
t-SNE Visualization



Text under **bold** are final models/functions/parameters used (or select examples)

Top 10 Words

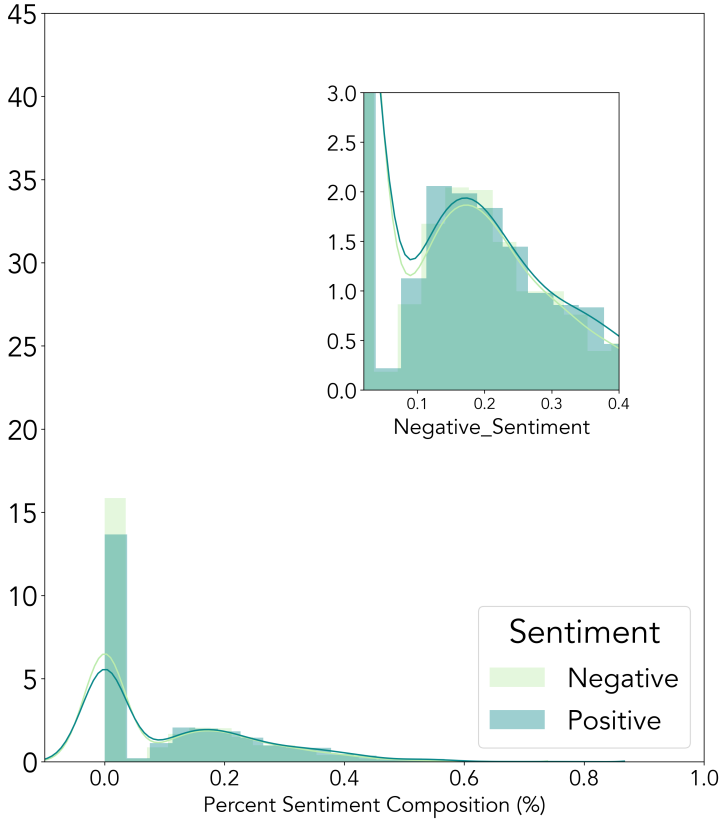




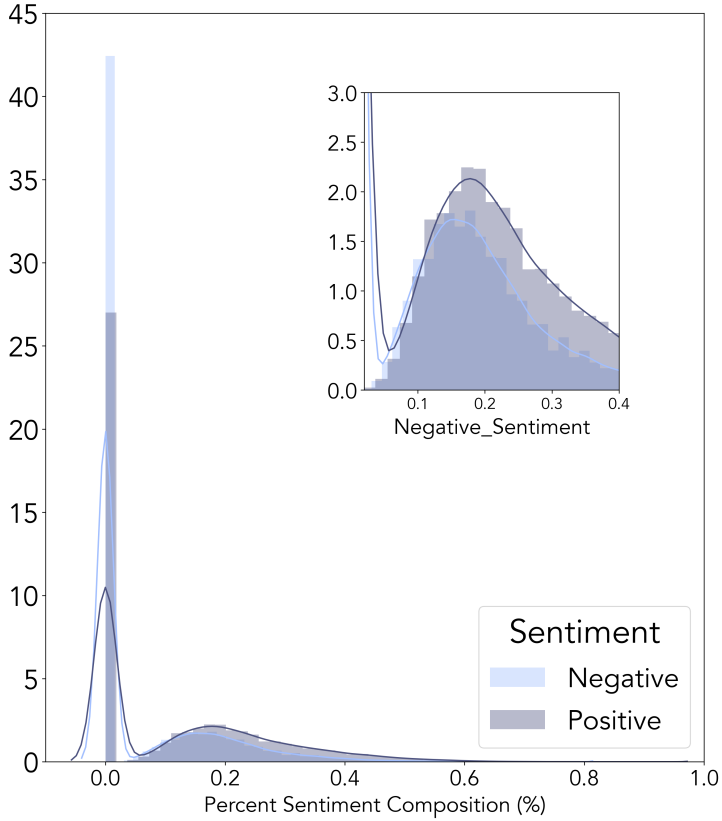
EDA

Distribution of Sentiment

/r/TheOnion



/r/NotTheOnion



Top Words from Topics

Toilet Paper

police ask
stuck nose
trial combat
police department
toilet paper
parking lot
amid covid
police stop
city council
jail time

Black Lives Matter

premier league
stay home
video posted
police officer
cream
black matter
climate change
social media
fox news
tested positive

Current Affairs

wearing mask
space force
super bowl
amid pandemic
president trump
animal crossing
trump campaign
trump rally
corona virus
tiger king

Police

viral video
police arrest
black people
grocery store
corona beer
police say
high school
prime minister
police brutality
law enforcement

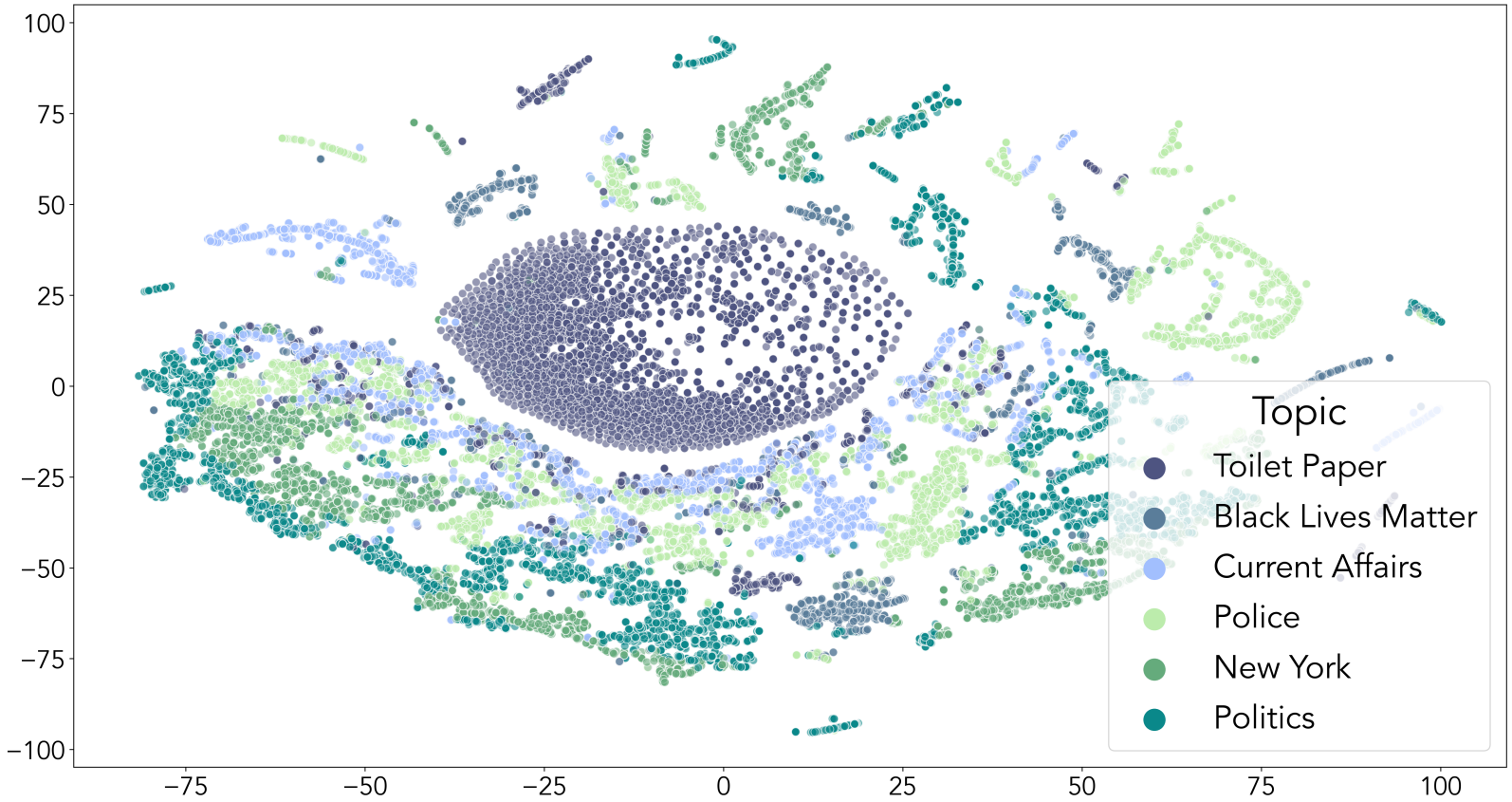
New York

trump new
test positive
face mask
positive covid
bin laden
new york
new jersey
police chief
supreme court
strip club

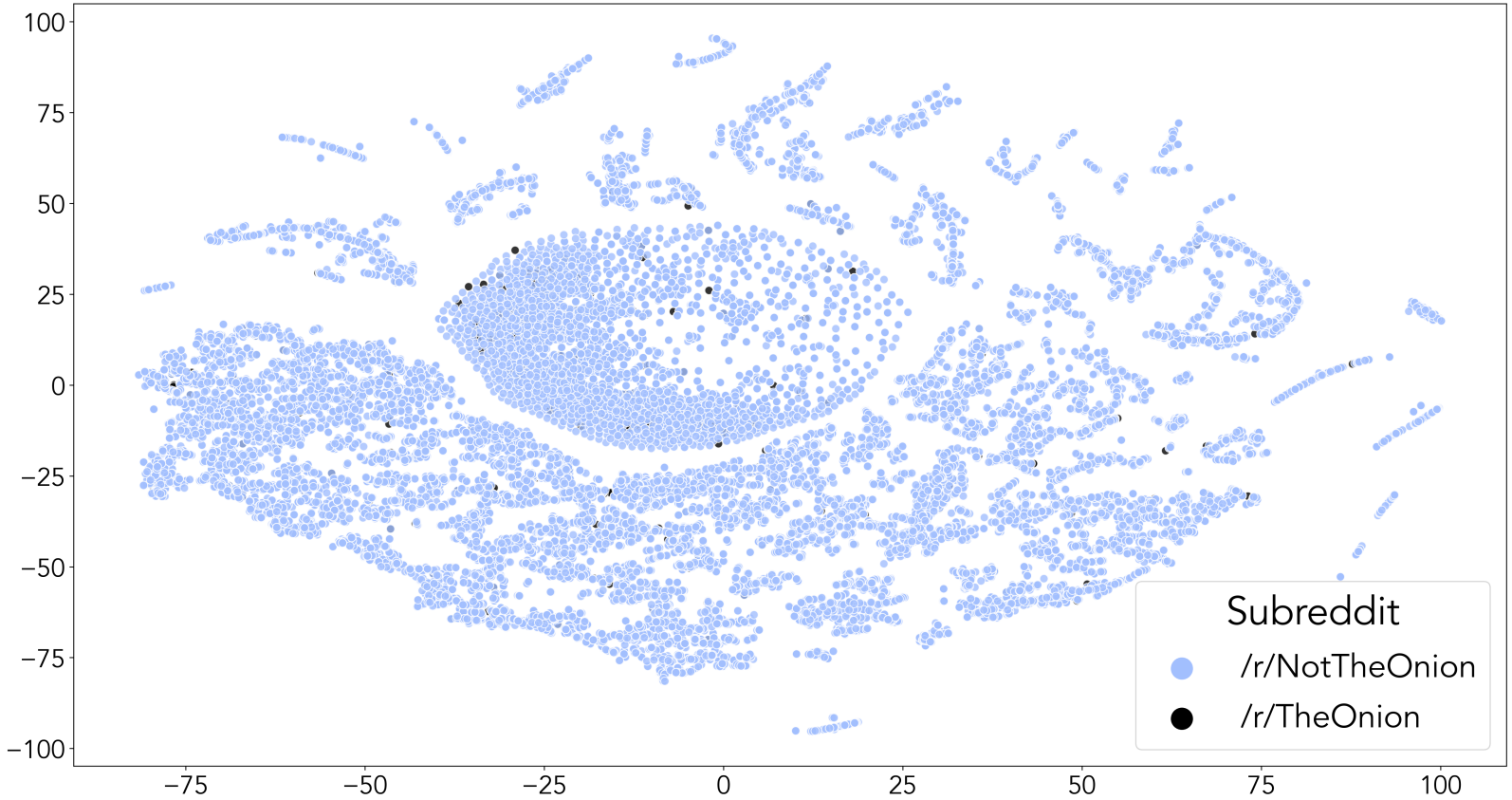
Politics

task force
video game
paw patrol
press secretary
conspiracy theory
white house
national guard
trump administration
world news
press conference

t-SNE Clustering by Topic



t-SNE Clustering by Subreddit



Classification with XGBoost

Topics Only	Predicted (–)	Predicted (+)
Actual (–)	76	148
Actual (+)	38	175

Accuracy = 55%

F1 Score = 63%

6 Topics

Topics and Sentiments	Predicted (–)	Predicted (+)
Actual (–)	103	125
Actual (+)	67	144

Accuracy = 53%

F1 Score = 57%

6 Topics + 3 Sentiments

Classification better than a coin toss

(+) Onion and (–) NotTheOnion

K-fold =10; Under-sampled 1:1

Future Work

REGULAR NEWS

Include “normal” news into analysis (/r/news)

Subreddits in this analysis are too alike

WORD2VEC

Doc-topic matrix from LSA were near-zero values

Google’s library was built from news headlines

NEURAL NETWORK

Using neural networks to improve classification

Use more data (from other years)

Thank you

Brian Nguyen
github.com/bt-nguyen

Conclusion

TOPICS

Distinguishable only through bigrams

Toilet paper fiasco is highly distinguishable (t-SNE)

WORDS

Frequently used words overlap (5 from top 10)

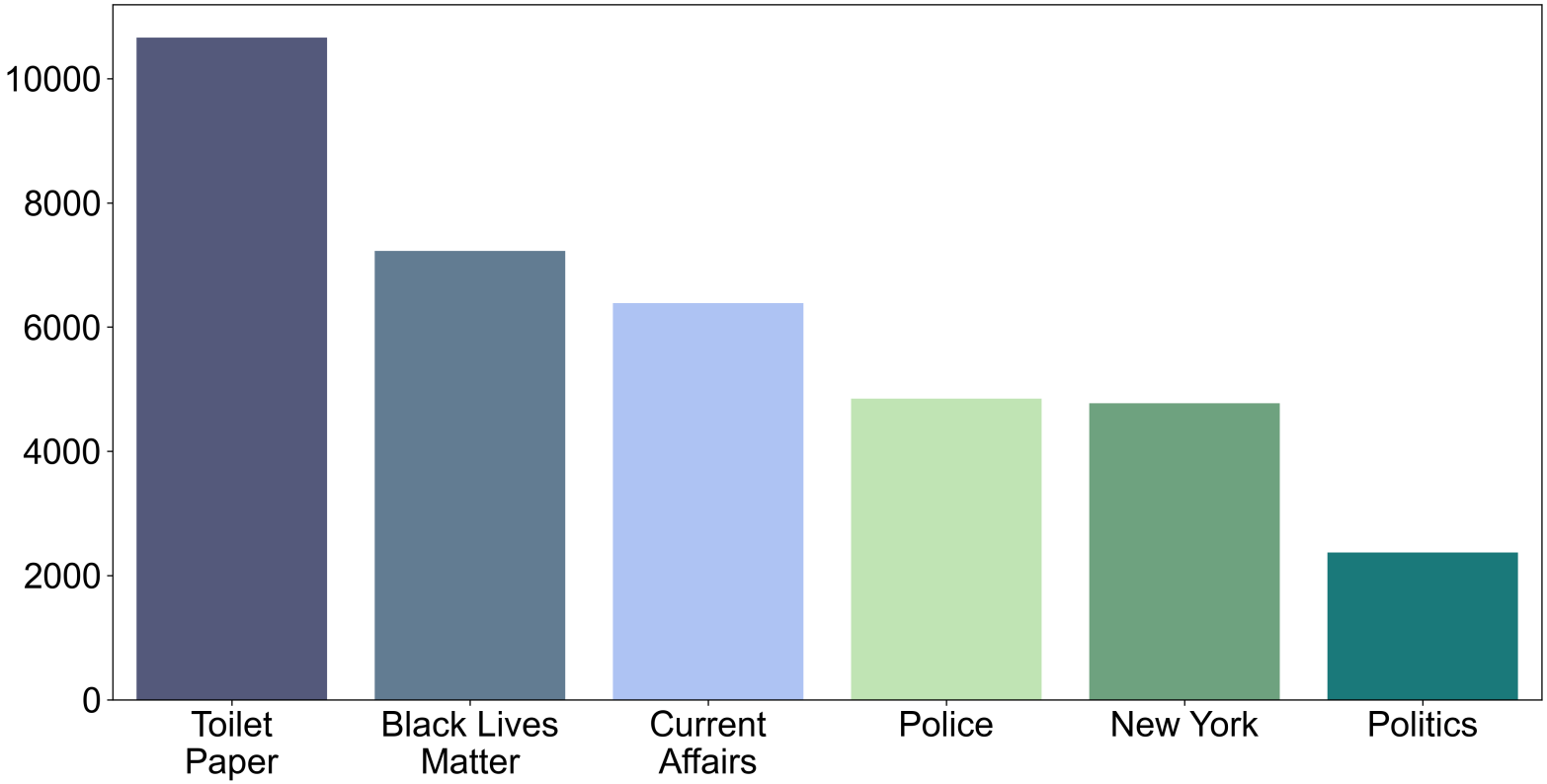
Distinguishable characteristic needs to be found

SENTIMENT

Real and fake is indistinguishable

Little to no sentiment across headlines

Topic Distribution



...and our set of editable icons

You can resize these icons, keeping the quality.
You can change the stroke and fill color; just select the icon and click on the paint bucket/pen.
In Google Slides, you can also use Flaticon's extension, allowing you to customize and add even more icons.

Business Icons

Avatar Icons

Creative Process Icons

Educational Process Icons

Help & Support Icons

Medical Icons

Nature Icons

Performing Arts Icons

SEO & Marketing Icons

Teamwork Icons

...and our sets of editable icons

You can resize these icons without losing quality.
You can change the stroke and fill color; just select the icon and click on the paint bucket/pen.
In Google Slides, you can also use Flaticon's extension, allowing you to customize and add even more icons.

Educational Icons

Medical Icons

Business Icons

Teamwork Icons

Help & Support Icons

Avatar Icons

Creative Process Icons

Performing Arts Icons

Nature Icons

SEO & Marketing Icons