FAKE News Detection

Our Team

Sarah Gamal

Bassant Samy

Amira Mohamed

Samy Ibrahim

Sara Gamil

Ahmed Sherif

1 Introduction
2 Datasets
3 Data Visualization
4 Key Findings from Visualization
5 Train-Test Split
6 Applying Models
7 Comparison of Model Accuracies
8 Confusion Matrices

9 Model Pipeline
10 Joblib Model Saving
11 Conclusion
12 Model Deployment

INTRODUCTION

Problem Statement

In today’s digital age, fake news is a critical problem. False information spreads quickly through social media, causing confusion and misinformation.

Project Goal

build a machine learning model that can detect whether a news article is real or fake by analyzing the content of the article.

Datasets

In this project, we aimed to create a robust Fake News Detection model by using five different datasets. Each dataset has its own characteristics in terms of content, format, and source, which allowed us to capture a wide range of fake and real news articles. Here's a detailed overview of each dataset.

Dataset 1: Fake and Real News Dataset (Kaggle)

- Articles from various online sources, labeled as either real or fake.
Focused mainly on political news.

- Balanced distribution between real and fake news.

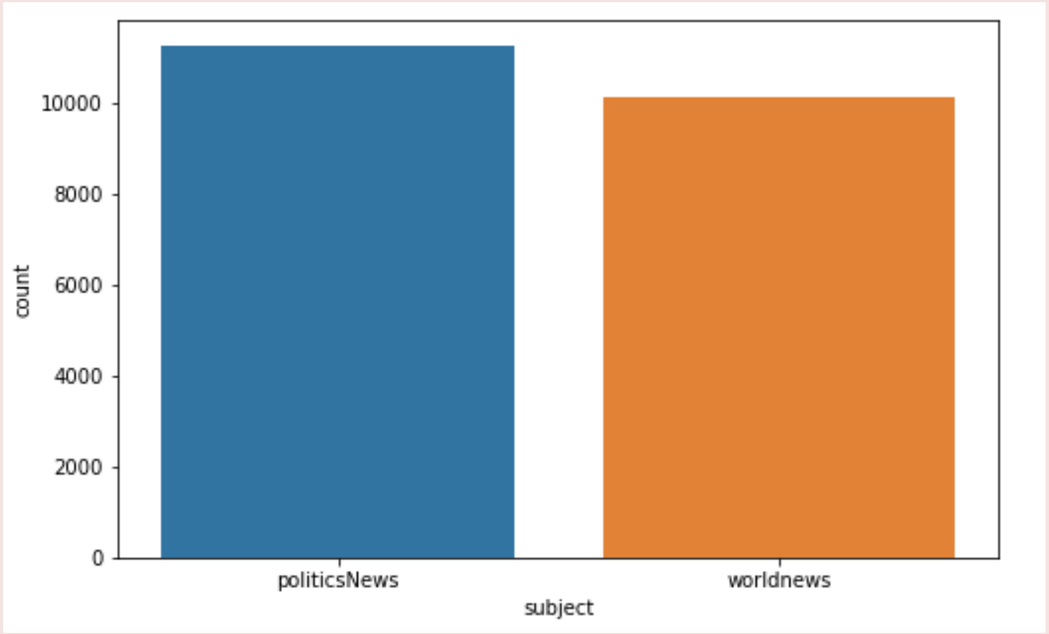
- Size: 44,898 articles.

	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL

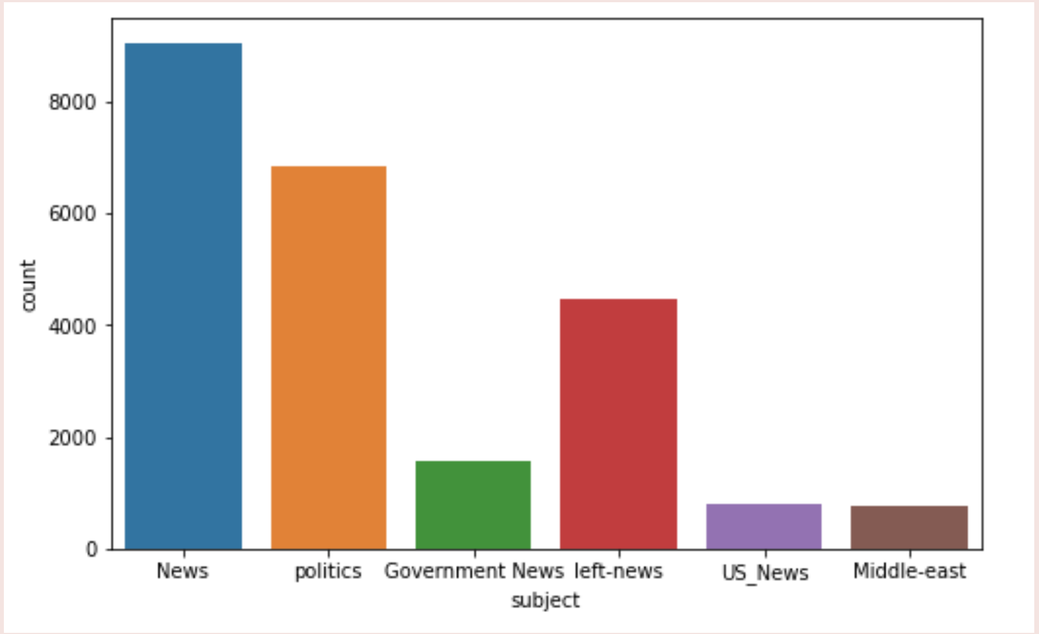
Dataset 2: ISOT Fake News Dataset

Mixture of real and fake news articles, with a focus on political content.

Size: 21,417 articles (10,417 real, 11,000 fake).



$fig_{(1)}$: Real News



$fig_{(2)}$: Fake News

Dataset 3: BuzzFeed News Dataset

- Articles curated and verified by BuzzFeed journalists.
Focus on fake news sites versus mainstream media.

- Size: 2,282 articles.

- Skewed towards fake news, as it was curated for investigative purposes.

	Article	label
0	National Federation of Independent Business	1
1	comments in Fayetteville NC	1
2	Romney makes pitch, hoping to close deal : Ele...	1
3	Democratic Leaders Say House Democrats Are Uni...	1
4	Budget of the United States Government, FY 2008	1
...	...	...
427	Who is affected by the government shutdown?	0
428	Lindsey Graham Threatens To Convert To Democra...	0
429	ELECTORAL COLLEGE ELECTOR COMMITS SUICIDE TO A...	0
430	Sarah Palin Calls To Boycott Mall Of America B...	0
431	Account Suspended	0

1056 rows × 2 columns

Dataset 4: LIAR Dataset

- Short claims made by politicians, fact-checked and labeled as true, half-true, or false.
Contains a mixture of short statements and verdicts.

- Imbalanced, with more false claims compared to true or half-true ones.

- Size: 12,836 statements.

	id	title	author	text	label
0	0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

Dataset 5: Politifact News Dataset

- Articles from Politifact, labeled as true, false, or mixed.
Political fact-checking articles.

- Balanced between true and false labels.

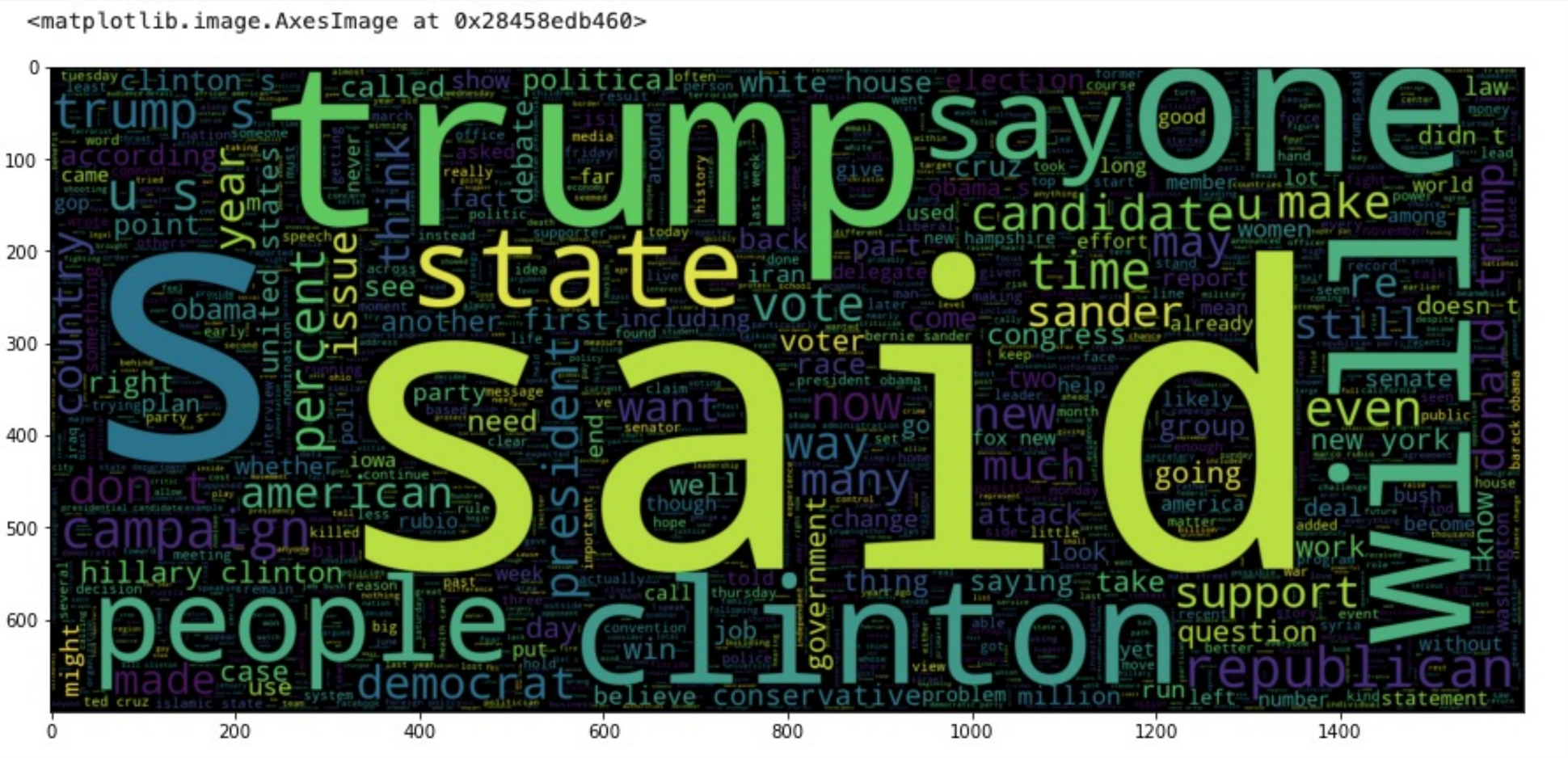
Size: 5,000 articles.

	URLs	Headline	Body	Label
0	http://www.bbc.com/news/world-us-canada-414191...	Four ways Bob Corker skewered Donald Trump	Image copyright Getty Images\nOn Sunday mornin...	1
1	https://www.reuters.com/article/us-filmfestiva...	Linklater's war veteran comedy speaks to moder...	LONDON (Reuters) - "Last Flag Flying", a comed...	1
2	https://www.nytimes.com/2017/10/09/us/politics...	Trump's Fight With Corker Jeopardizes His Legi...	The feud broke into public view last week when...	1
3	https://www.reuters.com/article/us-mexico-oil-...	Egypt's Cheiron wins tie-up with Pemex for Mex...	MEXICO CITY (Reuters) - Egypt's Cheiron Holdin...	1
4	http://www.cnn.com/videos/cnnmoney/2017/10/08/...	Jason Aldean opens 'SNL' with Vegas tribute	Country singer Jason Aldean, who was performin...	1
...	...	...	...	...
4004	http://beforeitsnews.com/sports/2017/09/trends...	Trends to Watch	Trends to Watch\n% of readers think this story...	0
4005	http://beforeitsnews.com/u-s-politics/2017/10/...	Trump Jr. Is Soon To Give A 30-Minute Speech F...	Trump Jr. Is Soon To Give A 30-Minute Speech F...	0
4006	https://www.activistpost.com/2017/09/ron-paul-...	Ron Paul on Trump, Anarchism & the AltRight	NaN	0
4007	https://www.reuters.com/article/us-china-pharm...	China to accept overseas trial data in bid to ...	SHANGHAI (Reuters) - China said it plans to ac...	1
4008	http://beforeitsnews.com/u-s-politics/2017/10/...	Vice President Mike Pence Leaves NFL Game Beca...	Vice President Mike Pence Leaves NFL Game Beca...	0

4009 rows × 4 columns

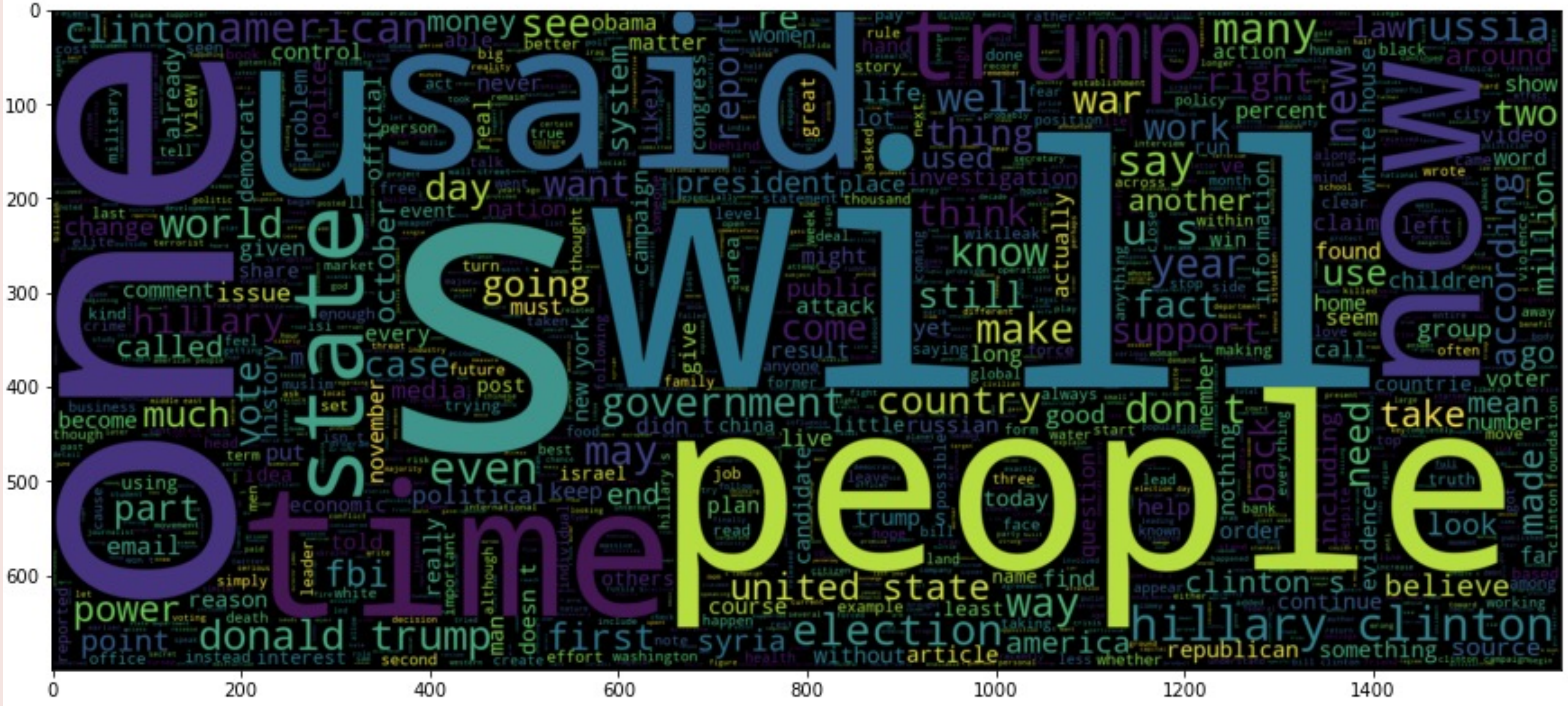
Data Visualization

Dataset 1: Fake and Real News Dataset



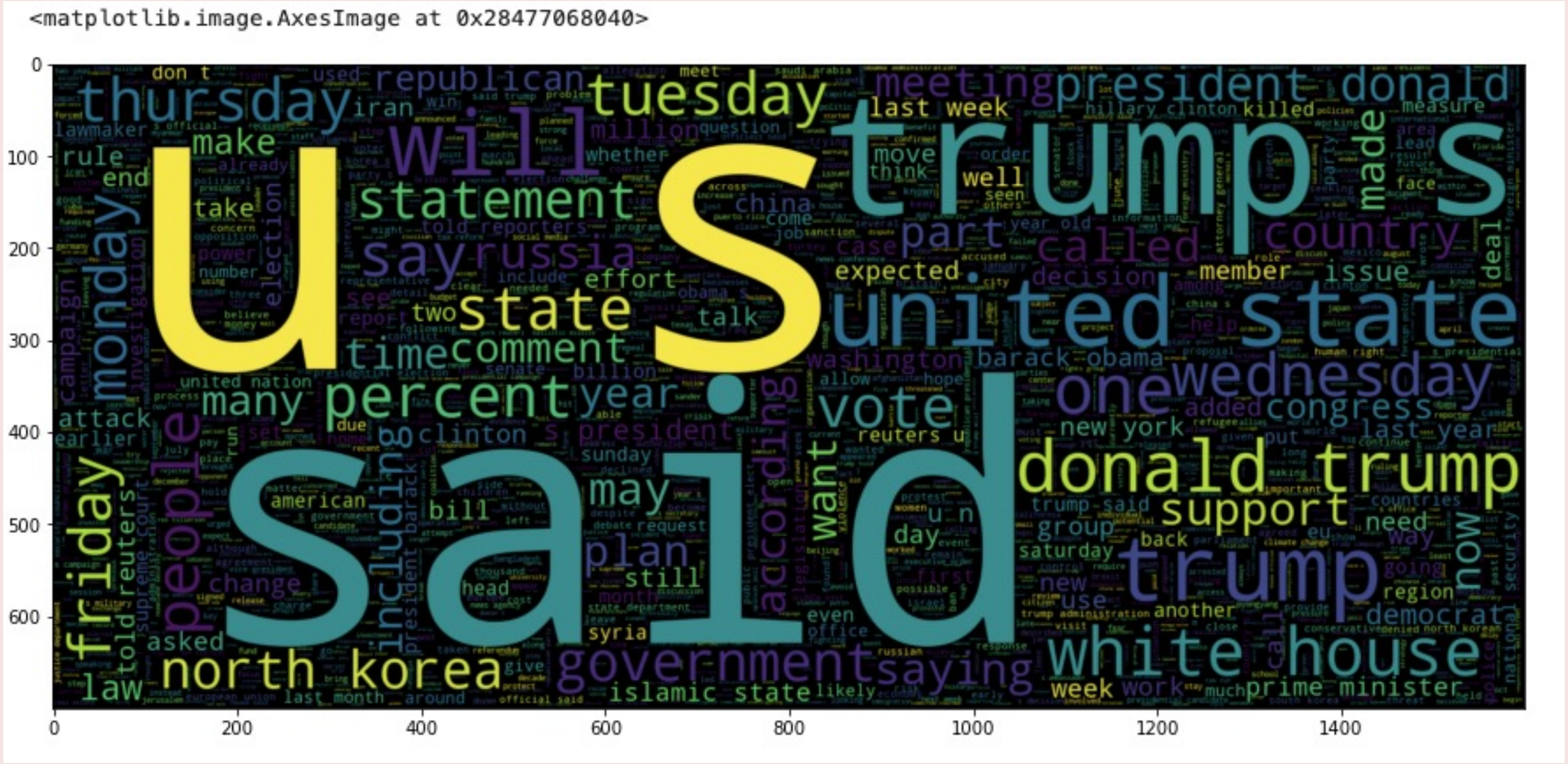
fig$_{(3)}$: Real News

<matplotlib.image.AxesImage at 0x28474209910>

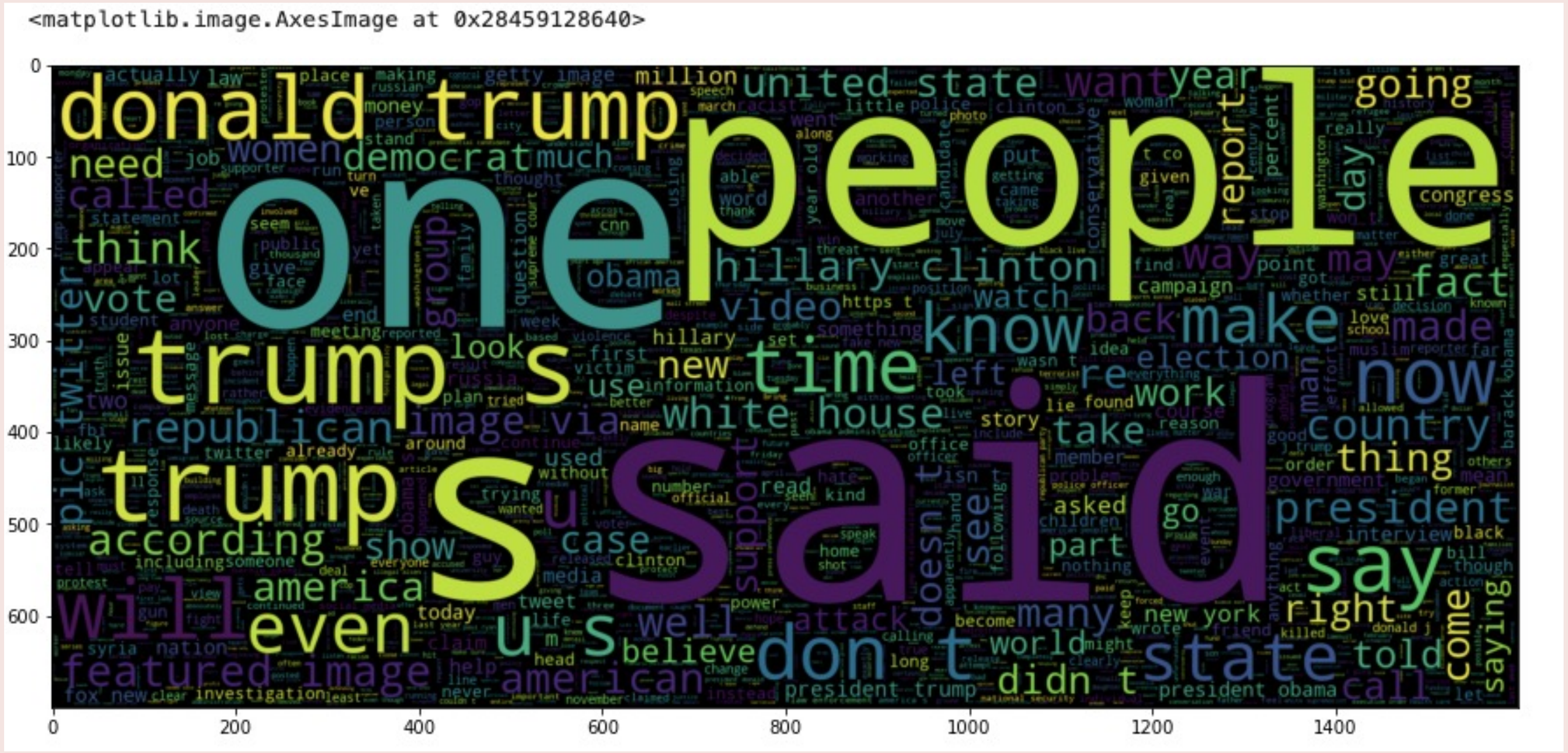


fig$_{(4)}$: Fake News

Dataset 2: ISOT Fake News Dataset

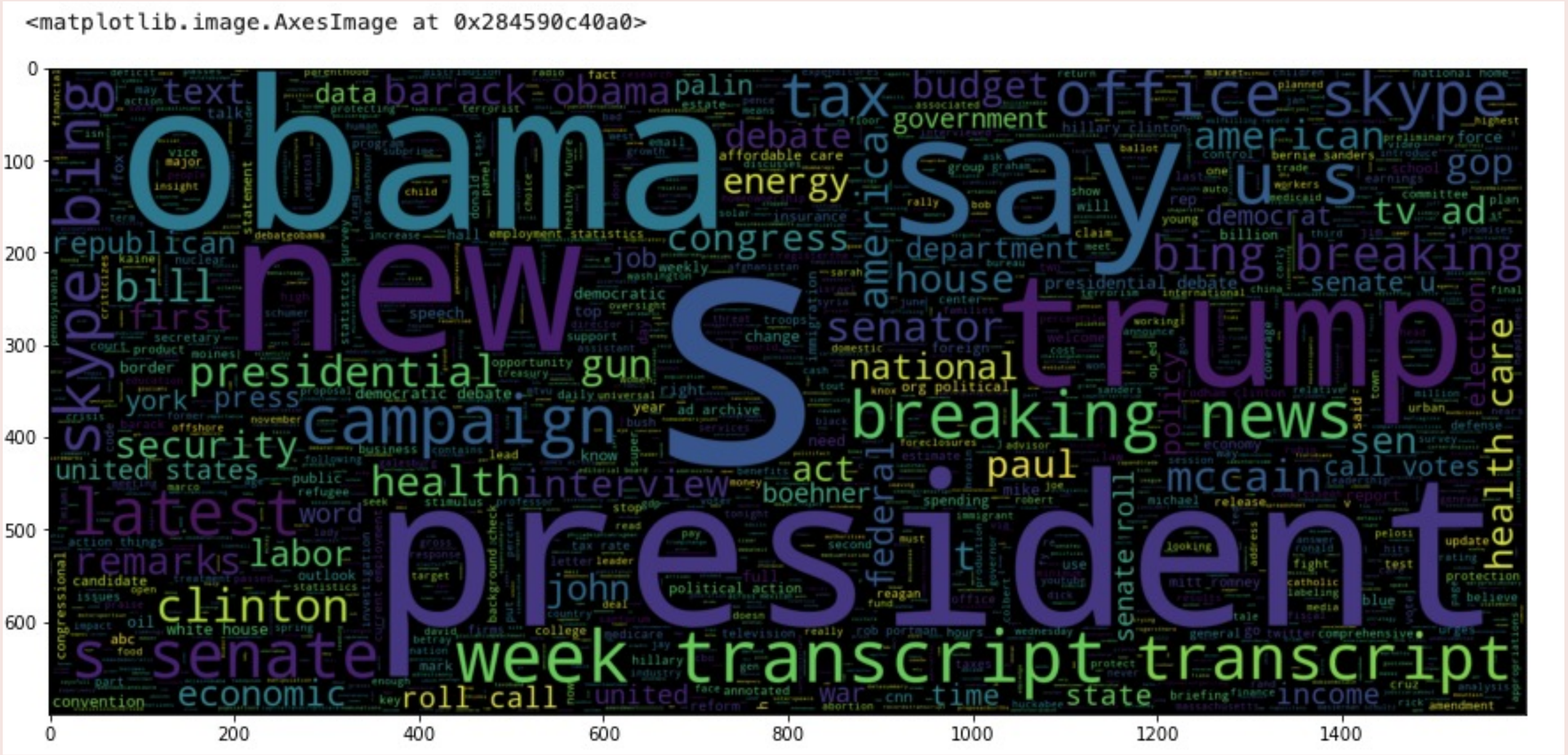


fig$_{(5)}$: Real News

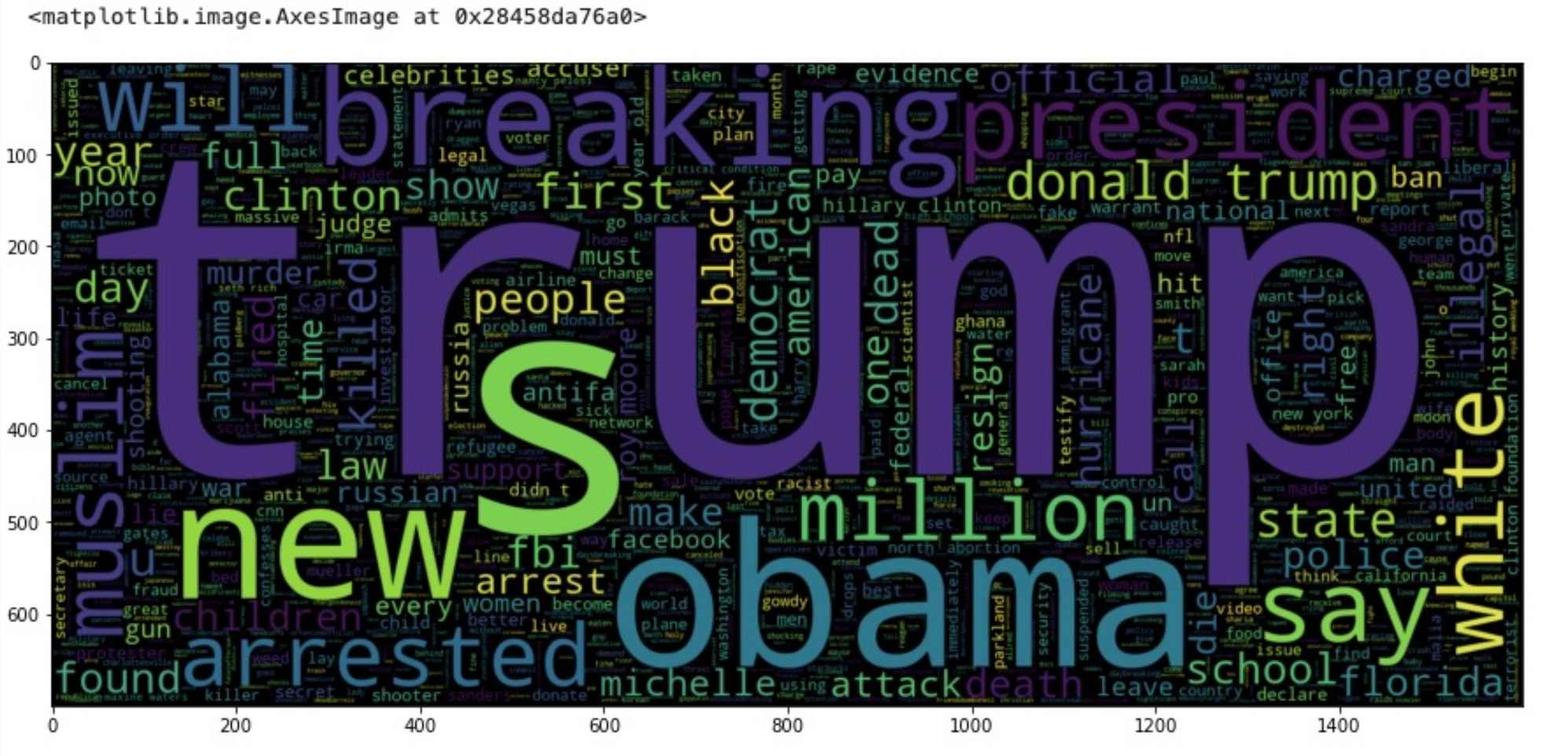


fig$_{(6)}$: Fake News

Dataset 3: BuzzFeed News Dataset

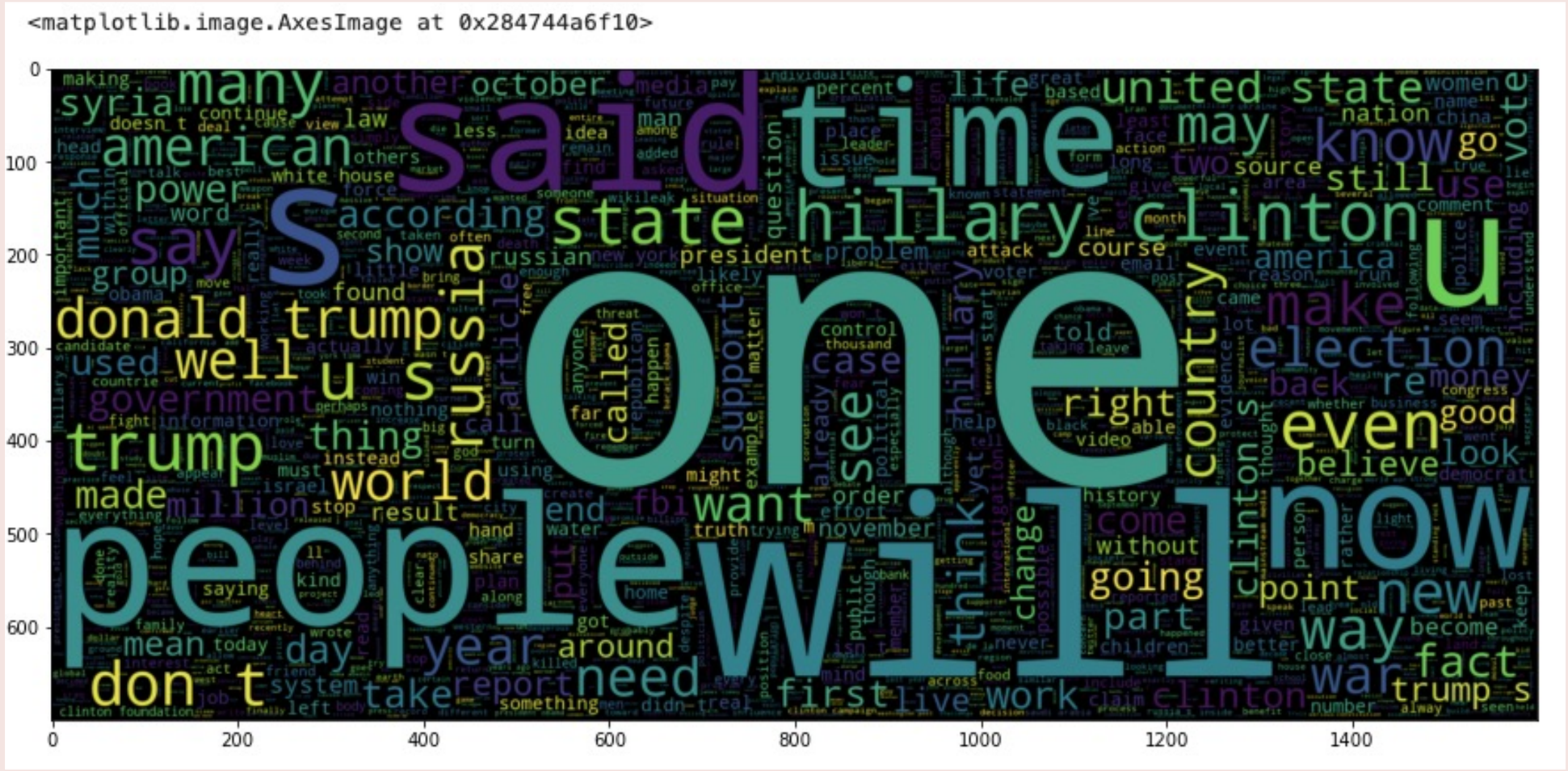


fig$_{(7)}$: Real News

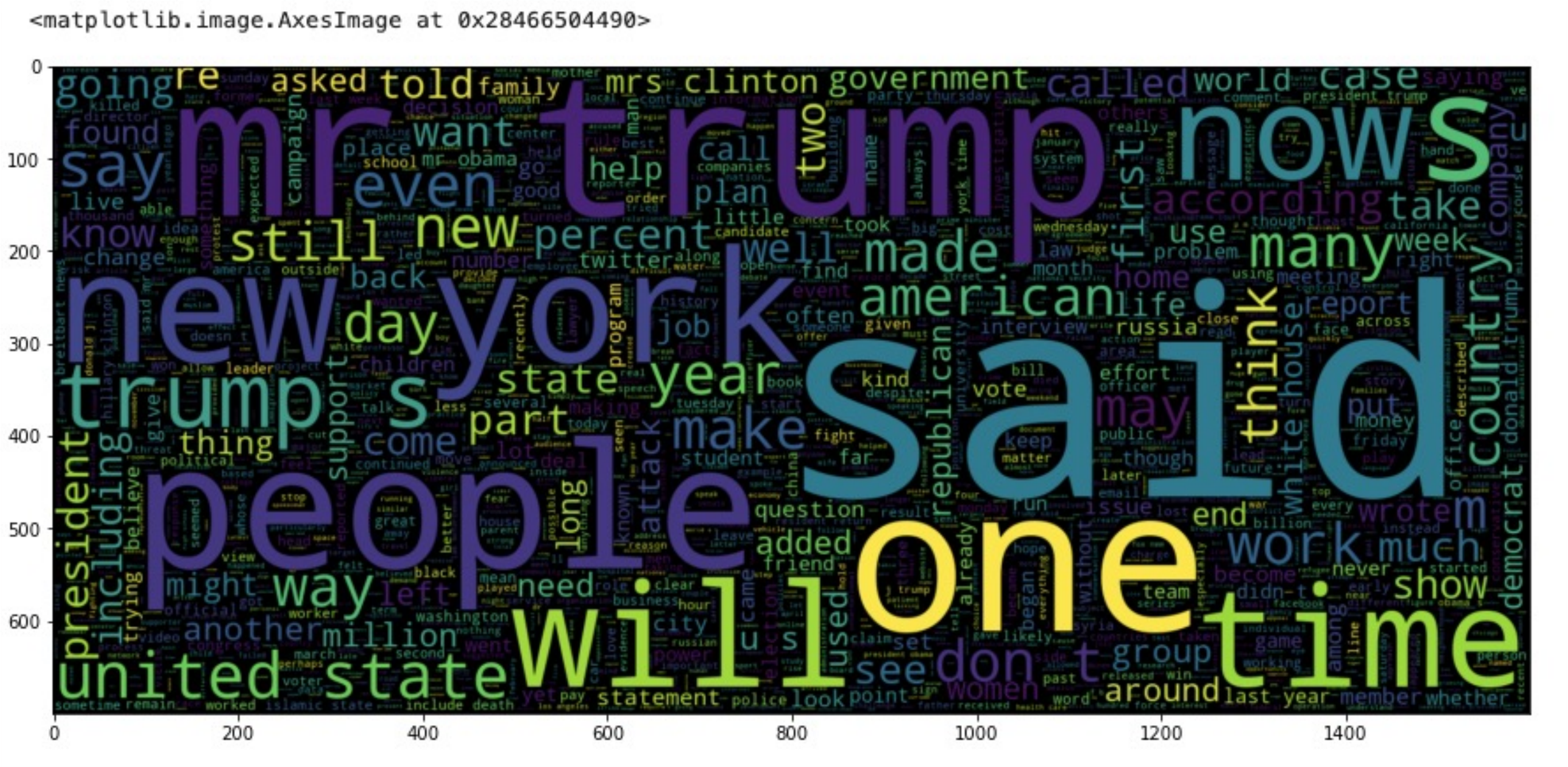


fig$_{(8)}$: Fake News

Dataset 4: LIAR Dataset

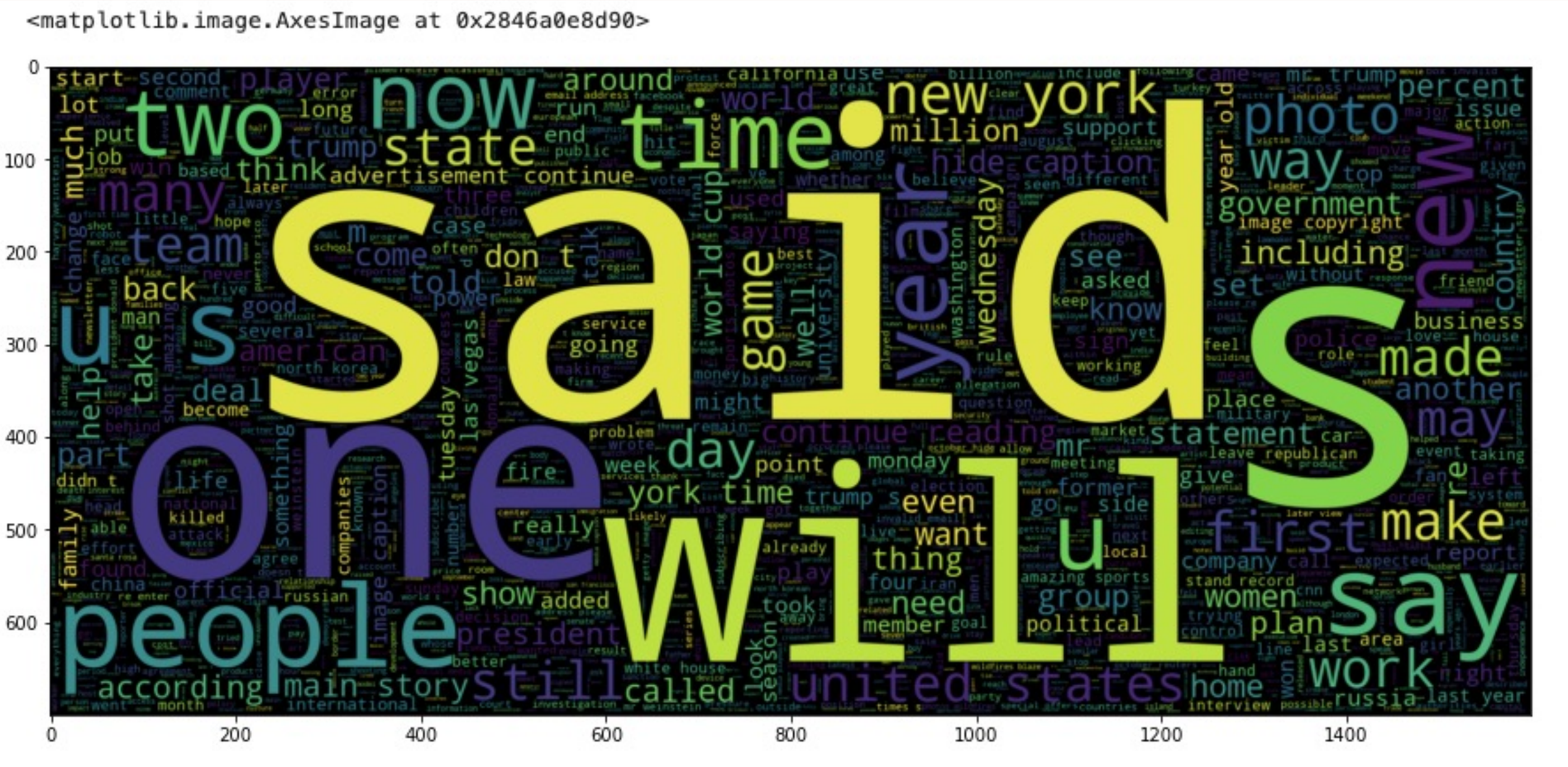


fig$_{(9)}$: Real News

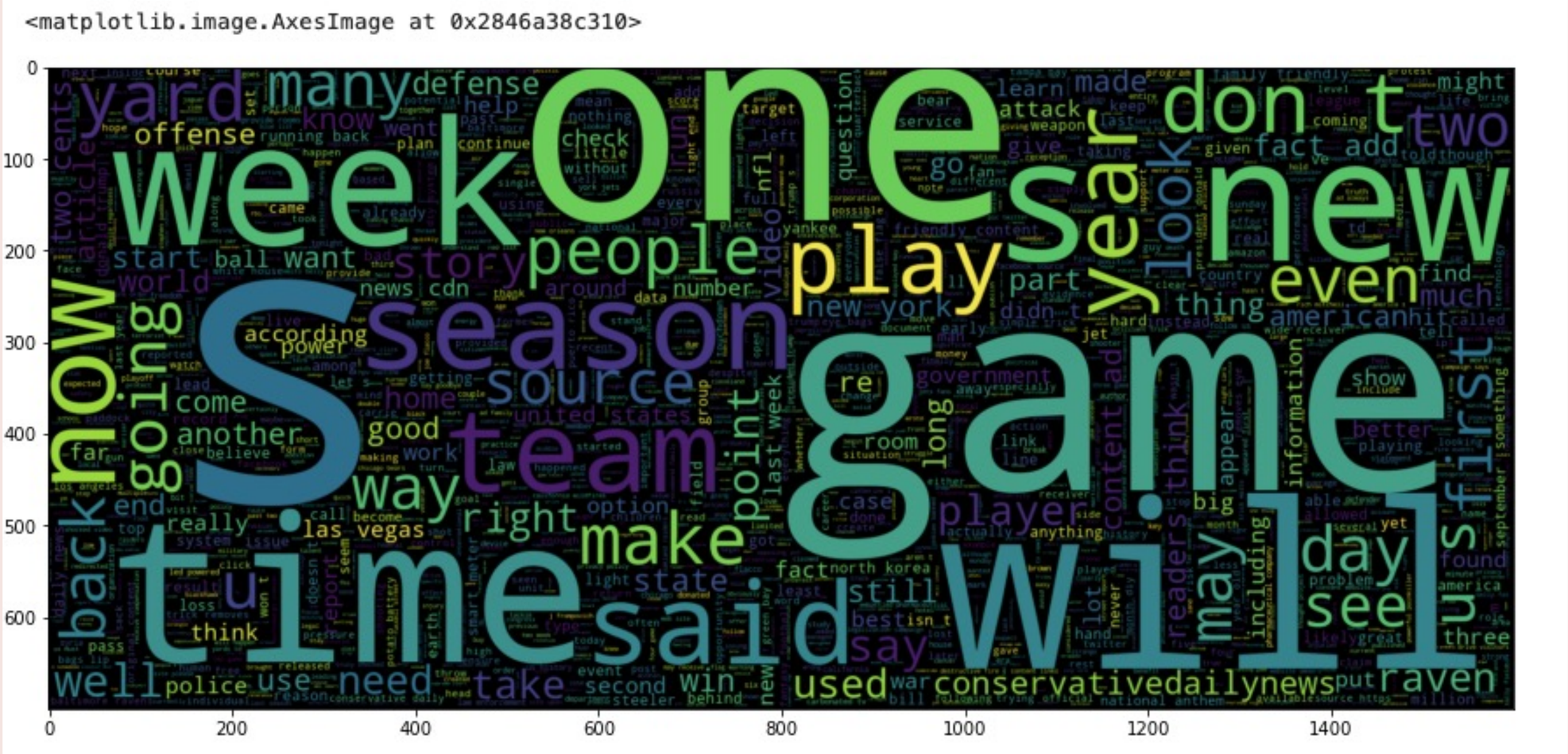


fig$_{(10)}$: Fake News

Dataset 5: Politifact News Dataset



fig$_{(11)}$: Real News



fig$_{(12)}$: Fake News

Key Findings from Visualization

- Dataset Comparisons: The distribution and common word analysis helped us understand the nature of fake and real news across different datasets.

- Word Cloud Insights: The most frequent words in fake news tend to be sensational and politically charged, while real news articles focus on factual reporting.

- Article Length: Fake news articles are often shorter and more direct, while real news tends to provide more context and details.

Train-Test Split

• Goal: Split the dataset into training and testing sets for model evaluation.

•Method:
•Used train_test_split from sklearn.model_selection.
•80% for Training, 20% for Testing to ensure robust model training while leaving sufficient data for evaluation.

•Result:
•Training set shape: x_train.shape and y_train.shape.
•Testing set shape: x_test.shape and y_test.shape.

Models

1- Logistic Regression Model

- Overview: A linear model for binary classification, ideal for distinguishing fake vs. real news

- Pipeline:
  Vectorization: CountVectorizer to convert text to vectors.
  Transformation: TfidfTransformer to weigh word importance.
  Classification: Logistic Regression (LogisticRegression()).

- Accuracy: Achieved Logisticmodel_accuracy 87.04% accuracy.

- Why Logistic Regression?
  Simple and efficient for binary classification.
  Handles high-dimensional sparse data well.

2- Decision Tree Classifier

- Overview: A decision-making model that splits data based on features, using entropy and depth constraints.

- Pipeline:

  Vectorization and TF-IDF transformation.
  Model: DecisionTreeClassifier with entropy criterion and a maximum depth of 10.

- Accuracy:

  Achieved DecisionTreemodel_accuracy 82.07% accuracy.

Pros & Cons:

Pros: Interpretable model, handles both categorical and continuous features.

Cons: Prone to overfitting without depth control.

3- Random Forest Classifier

- Overview: An ensemble of decision trees, combining predictions for better accuracy and generalization.

- Pipeline:

  Vectorization and TF-IDF transformation.
  Model: RandomForestClassifier.

- Accuracy:

  Achieved RandomForestmodel_accuracy 82.49% accuracy.

Why Random Forest?

Reduces overfitting and increases model stability.

Works well with complex datasets.

4- Stochastic Gradient Descent (SGD) Classifier

- Overview: An iterative, optimized linear model for classification.

- Pipeline:
  Vectorization and TF-IDF transformation.
  Model: SGDClassifier.

- Accuracy:
  Achieved SDGmodel_accuracy 86.23% accuracy.

Strengths:

- Efficient with large datasets.

- Suitable for high-dimensional data.

5- Gradient Boosting Classifier (GBC)

- **Overview**: A boosting method that builds weak learners (trees) sequentially, optimizing for log loss.

- **Pipeline**:

  Vectorization and TF-IDF transformation.

  Model: GradientBoostingClassifier with log loss and learning rate of 0.01.

- **Accuracy**:

  Achieved GBCmodel_accuracy 80.71% accuracy.

Why GBC?

Performs well with imbalanced data.

Can capture complex patterns in the dataset.

6- XGBoost Classifier

- Overview: An advanced gradient boosting method, optimized for speed and performance.

- Pipeline:
  Vectorization and TF-IDF transformation.
  Model: XGBClassifier with loss and learning parameters similar to GBC.

- Accuracy:
  Achieved xgboostmodel_accuracy 80.75% accuracy.

Why XGBoost?

-Regularization to avoid overfitting.

-Handles large datasets efficiently.

7- Multinomial Naive Bayes Classifier

- Overview: A probabilistic classifier based on the Bayes Theorem, assuming feature independence.

- Pipeline:

  Vectorization and TF-IDF transformation.

  Model: MultinomialNB().

- Accuracy:

  Achieved Multinomial_Naive_Bayes_accuracy 78.08% accuracy.

Why Multinomial Naive Bayes?

- Works well with text classification tasks.

- Simple and efficient for large-scale problems.

8- Bernoulli Naive Bayes Classifier

- Overview: Another variant of Naive Bayes, suited for binary/boolean feature sets.

- Pipeline:
  Vectorization and TF-IDF transformation.
  Model: BernoulliNB().

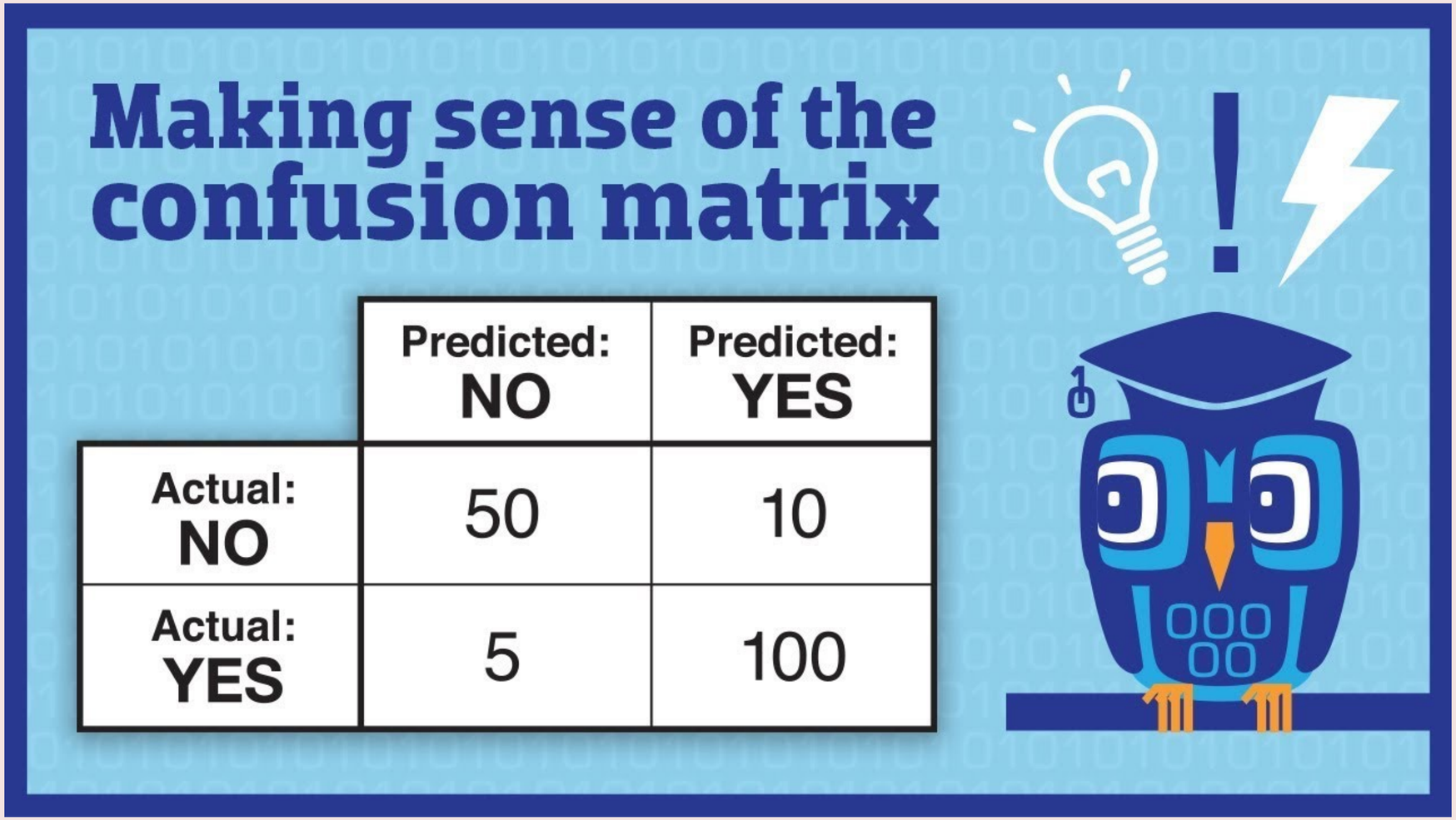
- Accuracy:
  Achieved Bernoulli_Naive_Bayes_accuracy 76.08% accuracy.

Pros & Cons:

Pros: Works well with binary features.

Cons: Assumes independent features, which may not always be true.

Confusion Matrices



Bernoulli Naive Bayes Classifier

- Purpose: To analyze the performance of the models beyond just accuracy.

- What is a Confusion Matrix?
  Shows True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

- Importance:
  Highlights where the model is making mistakes, such as labeling real news as fake (false positives).

Visualization of Model Confusion Matrices

- Visual Example:
  Display confusion matrices for the best-performing models (e.g., Logistic Regression, Random Forest).
  Highlight true positive/negative rates.
  Discuss any significant number of false positives or negatives.

- Insight:
  A good model minimizes false negatives in fake news detection.

Observations:

Fake news articles may focus on sensationalist topics, while real news focuses on factual events.

Certain keywords may be more prevalent in fake news.

Model Pipeline

- What is a Pipeline?

An automated workflow that combines several steps (vectorization, transformation, classification) in sequence

- Pipeline Steps:
  1. Count Vectorization.
  2. TF-IDF Transformation.
  3. Model Training and Prediction.

- Why Use Pipelines?

- Simplifies the workflow.
- Ensures consistency and repeatability

Joblib Model Saving

- Why Save the Model?
•To reuse the trained model without retraining.
•Useful for deployment and further analysis.

Method Used:
•joblib.dump(Logisticmodel, 'model.pkl’).

How it Works:
•Saves the Logistic Regression model to a file for future use.

Conclusion

- Multiple models were tested to detect fake news.

- Logistic Regression, Random Forest, and XGBoost showed the highest accuracies.

- Naive Bayes models were fast but slightly less accurate.

Key Takeaways:

Models like Random Forest and XGBoost provide robust performance.

Fake news detection is a challenging yet solvable task using machine learning.

Model Deployment

1. Flask Setup

The code begins by importing necessary libraries like Flask, joblib (for loading the model), re (for regular expressions), pandas, and os.

Flask(_name_): Initializes a new Flask web app instance

2. Model Loading

The path to the pre-trained model (Model.pkl) is dynamically constructed using os.

path to ensure compatibility across different file structures.The model is loaded using joblib for efficient deserialization.

3. Routes

@app.route('/') (GET method):Loads the home page (index.html) where users can input news text.

@app.route('/', methods=['POST']) (POST method):Handles form submissions, processes the user input, and predicts whether the news is fake or real.

4. Text Preprocessing Function (wordpre)

Purpose: Cleans and normalizes the input text before feeding it into the model.

Steps:

- Converts text to lowercase.
- Removes special characters, numbers, URLs, and HTML tags.
- Strips punctuation and unnecessary whitespace.

5. Prediction:

- The cleaned input is converted into a Pandas Series for compatibility with the machine learning model.
- Model Prediction: The pre-trained model predicts the class (fake/real) based on the input text.
- Error Handling: Any issues during prediction are caught and displayed as an error message.

6.Result Display

The application is run in debug mode (app.run(debug=True)), providing real-time feedback during development.

Thank you

If you have any questions,
please don’t hesitate to ask!