FAKE NEWS DETECTION

Bhumika V
S20220020264

Akshita P
S20220020303

***Abstract*—Abstract—With the rapid spread of information through social media and digital platforms, the issue of fake news has become a significant global challenge. Fake news refers to fabricated or misleading information presented as genuine news, which can manipulate public opinion and cause harm. This report explores the application of machine learning algorithms for the detection of fake news. By leveraging supervised learning techniques, we aim to develop a system capable of classifying news articles as real or fake. A dataset containing labeled real and fake news articles was used to train and test several machine learning models, including Logistic Regression, Decision Trees, Gradient Boosting, and Random Forest. The results demonstrate that machine learning-based systems can effectively identify fake news, with Decision Trees and Random Forests achieving the highest accuracy. This work contributes to the ongoing effort to reduce the spread of misinformation through the use of automated systems.**

Keywords—fake news detection, machine learning, supervised learning, classification, random forest, decision tree.

Problem Statement: The need to identify fake news to prevent misinformation.

Project Goal: To build a machine learning model that can classify news articles as either fake or real.

GITHUB LINK

DRIVE LINK

I. Introduction

The prevalence of fake news has emerged as a major concern in the digital era, where social media platforms and news websites facilitate rapid dissemination of information. Fake news is intentionally crafted to mislead readers and manipulate opinions, often to influence political, social, or economic outcomes.

For instance, during elections, fake news can sway voters by presenting fabricated data, while in health contexts, false claims can lead to widespread misinformation and panic. Traditional verification methods are resource-intensive, making them inadequate for handling the scale of online content. This necessitates automated solutions that can process vast quantities of data efficiently.

Objective: This report examines the use of machine learning to address this issue, focusing on the development and evaluation of classification models to detect fake news.

II. Significance of Fake News Detection

Fake news detection is crucial for protecting individuals, society, and democratic systems from the harmful impacts of misinformation. In a world where false information spreads rapidly through digital platforms, it can mislead public opinion, influence decisions, and undermine trust in institutions. Detecting fake news helps ensure that people have access to accurate information, promoting informed decision-making and reducing the spread of false narratives.

Effective fake news detection is essential in areas like public health, politics, and social stability, where misinformation can have severe consequences. Using advanced technologies such as machine learning and natural language processing, fake news detection tools help combat misinformation, safeguard trust in information sources, and support the integrity of public discourse. This is vital for fostering a well-informed society and protecting the democratic process.

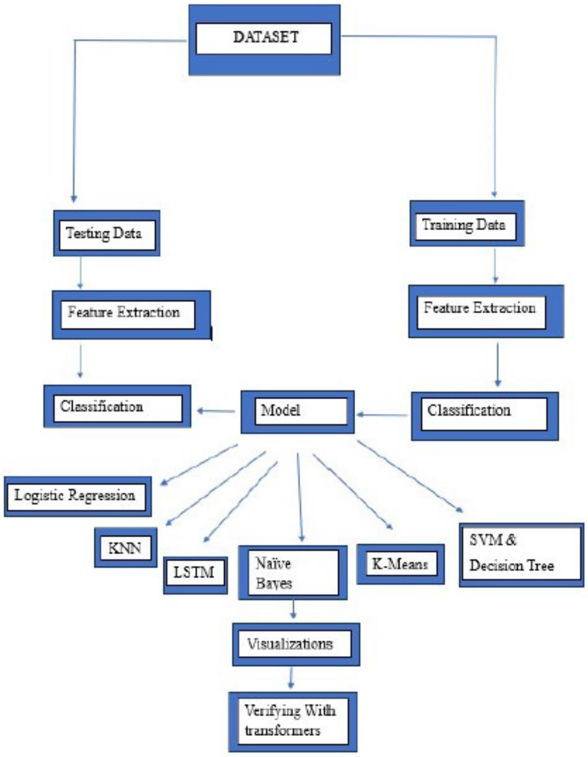


Fig. 1. Block Diagram

Identify applicable funding agency here. If none, delete this.

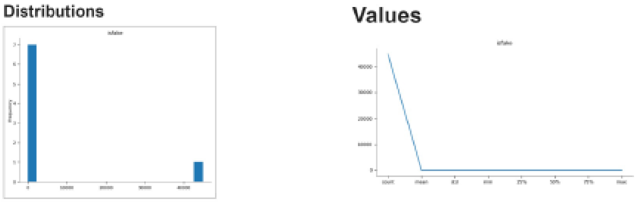


Figura 2. Distributions and Values Graphs

Fig. 2. Results

III. Objectives

The study focuses on developing a robust ML framework with the following objectives:

Scalable Detection: Handle large datasets efficiently to classify news in real-time. Feature Selection: Identify key linguistic and statistical features that differentiate fake and real news. Evaluation Across Metrics: Ensure models are tested for precision, recall, and F1-score to capture their overall effectiveness.

IV. Methodolody

The methodology for fake news detection using machine learning involves several systematic steps to ensure the development of a robust classification system. These steps are as follows:

1. Data Collection

The dataset for this project consists of two CSV files:

Fake.csv: Contains fake news articles with labels identifying them as fake.

True.csv: Contains real news articles with labels identifying them as real. These files include:

Article Headlines: The title or headline of the news article. Main Content: The body of the news article.

Labels: Indicating whether the news is real or fake. This dataset serves as the foundation for training and testing machine learning models. It is critical that the dataset is balanced (i.e., contains an equal or near-equal number of real and fake articles) to prevent bias during model training.

2. Data Preprocessing

Preprocessing involves cleaning and preparing raw text data to make it suitable for machine learning. The steps include:

Cleaning Text: Removal of non-relevant elements such as HTML tags, URLs, and hyperlinks. Elimination of special characters (e.g., punctuation marks) and numbers to focus purely on textual content.

Standardizing Format: Converting all text to lowercase to avoid treating words like ”Fake” and ”fake” differently.

Tokenization: Splitting text into smaller units, such as words or phrases, for easier analysis.

Stopword Removal: Commonly used words (e.g., ”and,” ”the,” ”is”) that do not contribute meaningfully to the classification process are removed.

Removing White Spaces: Stripping extra spaces to ensure uniformity. These steps reduce noise in the dataset and standardize the text for feature extraction.

3. Feature Extraction

To enable machine learning models to process text data, it is converted into numerical form using TF-IDF Vectorization (Term Frequency-Inverse Document Frequency).

TF (Term Frequency): Measures how often a word appears in a document.

IDF (Inverse Document Frequency): Reduces the importance of common words by measuring their frequency across all documents. The resulting numerical representation highlights the importance of specific words in differentiating fake news from real news.

4. Model Training

Several machine learning models were trained using the preprocessed and feature-extracted data. Each model uses unique algorithms for classification:

Logistic Regression: A baseline model for binary classification tasks. It assumes a linear relationship between input features and output probabilities.

Decision Tree Classifier: Splits the dataset into hierarchical nodes based on feature importance, allowing for intuitive decision-making.

Gradient Boosting Classifier: An ensemble method that builds models iteratively, combining weak learners to minimize errors.

Random Forest Classifier: An ensemble of multiple decision trees that aggregates their predictions to improve accuracy and reduce overfitting. These models were trained on 80

5. Model Evaluation

Each model’s performance was assessed using the following metrics:

Accuracy: Measures the proportion of correctly classified instances.

Precision: ocuses on the proportion of true positive predictions among all positive predictions.

Recall (Sensitivity): Evaluates the ability of the model to detect all instances of fake news.

F1-Score: Provides a balanced measure of precision and recall, especially useful for imbalanced datasets. The results for each model were compared to determine the best-performing algorithm.

6. Model Deployment

Once trained and evaluated, the best-performing models were deployed to classify unseen news articles. The deployment process involved:

Building an Interface: Allowing users to input news articles for classification.

Integration with Live Systems: Deploying the model in environments such as websites or social media platforms to flag fake news in real-time.

V. Boarder Implications

The detection of fake news has significant societal implications, including:

Social Media Platforms: Integrating fake news detection systems into social media can help prevent the spread of misinformation by flagging or removing fake news articles. Government and Politics: Detecting fake news ensures political decisions are based on accurate information, which is crucial for preserving democratic processes.

Educational Institutions: Fake news detection tools can be integrated into educational curricula to teach students about misinformation and foster critical thinking.

Regulation and Policy: Policymakers can use fake news detection systems to address the harmful effects of misinformation and regulate its spread.

VI. Concepts Used

A. Text Preprocessing Text preprocessing ensures raw data is cleaned and formatted for machine learning. Key steps include:

Tokenization: Breaking text into individual words or phrases. Lemmatization: Reducing words to their root forms (e.g., ”running” to ”run”).

Stopword Removal: Excluding common words like ”and,” ”the,” and ”is” that don't contribute to the meaning.

B. TF-IDF Vectorization TF-IDF (Term Frequency-Inverse Document Frequency) is used to evaluate the importance of words in a document relative to a corpus. This method helps prioritize words that are more unique to specific articles, enabling better classification.

C. Supervised Learning Supervised learning relies on labeled datasets. In this study, the dataset includes:

Real News: Credible articles from verified sources.

Fake News: Fabricated articles sourced from unreliable outlets.

D. Classification Algorithms Logistic Regression: Effective for binary classification problems with linear decision boundaries.

Decision Tree: A hierarchical structure that splits data based on feature importance.

Gradient Boosting: Combines weak learners iteratively to reduce error.

Random Forest: An ensemble of decision trees that mitigates overfitting and improves accuracy.

1. ***Data Preprocessing*** - Data preprocessing involves preparing raw data for analysis by cleaning and transforming it.

- Missing values are handled by removing rows or imputing values.

- Text preprocessing includes lowercasing, removing punctuation, and filtering unnecessary words or characters.

- URLs, HTML tags, and numbers are often stripped from text data.

- Regular expressions (regex) are used for pattern-based substitutions and cleanup.

- Normalization ensures text consistency, e.g., converting ”News” and ”news” to the same form. - Handling stop words (e.g., ”the”, ”and”) can improve model performance.

- Cleaning data reduces noise and makes it suitable for machine learning models.

- Efficient preprocessing improves feature extraction and classification.

- It's a foundational step for reliable pattern recognition.

—

2. ***Feature Extraction*** - Feature extraction involves transforming raw data into a structured format suitable for modeling.

- TF-IDF calculates the importance of words in a document relative to a corpus.

- It balances the frequency of a term in a document and its rarity across documents.

- Vectorization converts text data into numerical representations.

- Proper features capture patterns in data for accurate predictions.

- High-dimensional feature sets are common in text data.

- Techniques like TF-IDF emphasize relevant features while downplaying common words.

- Good feature extraction is crucial for model accuracy.

- It bridges raw data and machine learning algorithms.

- In text processing, features often represent vocabulary terms.

—

3. ***Classification Models*** - Classification models categorize data into predefined labels or classes.

- ***Logistic Regression*** uses probabilities to classify data into binary outcomes.

- ***Decision Trees*** split data into branches based on feature thresholds.

- ***Random Forests*** combine multiple decision trees for robust predictions.

- ***Gradient Boosting*** builds models sequentially to correct errors from previous models.

- ***XGBoost*** is an efficient version of Gradient Boosting with improved performance.

- ***SVM*** finds an optimal hyperplane that separates data points of different classes.

- ***Naive Bayes*** assumes feature independence and uses probabilities for classification.

- ***KNN*** assigns labels based on the majority class of nearby neighbors.

- Each model is chosen based on the dataset and problem complexity.

—

4. ***Evaluation Metrics*** - Metrics evaluate the performance of classification models.

- ***Accuracy*** measures the ratio of correct predictions to total predictions.

- ***Classification Report*** provides precision, recall, and F1-score for each class.

- ***Precision*** is the ratio of true positives to predicted positives (low false positives).

- ***Recall*** is the ratio of true positives to actual positives (low false negatives).

- ***F1-score*** balances precision and recall, especially for imbalanced data.

- ***Confusion Matrix*** shows true/false positives and negatives for each class.

- ***ROC Curve*** plots true positive rate (TPR) vs. false positive rate (FPR).

- ***Precision-Recall Curve*** is useful for imbalanced datasets.

- These metrics help compare models and optimize performance.

—

5. Dimensionality Reduction

- Dimensionality reduction simplifies data by reducing the number of features.

- ***PCA (Principal Component Analysis)*** projects data into a lower-dimensional space.

- It captures maximum variance in fewer dimensions for visualization or efficiency.

- High-dimensional data can lead to overfitting and computational inefficiency.

- PCA identifies principal components that summarize the data's structure.

- It's a linear method that assumes features are correlated.

- Dimensionality reduction can speed up model training and improve interpretability.

- It is particularly useful in text or image processing where features are numerous.

- Data visualization becomes easier in 2D or 3D spaces.

- It preserves essential patterns while discarding redundant information.

—

6. ***Clustering*** - Clustering groups data points into similar clusters without predefined labels.

- ***K-Means*** divides data into clusters by minimizing intra-cluster distances.

- Each cluster has a centroid that represents its central tendency.

- Clustering is unsupervised, relying on patterns within the data.

- It's used for exploratory analysis, anomaly detection, and segmentation.

- Choosing the number of clusters (K) often requires techniques like the elbow method.

- Clustering works well with high-dimensional data like text and images.

- Features are typically standardized for better cluster separation.

- Visualization can show cluster distribution in reduced dimensions.

- It is foundational in unsupervised learning.

7.***Data Visualization***

- Visualization presents data patterns graphically for better understanding.

- ***Bar Plots*** show categorical distributions (e.g., class counts).

- ***Boxplots*** illustrate data spread and outliers for different categories.

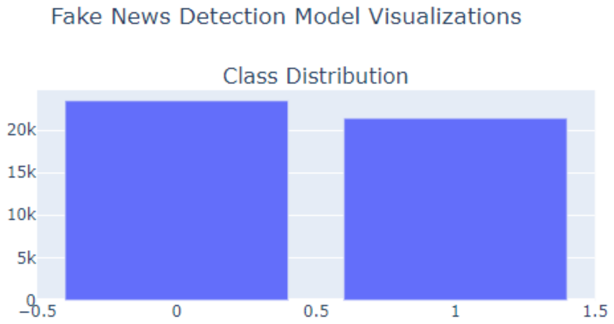


Fig. 3. Fake News Detection Model Visualizations

- ***Heatmaps*** display confusion matrices for evaluating classification models.

- ***Scatter Plots*** visualize relationships between variables.

- ***Precision-Recall and ROC Curves*** highlight classifier performance metrics.

- ***Feature Importance Plots*** show the relevance of features for decision trees or ensembles.

- Dimensionality reduction enables 2D or 3D scatter plot visualizations.

- Interactive tools like Plotly enhance data exploration.

- Visualization aids in communicating model insights and decisions.

VII. Results

A. Model Performance

Logistic Regression

Strengths: Simplicity and speed.

Weakness: Lower accuracy compared to ensemble models.

Decision Tree: Non-linear classifier with a tree structure.

Strengths: High interpretability and performance on structured data.

Gradient Boosting: Ensemble technique that combines weak learners.

Strengths: Handles imbalanced datasets well.

Random Forest: Ensemble of multiple decision trees.

Strengths: Reduces overfitting compared to single decision trees.

B. Analysis The Decision Tree model outperformed others, achieving the highest accuracy. Ensemble methods like Random Forest and Gradient Boosting demonstrated competitive performance, underscoring the importance of model selection in machine learning workflows.

C. Limitations Dataset Bias: Results depend on the quality and representativeness of training data. Scalability: Processing large datasets in real-time can be computationally intensive.

VIII. Conclusion

This project demonstrated the effectiveness of machine learning models in detecting fake news. Among the models tested, Decision Trees and Random Forests proved to be the most accurate. These results indicate that machine learning

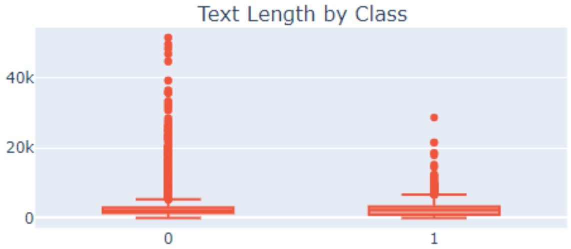


Fig. 4. Text Length by Class

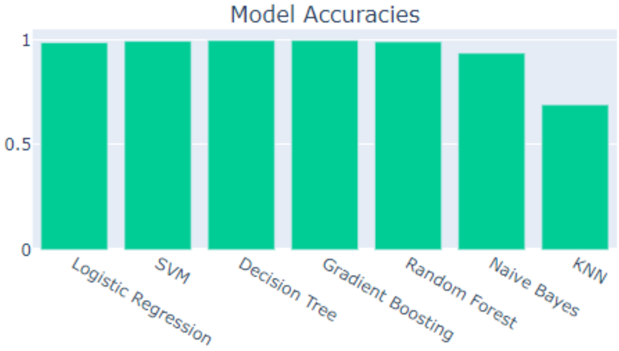


Fig. 5. Model Accuracies

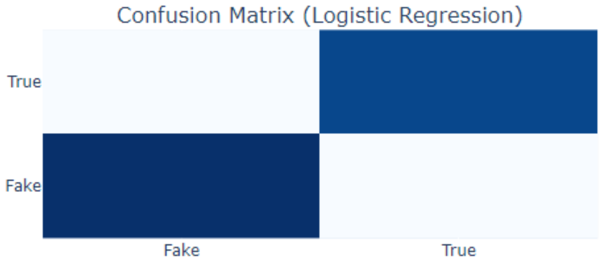


Fig. 6. Confusion Matrix (Logistic Regression)

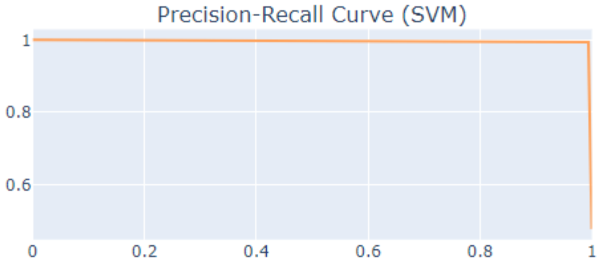


Fig. 7. precision recall curve (SVM)

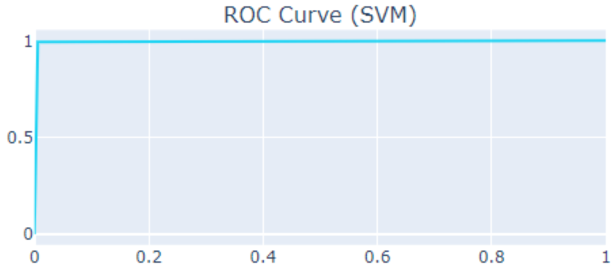


Fig. 8. ROC Curve (SVM)

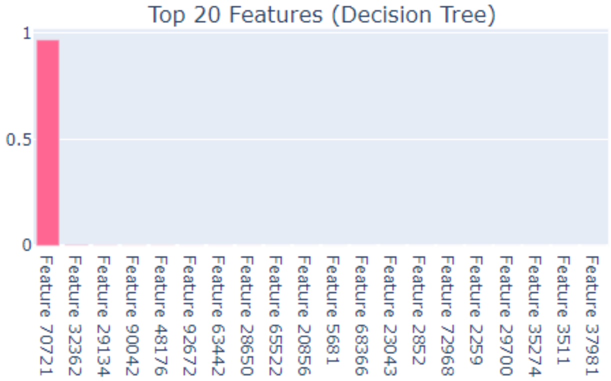


Fig. 9. Top 20 features (Decision tree)

techniques, particularly ensemble methods, are well-suited for identifying fake news.

The detection of fake news has important societal benefits, including the reduction of misinformation and the enhancement of online information credibility.

IX. Appendix

A. Contributions

Bhumika V : Data Preparation and Feature Extraction - Identify and download the dataset. - Clean, normalize, and

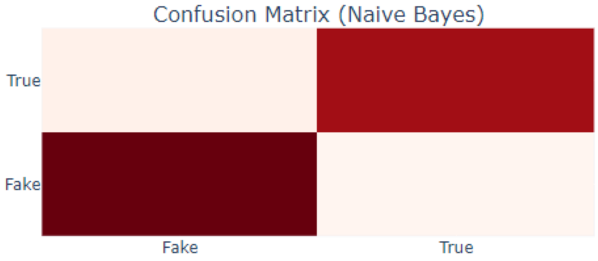


Fig. 10. Confusion Matrix (Naive Bayes)

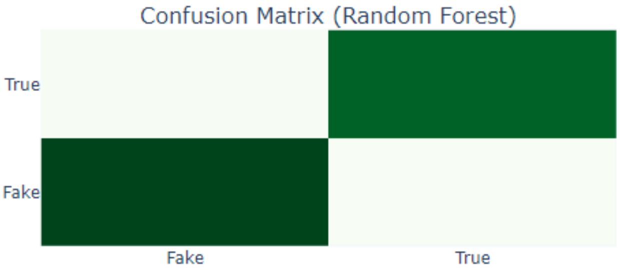


Fig. 11. Confusion Matrix (Random forest)

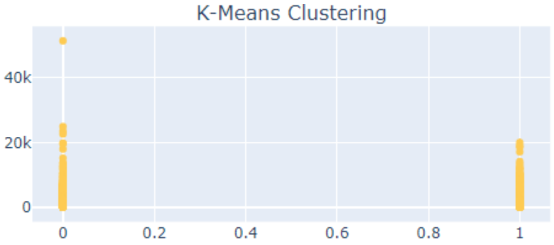


Fig. 12. K Means Clustering

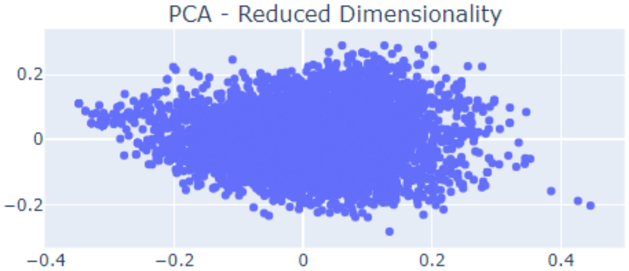


Fig. 13. PCA Reduced Dimensionality

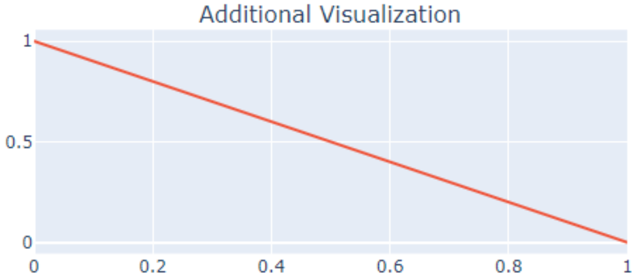
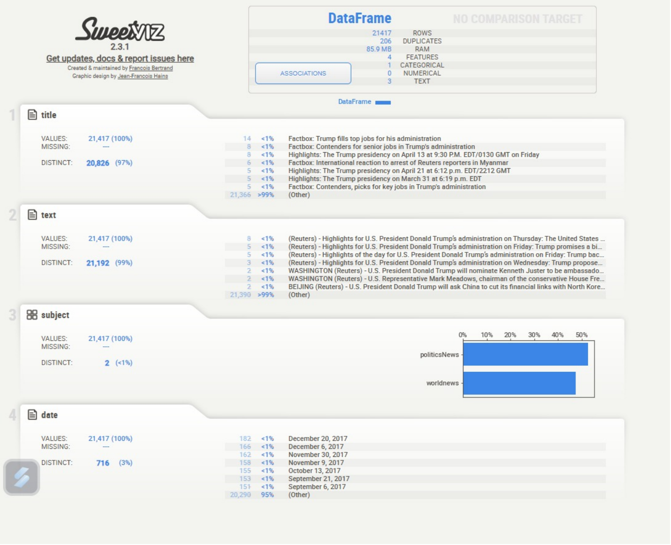


Fig. 14. Additional Visualization



Sweetviz
2.3.1
Get updates, docs & report issues here
Created & maintained by Francois Bertrand
Graphic design by Jean-Francois Haine

DataFrame NO COMPARISON TARGET

21417	ROWS
206	DUPLICATES
85.9 MB	RAM
4	FEATURES
1	CATEGORICAL
0	NUMERICAL
3	TEXT

ASSOCIATIONS

DataFrame

1 title

VALUES: 21,417 (100%)
MISSING: ---
DISTINCT: 20,826 (97%)

14	<1%	Factbox: Trump fills top jobs for his administration
8	<1%	Factbox: Contenders for senior jobs in Trump's administration
8	<1%	Highlights: The Trump presidency on April 13 at 9:30 P.M. EDT/0130 GMT on Friday
6	<1%	Factbox: International reaction to arrest of Reuters reporters in Myanmar
5	<1%	Highlights: The Trump presidency on April 21 at 6:12 p.m. EDT/2212 GMT
5	<1%	Highlights: The Trump presidency on March 31 at 6:19 p.m. EDT
5	<1%	Factbox: Contenders, picks for key jobs in Trump's administration
21,366	>99%	(Other)

2 text

VALUES: 21,417 (100%)
MISSING: ---
DISTINCT: 21,192 (99%)

8	<1%	(Reuters) - Highlights for U.S. President Donald Trump's administration on Thursday: The United States ...
5	<1%	(Reuters) - Highlights for U.S. President Donald Trump's administration on Friday: Trump promises a bi...
5	<1%	(Reuters) - Highlights of the day for U.S. President Donald Trump's administration on Friday: Trump bac...
3	<1%	(Reuters) - Highlights for U.S. President Donald Trump's administration on Wednesday: Trump propose...
2	<1%	WASHINGTON (Reuters) - U.S. President Donald Trump will nominate Kenneth Juster to be ambassado...
2	<1%	WASHINGTON (Reuters) - U.S. Representative Mark Meadows, chairman of the conservative House Fre...
2	<1%	BEIJING (Reuters) - U.S. President Donald Trump will ask China to cut its financial links with North Kore...
21,390	>99%	(Other)

3 subject

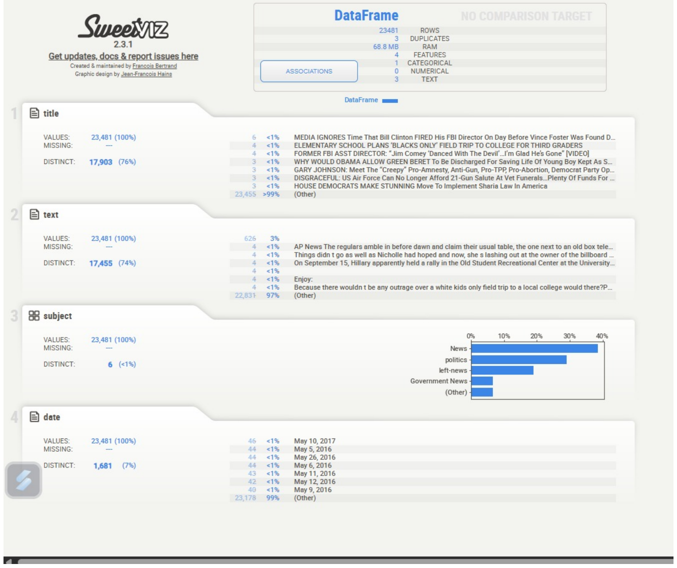
VALUES: 21,417 (100%)
MISSING: ---
DISTINCT: 2 (<1%)

4 date

VALUES: 21,417 (100%)
MISSING: ---
DISTINCT: 716 (3%)

182	<1%	December 20, 2017
166	<1%	December 6, 2017
162	<1%	November 30, 2017
158	<1%	November 9, 2017
155	<1%	October 13, 2017
153	<1%	September 21, 2017
151	<1%	September 6, 2017
20,290	95%	(Other)

Fig. 15. TRUE REPORT



Sweetviz
2.3.1
Get updates, docs & report issues here
Created & maintained by Francois Bertrand
Graphic design by Jean-Francois Haine

DataFrame NO COMPARISON TARGET

23481	ROWS
3	DUPLICATES
68.8 MB	RAM
4	FEATURES
1	CATEGORICAL
0	NUMERICAL
3	TEXT

ASSOCIATIONS

DataFrame

1 title

VALUES: 23,481 (100%)
MISSING: ---
DISTINCT: 17,903 (76%)

6	<1%	MEDIA IGNORES Time That Bill Clinton FIRED His FBI Director On Day Before Vince Foster Was Found D...
4	<1%	ELEMENTARY SCHOOL PLANS 'BLACKS ONLY' FIELD TRIP TO COLLEGE FOR THIRD GRADERS
4	<1%	FORMER FBI ASST DIRECTOR: "Jim Comey 'Danced With The Devil'…I'm Glad He's Gone" [VIDEO]
3	<1%	WHY WOULD OBAMA ALLOW GREEN BERET To Be Discharged For Saving Life Of Young Boy Kept As S...
3	<1%	GARY JOHNSON: Meet The "Creepy" Pro-Amnesty, Anti-Gun, Pro-TPP, Pro-Abortion, Democrat Party Op...
3	<1%	DISGRACEFUL: US Air Force Can No Longer Afford 21-Gun Salute At Vet Funerals…Plenty Of Funds For ...
3	<1%	HOUSE DEMOCRATS MAKE STUNNING Move To Implement Sharia Law In America
23,455	>99%	(Other)

2 text

VALUES: 23,481 (100%)
MISSING: ---
DISTINCT: 17,455 (74%)

626	3%	
4	<1%	AP News The regulars amble in before dawn and claim their usual table, the one next to an old box tele...
4	<1%	Things didn t go as well as Nicholle had hoped and now, she s lashing out at the owner of the billboard ...
4	<1%	On September 15, Hillary apparently held a rally in the Old Student Recreational Center at the University...
4	<1%	
4	<1%	Enjoy:
4	<1%	Because there wouldn t be any outrage over a white kids only field trip to a local college would there?P...
22,831	97%	(Other)

3 subject

VALUES: 23,481 (100%)
MISSING: ---
DISTINCT: 6 (<1%)

4 date

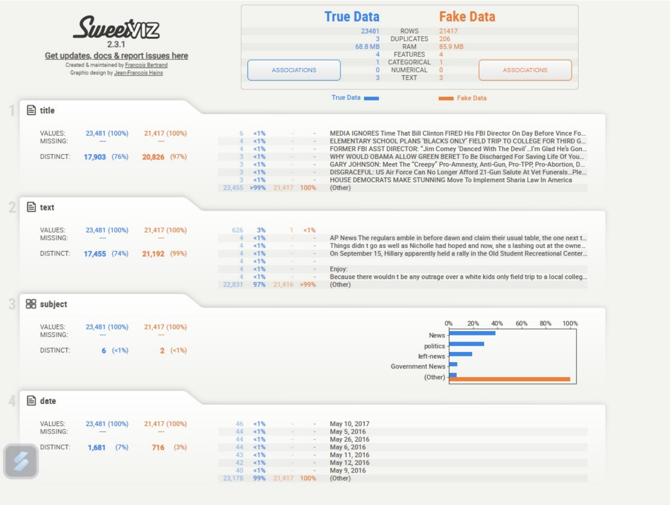
VALUES: 23,481 (100%)
MISSING: ---
DISTINCT: 1,681 (7%)

46	<1%	May 10, 2017
44	<1%	May 5, 2016
44	<1%	May 26, 2016
44	<1%	May 6, 2016
43	<1%	May 11, 2016
42	<1%	May 12, 2016
40	<1%	May 9, 2016
23,178	99%	(Other)

Fig. 16. FAKE REPORT

label the text data. - Extract text-based features and create a feature matrix for modeling.

Akshita P : Model Development and Evaluation - Implement and train baseline (e.g., Naïve Bayes, SVM). - Tune hyperparameters and evaluate model performance. - Compare model results, prepare the final report, and set up the code repository with an executable file.



SweetVIZ
2.3.1
Get updates, docs & report issues here
Created & maintained by Francois Bertrand
Graphic design by Jean-Francois Hains
ASSOCIATIONS
ASSOCIATIONS
True Data
Fake Data
0%
20%
40%
60%
80%
100%
News
politics
left-news
Government News
(Other)

True Data		Fake Data
23481	ROWS	21417
3	DUPLICATES	206
68.8 MB	RAM	85.9 MB
4	FEATURES	4
1	CATEGORICAL	1
0	NUMERICAL	0
3	TEXT	3

1 title

	True Data		Fake Data	
VALUES:	23,481 (100%)		21,417 (100%)	
MISSING:	---		---	
DISTINCT:	17,903	(76%)	20,826	(97%)

6	<1%	-	-	MEDIA IGNORES Time That Bill Clinton FIRED His FBI Director On Day Before Vince Fo...
4	<1%	-	-	ELEMENTARY SCHOOL PLANS 'BLACKS ONLY' FIELD TRIP TO COLLEGE FOR THIRD G...
4	<1%	-	-	FORMER FBI ASST DIRECTOR: "Jim Comey 'Danced With The Devil'...I'm Glad He's Gon...
3	<1%	-	-	WHY WOULD OBAMA ALLOW GREEN BERET To Be Discharged For Saving Life Of You...
3	<1%	-	-	GARY JOHNSON: Meet The "Creepy" Pro-Amnesty, Anti-Gun, Pro-TPP, Pro-Abortion, D...
3	<1%	-	-	DISGRACEFUL: US Air Force Can No Longer Afford 21-Gun Salute At Vet Funerals...Ple...
3	<1%	-	-	HOUSE DEMOCRATS MAKE STUNNING Move To Implement Sharia Law In America
23,455	>99%	21,417	100%	(Other)

2 text

	True Data		Fake Data	
VALUES:	23,481 (100%)		21,417 (100%)	
MISSING:	---		---	
DISTINCT:	17,455	(74%)	21,192	(99%)

626	3%	1	<1%	
4	<1%	-	-	AP News The regulars amble in before dawn and claim their usual table, the one next t...
4	<1%	-	-	Things didn t go as well as Nicholle had hoped and now, she s lashing out at the owne...
4	<1%	-	-	On September 15, Hillary apparently held a rally in the Old Student Recreational Center...
4	<1%	-	-	
4	<1%	-	-	Enjoy:
4	<1%	-	-	Because there wouldn t be any outrage over a white kids only field trip to a local colleg...
22,831	97%	21,416	>99%	(Other)

3 subject

	True Data		Fake Data	
VALUES:	23,481 (100%)		21,417 (100%)	
MISSING:	---		---	
DISTINCT:	6	(<1%)	2	(<1%)

4 date

	True Data		Fake Data	
VALUES:	23,481 (100%)		21,417 (100%)	
MISSING:	---		---	
DISTINCT:	1,681	(7%)	716	(3%)

46	<1%	-	-	May 10, 2017
44	<1%	-	-	May 5, 2016
44	<1%	-	-	May 26, 2016
44	<1%	-	-	May 6, 2016
43	<1%	-	-	May 11, 2016
42	<1%	-	-	May 12, 2016
40	<1%	-	-	May 9, 2016
23,178	99%	21,417	100%	(Other)

Fig. 17. COMBINED REPORT