Fake News Detection Using ML

Ronak Parikh, Ethan Park

Emory University

Abstract

The rapid spread of misinformation and fake news has become a major challenge in modern society, influencing public opinion, political discourse, and public health decisions. Detecting fake news is difficult because misleading articles often imitate legitimate journalism and adopt language designed to appear credible. In this project, we study how different machine learning approaches perform on the task of fake news classification, with an emphasis on predictive accuracy, interpretability, and feasibility under realistic computational constraints.

We evaluate multiple models on the WELFake dataset, which contains over 72,000 labeled news articles from diverse sources. Our experiments compare classical TF-IDF based models, including logistic regression and random forest, with a fine-tuned transformer model, BERT, using metrics such as F1 score, ROC-AUC, and PR-AUC. We also apply interpretability techniques to analyze which textual features most influence model predictions. Results show that TF-IDF with logistic regression provides a strong and competitive baseline despite its simplicity, while more complex models offer potential gains on context-dependent cases at higher computational cost. These findings highlight important trade-offs between model complexity, performance, and transparency in real-world fake news detection systems.

Introduction

Machine learning has become a widely used approach for fake news detection as misleading articles increasingly resemble legitimate journalism in tone and structure. Because fake content is often written to appear credible, simple keyword-based rules or manual review are rarely effective. Prior work has shown that learning-based models can identify linguistic patterns that help distinguish real news from fake content (Pérez-Rosas et al. 2018).

While many studies report strong classification performance, fewer examine how different model families compare under realistic constraints. Classical models are typically efficient and interpretable, while transformer-based models often achieve higher accuracy but require greater computational resources but are harder to analyze (Shu et al. 2019). As fake news

detection systems move closer to real-world use, understanding these trade-offs becomes increasingly important.

In this project, we compare classical machine learning models with transformer-based models for fake news detection using a shared dataset and evaluation setting. Specifically, we evaluate Logistic Regression and Random Forest models trained on TF-IDF features alongside a fine-tuned BERT model on the WELFake dataset (Shahane 2021). We also apply SHAP-based interpretability analysis to examine how different models make decisions (Ribeiro et al. 2016). Our contribution lies in providing a practical comparison that highlights when simpler models remain competitive and when more complex approaches offer meaningful benefits.

Background

Fake news detection has been actively studied in recent years, with early work focusing on linguistic and statistical features. Pérez-Rosas et al. (2018) showed that word frequency, syntactic patterns, and n-grams combined with classifiers such as Logistic Regression and Support Vector Machines can achieve strong performance. These results demonstrated that relatively simple text representations can capture signals of deception, though they may struggle when fake news closely imitates professional reporting.

Later studies incorporated broader contextual information. Shu et al. (2019) introduced the FakeNewsNet dataset, highlighting the importance of metadata such as publisher information and temporal patterns for improving model robustness. Building on this direction, Shahane (2021) released the WELFake dataset, which merged multiple datasets into a large and balanced corpus and enabled more systematic evaluation of fake news detection models.

TF-IDF remains popular due to its simplicity and interpretability, but it does not capture word order or semantic context. Models such as Random Forests can partially address this limitation by modeling nonlinear feature interactions, though they still rely on handcrafted representations. More recent work has focused on transformer-based models such as BERT, which use contextual embeddings to capture deeper semantic relationships. While these models often outperform traditional approaches, their higher computational cost and reduced interpretability motivate continued comparison with simpler methods (Ribeiro et al. 2016).

Methods

Dataset and Problem Setup

We use the publicly available WELFake dataset from Kaggle, which contains 72,134 labeled news articles, with roughly equal numbers of real and fake examples. Each article includes a title, full body text, and a binary label. Because the dataset combines articles from multiple sources with varying writing styles and lengths, it provides a realistic setting for evaluating fake news detection models.

To capture both headline-level cues and full-article semantics, we concatenate the title and body text into a single input string. The dataset is split into 70 percent training, 15 percent validation, and 15 percent test sets. Five-fold cross-validation is used on the training data for model selection and hyperparameter tuning.

Preprocessing and Feature Engineering

Text data was preprocessed using the following pipeline:

- All text was lowercased and English stop words were removed to reduce redundancy and noise.
- Missing values in article titles or body text were replaced with empty strings to ensure consistent input formatting.
- No stemming or lemmatization was applied, as preserving original word forms improves interpretability for later SHAP-based analysis.
- TF-IDF vectorization was applied using both unigrams and bigrams (ngram_range = (1, 2)), allowing the models to learn from individual words as well as short phrases such as “breaking news” or “claims that.” (This parameter was tuned).
- The resulting TF-IDF representation produced a sparse, high-dimensional feature space that emphasizes informative terms while reducing the influence of overly common words.

Models

Since the task is binary classification, we evaluate three supervised learning models: Logistic Regression, Random Forest, and a fine-tuned BERT transformer.

1. Logistic Regression (TF-IDF)

Logistic Regression serves as a strong and interpretable baseline for high-dimensional sparse data. It models the probability that an article is fake using a linear combination of TF-IDF

features. We apply class weighting to address minor label imbalance and tune the regularization strength, penalty type, solver, and iteration limit using cross-validation.

2. Random Forest (TF-IDF)

Random Forest is an ensemble method that aggregates predictions from multiple decision trees trained on different feature subsets. This allows the model to capture nonlinear interactions between words that linear models may miss. We tune the number of trees, maximum tree depth, and minimum samples per leaf to balance performance and overfitting.

3. BERT Fine-Tuned Model

To capture deeper contextual meaning, we fine-tune the pre-trained *bert-base-uncased* model using the Hugging Face Transformers library. Unlike TF-IDF, BERT represents each word based on its surrounding context, allowing it to capture nuanced semantic relationships. Input text is tokenized with fixed padding and truncated to a maximum length of 256 tokens. The model is optimized using AdamW with a linear learning rate schedule. Due to computational constraints, fine-tuning is performed for a limited number of epochs, with evaluation focusing on recall-oriented metrics to prioritize fake news detection.

Hyperparameter Tuning

Hyperparameters were selected using five-fold cross-validation (cv = 5) on the training set. Because the dataset exhibits slight class imbalance and false negatives are costly in fake news detection, F1 score was used as the primary metric for model selection.

Logistic Regression (TF-IDF)

- Grid search was used for hyperparameter tuning.
- Inverse regularization strength $C \in \{0.25, 0.5, 1.0, 2.0\}$
- TF-IDF n-gram range tested: unigrams (1, 1) and bigrams (1, 2).
- Solver fixed to liblinear with L2 regularization.
- Maximum number of iterations set to 3000 to ensure convergence on high-dimensional data.

Random Forest (TF-IDF)

- Randomized search was used to balance performance and computational cost.
- Number of trees tested: 100 and 200.
- Maximum tree depth: None or 10.
- Minimum samples required to split a node: 2 or 5.
- Maximum number of features per split set to sqrt.

- Unigram TF-IDF features were used to reduce feature dimensionality and overfitting.

BERT Fine-Tuning

- Hyperparameter tuning was limited due to computational constraints (running on Apple M1 Pro)
- Fine-tuning was performed for a small number of epochs with early stopping to prevent overfitting.

Evaluation Metrics

All models are evaluated on a held-out test set that is not used during training or hyperparameter tuning. We report the following metrics:

- Accuracy: Overall proportion of correct predictions; easy to interpret but can be misleading under class imbalance.
- Precision (Positive class = Fake): Fraction of articles predicted as fake that are truly fake; controls false positives.
- Recall (Positive class = Fake): Fraction of truly fake articles that are correctly identified; controls missed detections.
- F1 Score: Harmonic mean of precision and recall; used as the primary metric for model selection during tuning.
- ROC-AUC: Area under the ROC curve measuring ranking quality across thresholds; relatively robust to class imbalance.
- PR-AUC (AUPRC): Area under the precision–recall curve emphasizing the positive (fake) class; more informative for skewed data and the primary metric for BERT tuning.
- Confusion Matrix: Counts of true positives, false positives, true negatives, and false negatives used to analyze error patterns.

Experiment / Results

Original Data Description and Experimental Setup

We evaluate our models on the WELFake dataset, which contains 72,134 news articles labeled as real or fake, with a roughly balanced class distribution. Each sample includes a headline, full article text, and a binary label. Because the dataset aggregates articles from multiple sources with varying writing styles and lengths, it provides a realistic benchmark for fake news detection.

To capture both headline-level signals and full-article context, we concatenate the title and body text into a single input sequence. The dataset is split into 70% training, 15% validation, and 15% test sets, with all reported results computed on the held-out test set only.

For Logistic Regression and Random Forest, text was lowercased and vectorized using TF-IDF with unigrams and bigrams. No stemming or lemmatization was applied to preserve interpretability and consistency with explainability analyses. For BERT, raw text was tokenized using the pretrained tokenizer without manual feature engineering, allowing the model to learn contextual representations directly from the data.

Empirical Results

After selecting optimal hyperparameters through five-fold cross-validation, we evaluate Logistic Regression, Random Forest, and BERT using standard binary classification metrics. Model selection during tuning is based on F1 score, as it balances precision and recall and better reflects performance on fake news detection where missed detections are costly.

Logistic Regression Results

- Accuracy: 0.9567
- Precision (Fake): 0.9572
- Recall (Fake): 0.9586
- F1 Score: 0.9579
- ROC-AUC: 0.9903
- PR-AUC: 0.9906

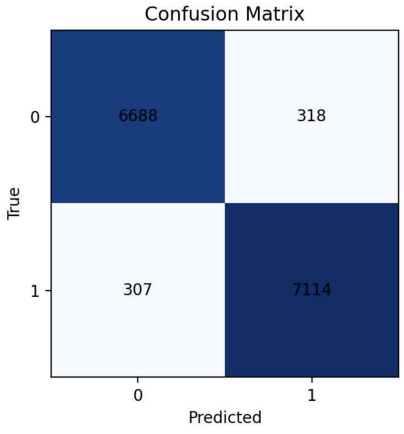


Fig. 1

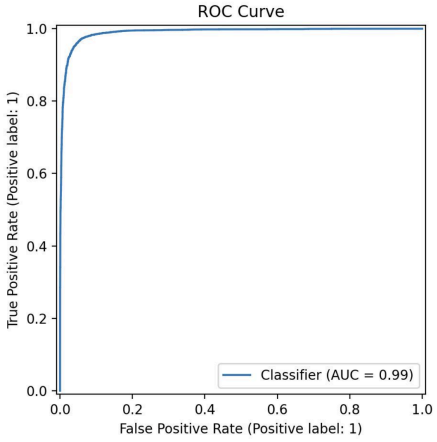


Fig. 2

The confusion matrix (Fig. 1) shows balanced error rates across classes, with relatively few false positives and false negatives. The ROC curve (Fig. 2) approaches the top-left corner,

indicating strong ranking performance. These results confirm that TF-IDF paired with a linear classifier remains a highly competitive and efficient baseline for fake news detection.

Random Forest Results

- Accuracy: 0.9393
- Precision (Fake): 0.9354
- Recall (Fake): 0.9476
- F1 Score: 0.9414
- ROC-AUC: 0.9884
- PR-AUC: 0.9895

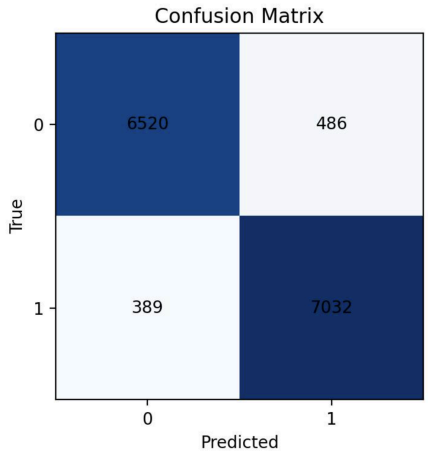


Fig. 3

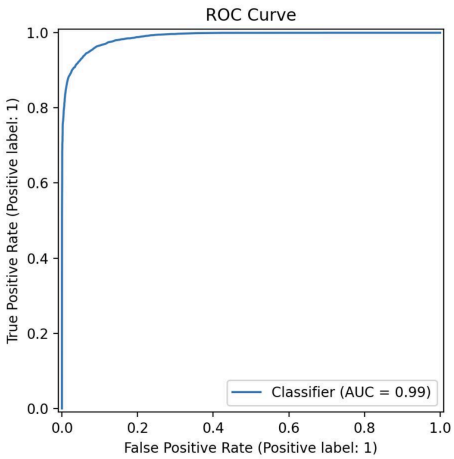


Fig. 4

Random Forest captures nonlinear interactions between textual features, but overall performance is slightly lower than Logistic Regression. The confusion matrix (Fig. 3) indicates a higher false positive rate, suggesting that added model complexity does not consistently translate to better generalization in this high-dimensional sparse feature space. The ROC curve (Fig. 4) approaches the top-left corner similar to the ROC curve from the logistic regression (Fig. 2).

BERT Fine-Tuning Results

- Accuracy: 0.9942
- Precision (Fake): 0.9953
- Recall (Fake): 0.9933
- F1 Score: 0.9943
- PR-AUC: 0.9999

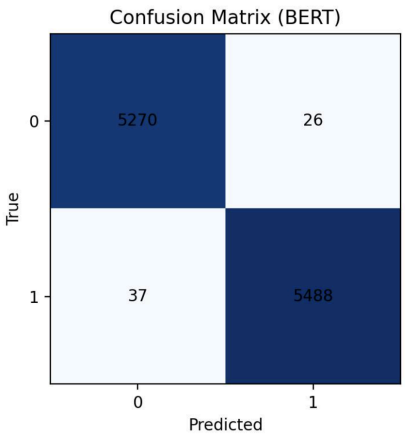


Fig. 5

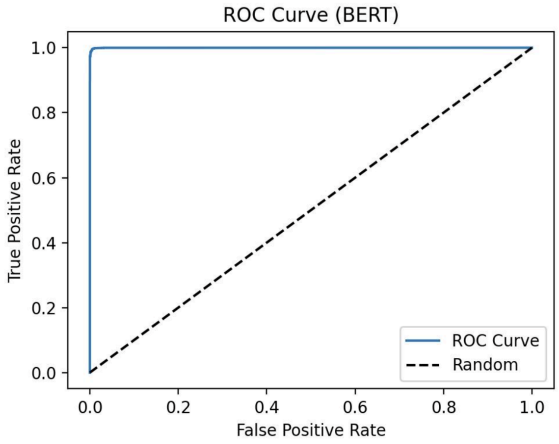


Fig. 6

The BERT model significantly outperforms both TF-IDF–based approaches across all evaluation metrics. The confusion matrix (Fig. 5) shows extremely low error rates, with only 26 false positives and 37 false negatives, indicating strong and balanced performance across both classes.

The ROC curve (Fig. 6) closely follows the top-left corner, reflecting near-perfect ranking quality and clear separation between real and fake articles. These results highlight the advantage of contextualized language representations, as BERT is able to capture subtle semantic cues and long-range dependencies that are difficult for bag-of-words models to model effectively.

Comparative Analysis

	Accuracy	Precision	Recall	F1 Score	ROC-AUC	PR-AUC
Logistic Regression	0.9567	0.9572	0.9586	0.9579	0.9903	0.9906
Random Forest	0.9393	0.9354	0.9476	0.9414	0.9884	0.9895
BERT Fine-Tuning	0.9942	0.9953	0.9933	0.9943	0.9999	0.9999

Table 1

Looking across the three models (Table 1), a clear pattern emerges: performance improves as the models gain a deeper understanding of language. Logistic Regression performs surprisingly well given its simplicity. Using TF-IDF features, it achieves strong and balanced

results, showing that a straightforward linear model can still be highly effective for fake news detection. This makes it a solid and efficient baseline, especially when interpretability and computational cost matter.

Random Forest adds model complexity by capturing nonlinear feature interactions, but the results suggest that this added flexibility does not consistently improve performance. In fact, its slightly lower precision and F1 score indicate that more complex models can struggle to generalize when working with sparse, high-dimensional TF-IDF features. This reinforces the idea that complexity alone does not guarantee better outcomes in text classification tasks.

BERT Fine-Tuning clearly stands out. By leveraging contextualized language representations, BERT is able to capture subtle semantic cues and long-range dependencies that simpler models miss. This leads to substantially higher accuracy and near-perfect ranking performance. Overall, the comparison highlights an important trade-off: traditional models remain strong and efficient baselines, but transformer-based models provide a clear performance advantage when computational resources allow.

Discussion

The results of our experiments suggest that how text is represented plays a more important role in fake news detection than model complexity alone. Even though TF-IDF relies on surface-level word frequencies, it captures many strong signals associated with deceptive writing. This helps explain why Logistic Regression performs so well despite its simplicity, showing that carefully chosen features can still be highly effective for this task.

Our findings also show that increasing model complexity without changing the underlying representation does not automatically improve performance. While Random Forest introduces nonlinear decision boundaries, pairing it with sparse TF-IDF features does not consistently lead to better generalization. This highlights a key limitation of bag-of-words style representations, which struggle to encode deeper semantic relationships regardless of the classifier placed on top of them.

In contrast, the strong performance of BERT highlights the impact of contextualized language representations. By modeling words in relation to their surrounding context, BERT is able to capture subtle meanings, implied claims, and long-range dependencies that frequency-based approaches often miss. This suggests that future gains in fake news detection are more likely to come from advances in representation learning rather than incremental changes to traditional machine learning models.

Overall, these results point to an important practical lesson. Simpler models remain attractive when interpretability, efficiency, and limited computational resources are priorities. At the same time, transformer-based models provide clear benefits when accuracy is the primary concern. Recognizing and balancing these trade-offs is essential when deploying fake news detection systems in real-world settings, where constraints and goals can vary widely.

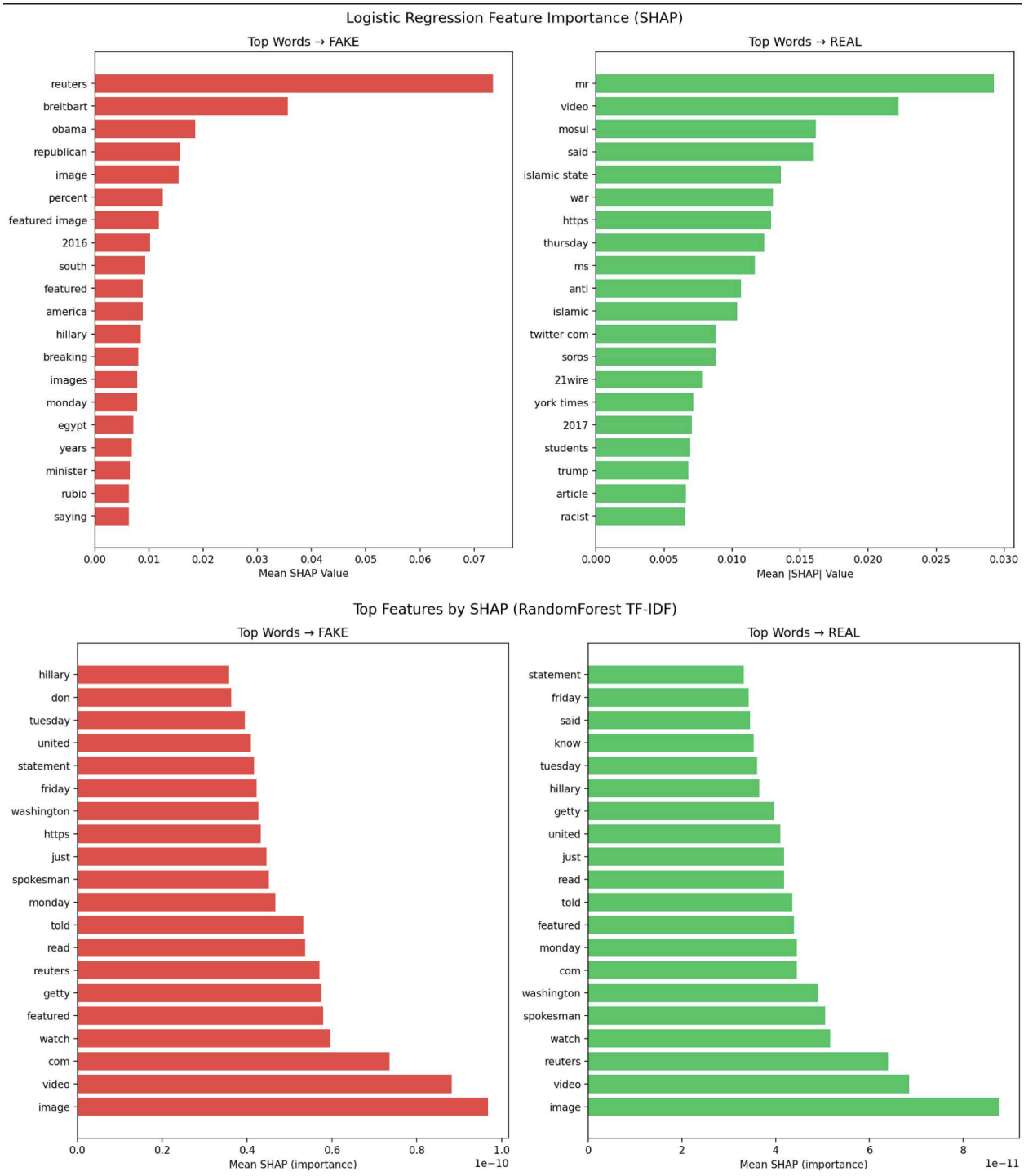
Contributions

Ronak Parikh
Focused on building and evaluating the traditional machine learning models used in the project, specifically Logistic Regression and Random Forest. Worked on setting up the TF-IDF feature representation, training the models, and tuning hyperparameters using five-fold cross-validation with F1 score as the main selection criterion. Analyzed model performance using accuracy, precision, recall, F1 score, ROC-AUC, PR-AUC, and confusion matrices, and examined common error patterns. Took the lead on writing the Methods sections related to these models and documenting their experimental results. Responsible for coding, analysis, presentation, and reporting of this work.

Ethan Park
Focused on the transformer-based approach by implementing and fine-tuning a BERT model for fake news detection. Handled text tokenization, model setup, and training within available computational limits, and selected hyperparameters based on validation performance with attention to F1 score for fake news articles. Compared BERT's performance with traditional models, highlighting strengths, limitations, and computational trade-offs. Took the lead on documenting the BERT methodology and results, and contributed to the overall discussion and presentation materials. Responsible for coding, analysis, presentation, and reporting of this work.

Code/dataset

Dataset Link: https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification
Code: https://github.com/ronakparikh52/FakeNewsClassifier


Logistic Regression Feature Importance (SHAP)
Top Words → FAKE
reuters
breitbart
obama
republican
image
percent
featured image
2016
south
featured
america
hillary
breaking
images
monday
egypt
years
minister
rubio
saying
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Mean SHAP Value
Top Words → REAL
mr
video
mosul
said
islamic state
war
https
thursday
ms
anti
islamic
twitter com
soros
21wire
york times
2017
students
trump
article
racist
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Mean |SHAP| Value
Top Features by SHAP (RandomForest TF-IDF)
Top Words → FAKE
hillary
don
tuesday
united
statement
friday
washington
https
just
spokesman
monday
told
read
reuters
getty
featured
watch
com
video
image
0.0
0.2
0.4
0.6
0.8
1.0
1e−10
Mean SHAP (importance)
Top Words → REAL
statement
friday
said
know
tuesday
hillary
getty
united
just
read
told
featured
monday
com
washington
spokesman
watch
reuters
video
image
0
2
4
6
8
1e−11
Mean SHAP (importance)

CS 334, supervised by Dr. Li Xiong, December 15, 2025