

A conceptual illustration on a light gray textured background. A woman with brown hair, wearing a white shirt and light gray pants, is running to the left on a bright yellow ground. A large, realistic-looking hand from the right side of the frame holds several thin black strings. These strings are attached to small black rectangular labels, each with the word 'fakes' written in white. The strings lead from the hand towards the woman, suggesting she is being pulled or ensnared by these 'fake' elements.

# NLP Project - Fake News

Mauricio & Kepa

## 1. Data Loading and Exploration

- A dataset of approximately 40,000 news articles (50% real, 50% fake).
- Main columns: **label**, **title**, **text**, subject, and date.
- Initial exploration: checked data structure, missing values, and label distribution.

## 2. Preprocessing

- Text cleaning: converted to lowercase, removed punctuation, numbers, and stopwords.
- Lemmatization.
- TF-IDF vectorization (with bigrams), with vocabulary of about 8,500 relevant words.

## 3. Feature Engineering

- TF-IDF features with an n-gram range of (1,2) to capture word combinations.
- Filtered out biased and overly rare words to improve model generalization and avoid data leakage.

## 4. Models

Model	Accuracy	Observation
Logistic Regression	0.877	Good balance between precision and recall
Naive Bayes	0.839	Slight drop in recall performance
Random Forest	0.881	Best overall performance and stability

## 5. Key results

- F1-score average  $\approx 0.88$
- Important Words (LR): *trump, said, state, president*

# Validation

- It does not perform well (tendency seems correct)
- Presence of Leakage words

# Deployment

Get it done easily

What we wanted... (docker + own server)  
[Spoiler!](#)

**Fake News Detection**

Enter the title and text of a news article. Predictions will be made using Naive Bayes, Logistic Regression, and Random Forest.

<b>News Title</b> <input type="text" value="Enter the news title here..."/>	<b>Output</b> <div></div>
<b>News Text</b> <input type="text" value="Enter the news text here..."/>	<div>Flag</div>

**Fake News Detector**  
AI-Powered News Verification System

**News Title**

**News Content**

**Select Model**  
☒ Ensemble (Best) ☐ Logistic Regression ☐ Naive Bayes

# Learning Takeaways

- **Data caution:** subjects affect real/fake likelihood and were considered in modeling.
- **Advanced preprocessing:** Handling emojis, contractions, corrections, and sentiment improves text understanding.
- **Embeddings:** Word embeddings capture semantic meaning and boost model performance beyond TF-IDF. Implementation?
- **Docker** deployment
- **Clean data** is essential to avoid overfitting.
- **Complexity** of the project from scratch
- **Time management**
- High level of **accuracy** doesn't mean the model is correct

