Factify – Source-Based Fake News Classifier

A PROJECT REPORT
BY

Charvi Singh (E23CSEU0843)

SUBMITTED TO

SCHOOL OF COMPUTER SCIENCE ENGINEERING AND TECHNOLOGY, BENNETT UNIVERSITY

GREATER NOIDA, 201310, UTTAR PRADESH, INDIA

April 2025

📌 What This Project Is About

In today's fast-paced digital world, **misinformation travels faster than ever**. So, I set out to build a model that can classify news articles as **real** or **fake**, but with a twist:
Instead of analyzing the article content, I focused on the **source metadata**—like the author, publication date, and credibility of the platform.
This approach helps us understand if we can trust a piece of news just based on **where it's coming from**, not what it says. It's a simpler, more explainable way of judging news credibility.

Most fake news detection systems focus heavily on **natural language processing (NLP)** to study the **text content** of an article—looking for unusual writing patterns, sentiment, grammar, or semantic inconsistencies. While this is definitely useful, I took a slightly different approach:
What if we could tell whether a news article is trustworthy just by looking at where it's coming from?
In this project, I trained models using **source-related features**—such as:

- The **author** of the article
- The **publication date and time**
- The **source or platform** where the article was published
- Whether the source was previously known to publish reliable or misleading content

🛠 Tools & Techniques Used

1. **Python (Jupyter Notebook):**
   Used for coding, experimenting, and visualizing results in an interactive environment.

2. **scikit-learn:**
   Implemented traditional ML models like Random Forest for fast, interpretable classification.

3. **Pandas:**
   Handled data cleaning, preprocessing, and transformation of metadata features.

4. **CNN + Embeddings (Keras):**
   Built a deep learning model to capture patterns in metadata using embedded representations.

5. **TF-IDF:**
   Converted text-based metadata (like author names) into numerical features for ML models.

6. **Evaluation Metrics (Accuracy & F1-score):**
   Measured overall model performance and balanced prediction quality for both classes.

Dataset Used

Name: *Getting Real about Fake News*

Source: Kaggle
This dataset contains structured metadata about news articles — without relying on the actual article text — making it perfect for source-based credibility classification.

Features Included:

- **Author -** Ruchi
- **Label** – Ground truth indicating if the news is *Real* or *Fake*
-

What I Did:
I cleaned the data, handled missing values, and preprocessed metadata fields like authorship and publication date to make the dataset model-ready.

Why This Dataset?
Unlike many datasets that focus only on content, this one emphasizes **source information**—a key factor in assessing credibility. It's already preprocessed for skew and fits perfectly for experimenting with source-based fake news detection.

Reference Paper:
You can check out the original research work that inspired this dataset here:
http://www.ijirset.com/upload/2020/june/115_4_Source.PDF

Models I Built

1. RandomForestClassifier with TF-IDF

- Transformed article metadata (like author/source) into vectors using TF-IDF.
- Trained a Random Forest model on this structured text data.
- Easy to interpret and quick to train.
- **Limitation:** Doesn't understand context or deeper semantics.
-

2. Embeddings + CNN

- Used word embeddings to capture **semantic relationships** between words.
- CNN layers helped detect patterns and structure in the input features.
- Performed better on both training and testing data.
- **Limitation:** Needs more training data and resources.

Comparison of Both Models

Model	Pros	Cons
Random Forest + TF-IDF	Simple & efficient	Lacks context understanding
Embeddings + CNN	Captures deeper meaning	Heavier on resources & time

Results & Learnings

- The CNN model outperformed Random Forest, especially on test data.
- Structured metadata (like author & source) can be a useful signal for credibility.
- Explainable models are just as important as accurate ones, especially in sensitive domains like misinformation.
- This project taught me how technical tools can contribute to digital literacy and help users become more critical consumers of news.

⚠️ **Disclaimer**
This project was built for **educational purposes** only.
It's not a replacement for fact-checking or journalistic processes. Please don't rely on it for real-world verification—**always cross-check with trusted sources**!