HEADLINE HUNTER

MSML 606 PROJECT

Nikita Miller Yatish Sikka
Anisha Katiyar Aariz Faridi

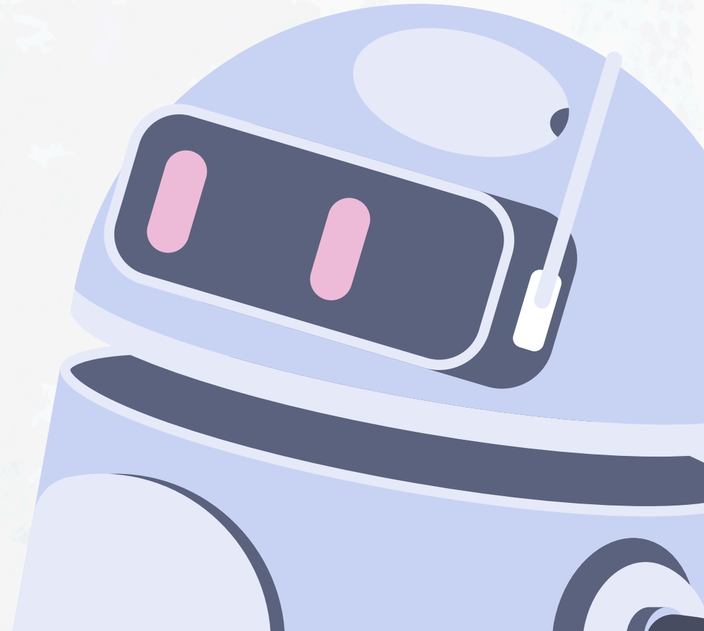


TABLE OF CONTENTS

01

Introduction

MOTIVATION

Fake news spreads rapidly across social media and online platforms, especially with people relying on just titles.

Traditional detection methods either rely on black-box models or are too resource-intensive for scalable deployment.

There's a growing need for efficient, interpretable systems that can work in real time and under resource constraints.

Problem Statement

We aim to develop an efficient and interpretable fake news detection system that balances accuracy, resource usage, and scalability for real-world deployment.

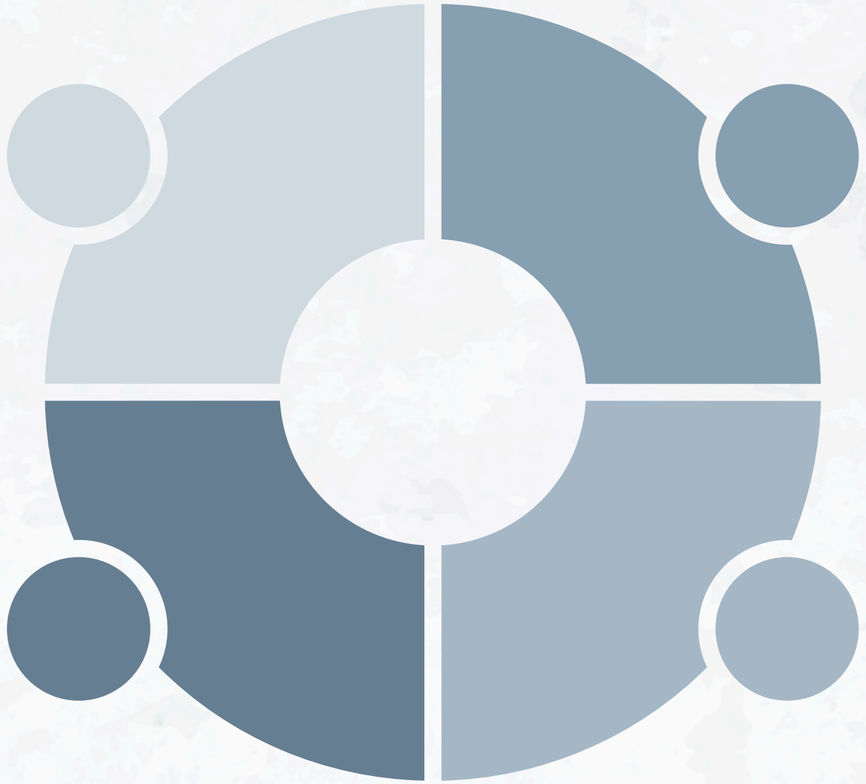
02

Dataset and Preprocessing

DATASET

We used the Fake News Detection Dataset on Kaggle

Contained title and text of real and fake articles

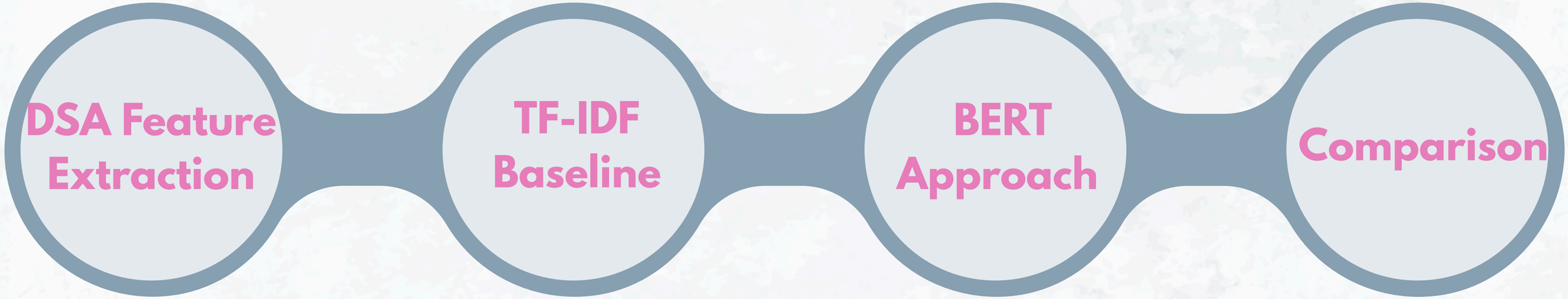


<u>Cleaned dataset</u>: Converted all text to lowercase, removed HTML tags, URLs, bracketed citations, punctuation, and extra spaces

Balanced dataset with 2 labels, 0 for fake and 1 for real article

Fake title often differs in tone from text

METHODOLOGY



PREPROCESSING

For the DSA approach:

- Trie → detect clickbait phrases in title
- Heap → extract tp-k longest word lengths
- Set → compute Jaccard similarity
- Hashmap → max word frequency ratio
- Regex → punctuation emphasis
- Capital word scan → longest uppercase word

PREPROCESSING

TF-IDF Vectorization

- Combined title + text
- Transformed using TfidfVectorizer(max_features=5000)
- Input to Random Forest classifier

PREPROCESSING

BERT

- Combined title + text
- Tokenized using BertTokenizer
- Truncation + padding to max length
- Input to BertForSequenceClassification via HuggingFace Trainer

03

Data Structures Used

Trie

- A Trie or a prefix tree used for efficient string matching.
- Enables O(m) time search (m = length of phrase).
- Stored a curated list of common clickbait phrases in a Trie.
- E.g., “You won’t believe”, “Shocking truth”, “This one trick...”.
- Each news headline is tokenized and checked for these phrases.
- Helps flag sensationalized content during early-stage analysis.

Heap

- A max-heap allows constant-time access to the longest element.
- Supports maintaining top-k elements in O(n log k) time.
- Used a max-heap to identify the top-k longest words in news articles and titles.
- Acts as a heuristic for technical or complex vocabulary, potentially signaling article credibility by checking for over exaggeration.

Sets

- A collection of unique, unordered elements.
- Supports fast membership tests and set operations (e.g., union, intersection).
- Ideal for computing overlap between distinct token sets.
- Used to compute Jaccard Similarity between title and body of an article:
  - $J(A, B) = |A \cap B| / |A \cup B|$
- Low similarity scores flag possible headline-content mismatch, common in fake news.

HashMap

- A key-value data structure for fast insertion and lookup.
- Python dict is a common implementation.
- Ideal for frequency counting and aggregation tasks.
- Store word frequencies from article title and body.
- Compute the ratio of most frequent word to total word count.
- Helps detect overused terms, common in spammy or low-effort content.

SUMMARY

Feature Name	Purpose	Underlying DSA	How DSA Is Used
clickbait_score	Detect if title contains known clickbait phrases	Trie (Prefix Tree)	Phrases like “you won’t believe” are inserted; title is scanned for prefix matches
title_topk_word_len	Avg. length of longest k words in title	Heap logic (nlargest)	heapq.nlargest() simulates a max heap to extract top k word lengths
text_topk_word_len	Same as above, for text	Heap logic (nlargest)	Same method applied to body text
title_max_word_ratio	Detects repetition	Dict (word count map)	A hash map is built to count word frequencies, then max() is applied
text_max_word_ratio	Most frequent word ÷ total words in title	Dict (word count map)	Same hash map and max() logic on full text
jaccard_title_text	Measure title-text token overlap	Set	Converts both to sets; computes intersection/union for Jaccard index
title_punct_emphasis	Find longest sequence of punctuation (“!!”)	Insertion Sort + Regex	Regex extracts runs of punctuation, then insertion sort finds the longest one
title_longest_cap_word	Longest all-caps word in title	Linear Scan + max()	Iterates through words, filters for isupper(), and uses max(len)

04

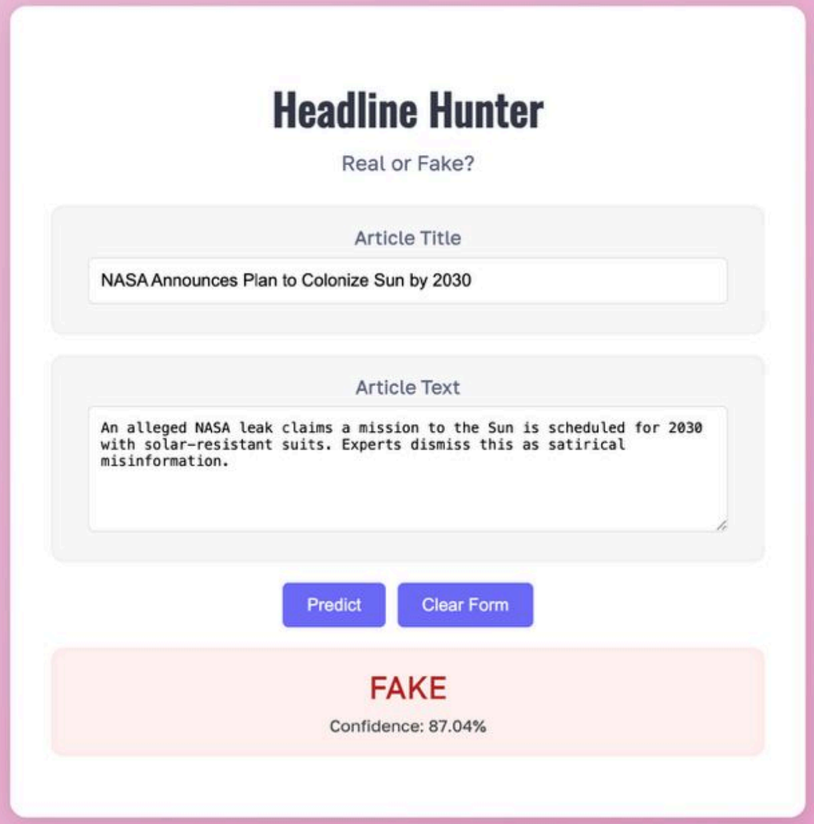
Results and Discussion

Results

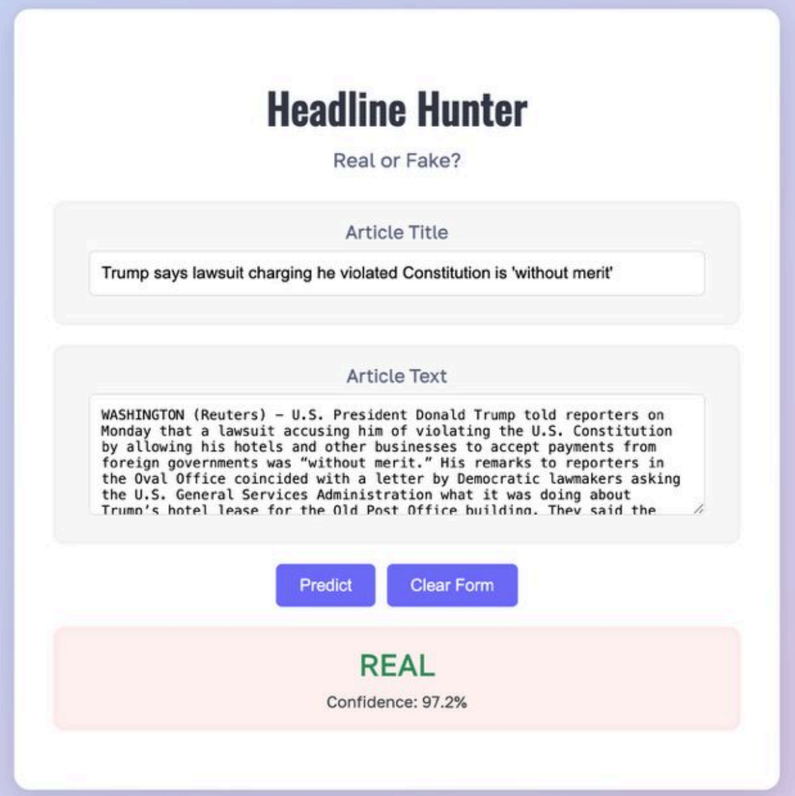
Model	Accuracy	F1 Score	Train Time	Inference Time	Model Size
DSA	0.9102	0.9211	1.94 sec	0.000008 sec	10.32 MB
TF-IDF	0.9597	0.9632	5.06 sec	0.000016 sec	5.01 MB
BERT	0.9795	0.9809	3.5 hours	0.1171 sec/sample	437.96 MB

- **DSA excels in simplicity and speed, making it ideal for real-time, low-resource environments.**
- **TF-IDF strikes a middle ground, offering decent accuracy with manageable resource demands, but struggles with subtle semantic relationships.**
- **BERT outperforms in understanding context, but its heavy computational load limits practical deployment in lightweight or real-time applications.**

Demo


Headline Hunter
Real or Fake?
Article Title
NASA Announces Plan to Colonize Sun by 2030
Article Text
An alleged NASA leak claims a mission to the Sun is scheduled for 2030 with solar-resistant suits. Experts dismiss this as satirical misinformation.
Predict
Clear Form
FAKE
Confidence: 87.04%

Demo


Headline Hunter
Real or Fake?
Article Title
Trump says lawsuit charging he violated Constitution is 'without merit'
Article Text
WASHINGTON (Reuters) - U.S. President Donald Trump told reporters on Monday that a lawsuit accusing him of violating the U.S. Constitution by allowing his hotels and other businesses to accept payments from foreign governments was “without merit.” His remarks to reporters in the Oval Office coincided with a letter by Democratic lawmakers asking the U.S. General Services Administration what it was doing about
Predict
Clear Form
REAL
Confidence: 97.2%

05

Limitations and Ethical Concerns

Limitations

DSA (Data Structures-Based):

- Relies on surface-level heuristics (e.g., word frequency, token overlap).
- Fails to capture semantics, context, or subtle misinformation tactics.
- Limited generalization beyond predefined patterns.

TF-IDF:

- Offers better accuracy than DSA, but...
- Suffers from high-dimensional vectors → more memory usage.
- Hard to interpret and doesn't account for word meaning or order.

BERT (Transformer-Based):

- State-of-the-art accuracy, but at a cost:
- Large model size: ~438 MB
- Slow inference: ~0.1171 sec/sample
- Not ideal for real-time or resource-constrained environments (e.g., mobile, edge devices).

Ethical Concerns

- False Positives: Misclassifying real news as fake can damage media credibility and threaten freedom of expression.
- Training Data Bias: Datasets may reflect imbalances in topic or source representation, leading to biased outcomes.
- Censorship Risks: Without transparency, automated systems could enable unjust content suppression, especially in sensitive contexts.

06

Impact and Application

Social Impact

- **Real-World Use Cases: Integration into fact-checking plugins, browser extensions, or messaging platforms, use by AI assistants to flag questionable content in real-time**
- **High-Stakes Examples: COVID-19 misinformation, election manipulation**
- **Long-Term Potential: Scalable, interpretable models enable use on edge devices (e.g., phones, browser tools)**
- **Helps build public trust through transparent, explainable detection**

Proposed Enhancements

- Stylometric Feature Extensions: Capitalization ratio (number of capital letters/total), Punctuation ratio (exclamations, ellipses)
- Average sentence length: Verbosity or simplicity patterns
- Semantic + Entity Features: Compare named entities in title vs. body (e.g., “Biden” in title, not in text)
- Use sentence embeddings (e.g., SBERT) to detect title-text drift

Thank You!