Fake vs True News – Data Analysis Report

---

1. Executive Summary

This project analyzes a dataset of ~45,000 news articles with the goal of distinguishing between **True** and **Fake** news. Using exploratory data analysis (EDA), text processing, SQL queries, and Power BI dashboards, we examined subject distributions, writing style, sentiment, and linguistic features.

The dataset is well balanced (52% Fake, 48% True), making it suitable for training machine learning models. Key insights reveal that **True news uses agency-style reporting language**, while **Fake news adopts a blog/social media tone with more visual and emotional cues**.

Source: https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets/data

---

2. Objectives

- Identify distinguishing characteristics between Fake and True news.
- Analyze word usage, article length, subject distribution, and sentiment.
- Highlight useful features for predictive modeling.
- Create dashboard summaries for subject, month, and class distribution.

---

3. Dataset Overview

- **Total Records:** 44,898
- **Columns:** title, text, subject, date, label, clean_text, label_str, word_count, char_count, avg_word_length, sentence_count, polarity
- **Subjects (8):** politicsNews, worldnews, News, politics, Government News, left-news, US_News, Middle-east

4. Data Cleaning

- Removed duplicates and null values.
- Converted date into proper YYYY-MM-DD format.
- Created new columns:
    - **is_news** (Yes = True, No = Fake)
    - **Month** (from date)
- Normalized label column (True, Fake).

---

5. Class Distribution

- **Fake:** 52.3%
- **True:** 47.7%
  The dataset is balanced, reducing bias in training.

---

6. Subject Distribution (Top 5)

1. politicsNews – 11,272
2. worldnews – 10,145
3. News – 9,050
4. politics – 6,841
5. left-news – 4,459

Observation: **Politics dominates** both fake and true categories.

7. Article Length & Style

- Avg. word count: **234.6**
- Median word count: **207**
- Max word count: **4958**
- Avg. char count: **1706.2**
- Sentence count stuck at **1.0** → text preprocessing merged articles into single blocks.

---

8. Sentiment Analysis

- Overall polarity: ~0.047 (neutral).
- True news avg polarity: 0.046
- Fake news avg polarity: 0.047
  Sentiment is **not a reliable differentiator**.

---

9. Vocabulary Insights

- Unique words in True news: **71,880**
- Unique words in Fake news: **72,082**
- Overlap: **32,897 words**
  Both categories are rich in vocabulary; preprocessing is critical.

---

10. Top Words & Phrases

- True News

- **Words:** said, Reuters, state, president, government

- **Bigrams:** united states, white house, donald trump
- **Trigrams:** president donald trump, washington reuters president

- Fake News

- **Words:** trump, people, obama, clinton, american
- **Bigrams:** donald trump, featured image, hillary clinton
- **Trigrams:** pic twitter com, new york time, black life matter

True news = agency & institutional language.
Fake news = blog/social cues, visual references, political spin.

---

11. Feature Correlations

- Word count ↔ Char count: **0.998 (very high)**
- Polarity has near-zero correlation with others.
  Word/char length are redundant; polarity is weak.

---

12. Predictive Features (TF-IDF + Logistic Regression)

- **True indicators:** reuters, washington, friday, tuesday, trumpâ
- **Fake indicators:** image, read, featured, pic, hillary, wire

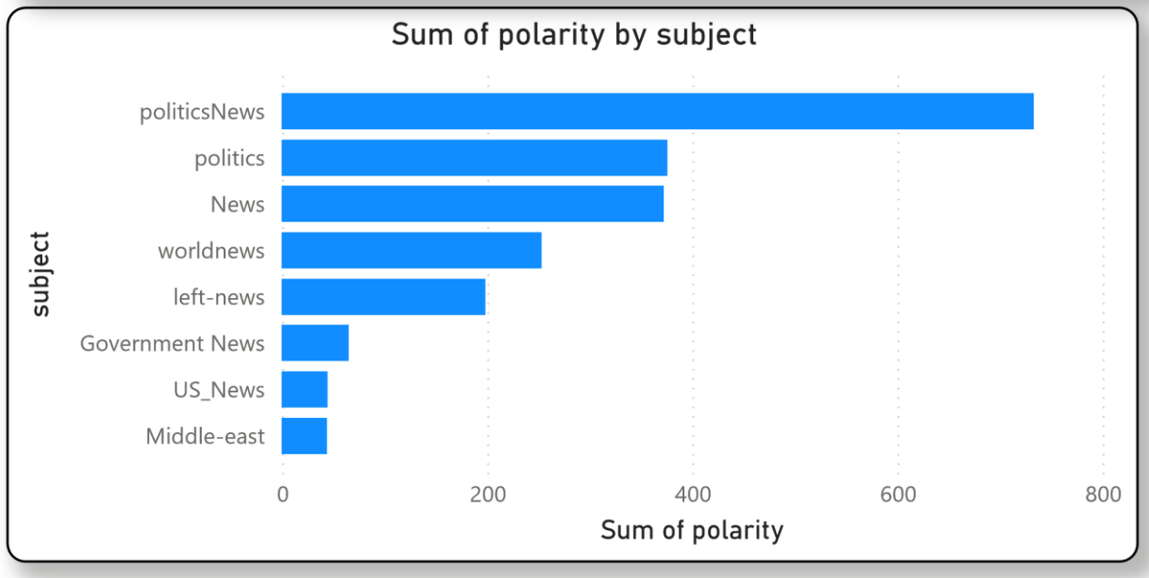
Observation:

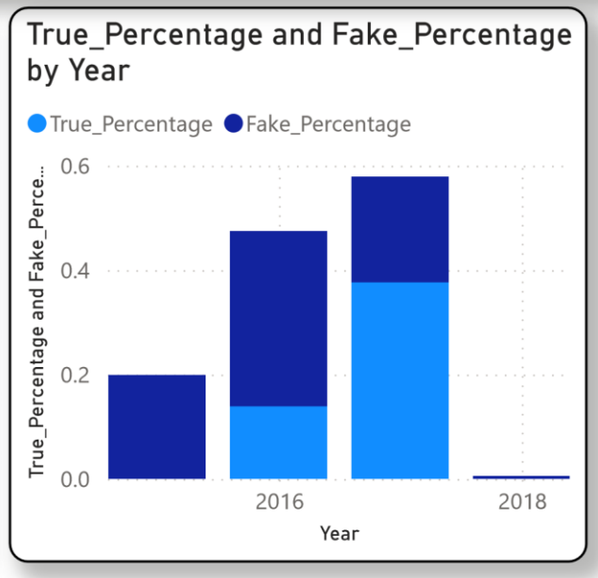
- **True news = facts + structured reporting**
- **Fake news = visuals + opinion-heavy language**

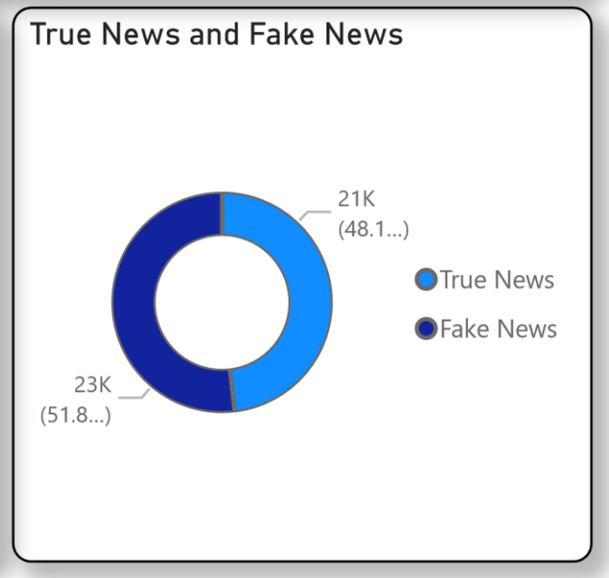
---

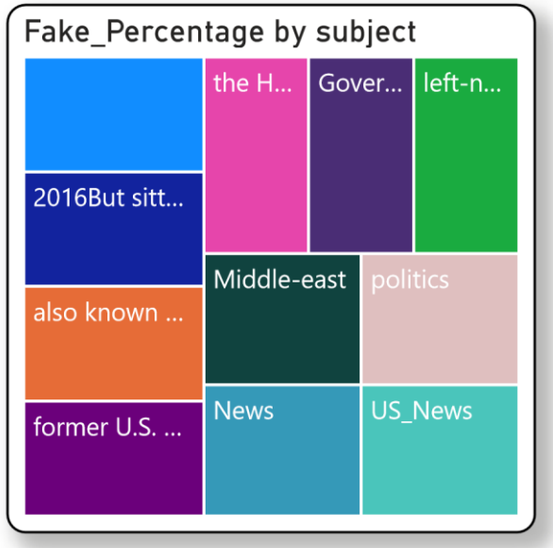
13. Power BI Dashboard Highlights

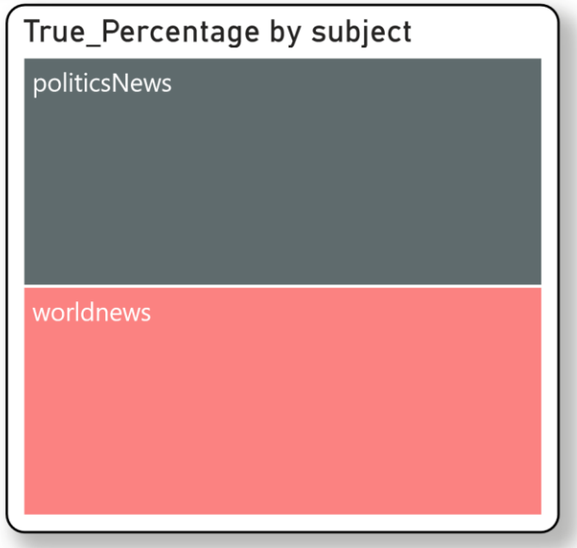
Fake News Detection - Dashboard

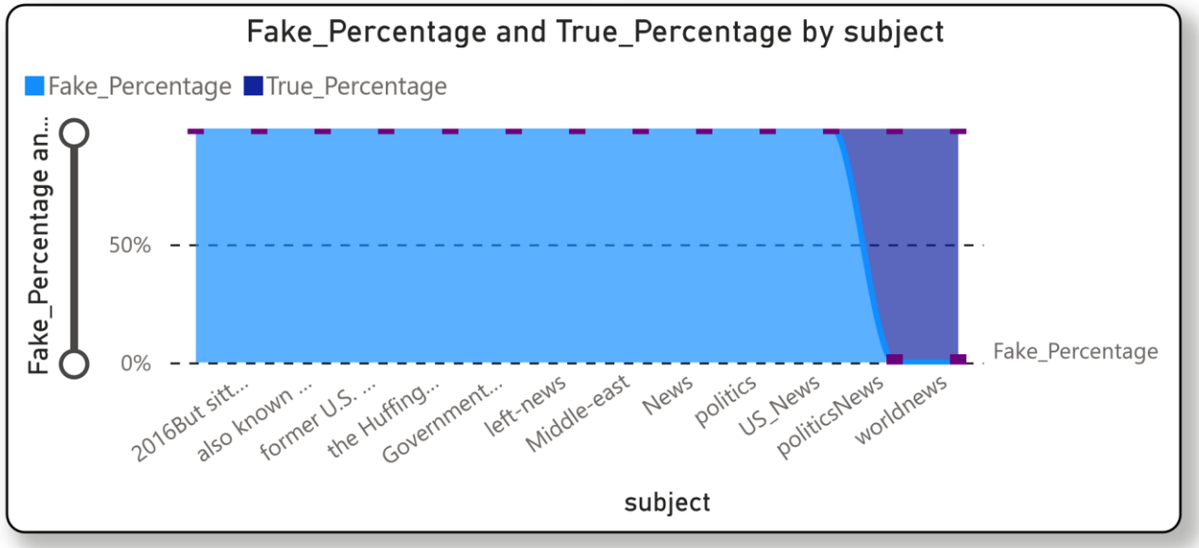












44K
Total Rows

0.48
True_Percentage

233.10
Avg Word Count True news

1.72K
Avg Char Count True n...

76M
Sum of char_count

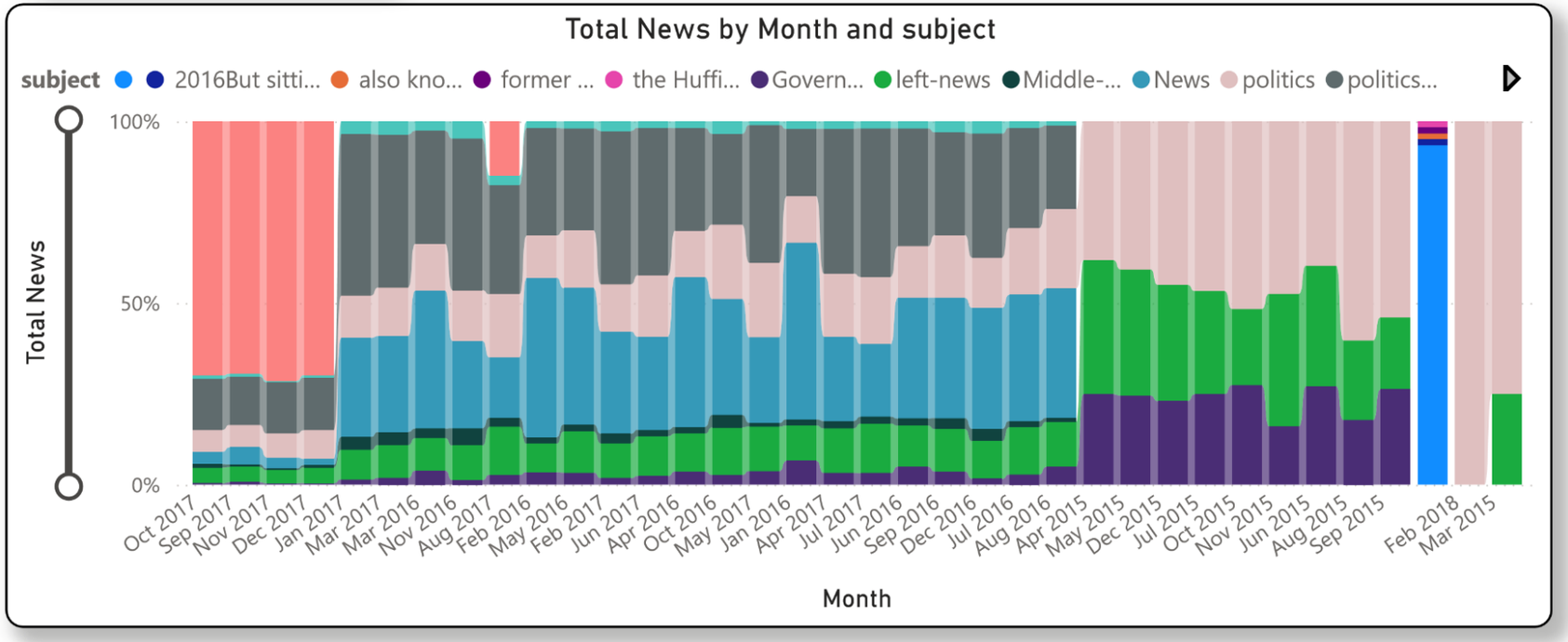
2.09K
Sum of polarity

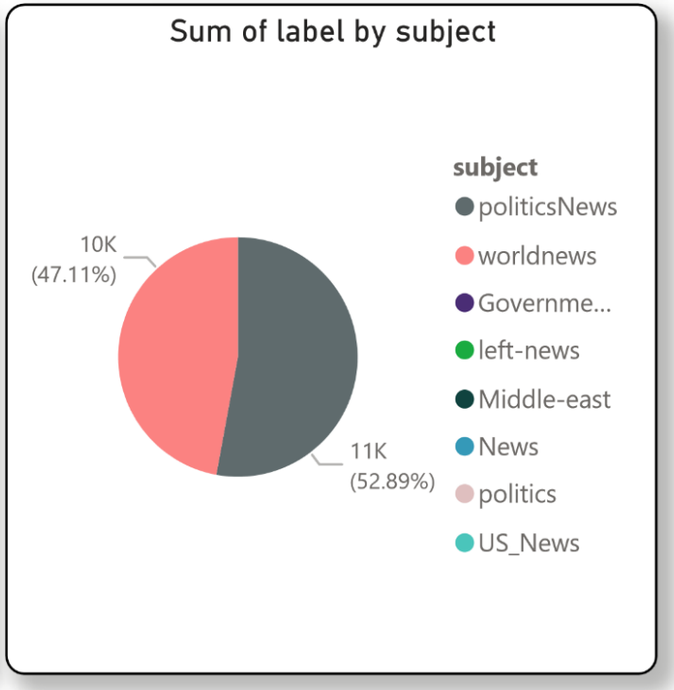
0.52
Fake_Percentage

242.91
Avg Word Count Fake news

1.74K
Avg Char Count Fake n...

274.96K
Sum of avg_word_length





14. Findings & Insights

- Dataset is balanced → good for modeling.
- Length metrics (word/char) = weak signals.
- Sentiment adds no value.
- Vocabulary and n-grams are **key differentiators**.
- Fake news is **shorter, more visual, emotional**.
- True news is **longer, agency-style, fact-based**.

---

15. Recommendations

- Focus on **n-grams, named entities, and TF-IDF features** for classification.
- Exclude polarity and redundant length metrics.
- Extend analysis by testing ML classifiers (Logistic Regression, Random Forest, BERT).
- Improve text preprocessing (fix sentence segmentation).

---

16. Conclusion

This project shows how linguistic style, subject matter, and word patterns can reliably distinguish Fake from True news. While sentiment and article length offer little insight, vocabulary signals (agency terms vs. social cues) are powerful predictors. The dataset, with its balance and diversity, is a strong foundation for building robust fake news detection models.