Fake vs Real News Analysis

DSK 822: News and Market Sentiment Analytics

Szimona Szternak - szszt24@student.sdu.dk

19 Dec 2025

1 Objective

The goal of this analysis is to explore stylistic and sentiment differences between fake and real news, and to build machine learning models to classify news articles based on text content. Additional insights relate to how sentiment-driven trading models could be biased by fake news.

2 Data

The analysis uses a publicly available fake and real news dataset from Kaggle [3].

- **Fake news dataset:** `Fake.csv`
- **Real news dataset:** `True.csv`
- **Columns:** `title, text, subject, date`
  - `title`: title of news article
  - `text`: body text of news article
  - `subject`: subject of news article
  - `date`: publish date of news article
- **Total samples:** 44,898 (before balancing)

3 Data Preprocessing

Before modeling, the title and text fields of each news article were combined into a single content field to provide a unified representation of the article's information. The text was then cleaned to improve quality and consistency: all characters were converted to lowercase, URLs and numbers were removed, punctuation was stripped, and extra whitespace was normalized [2].

In addition to the cleaned content, several stylistic and structural features were computed to capture differences in writing style between fake and real news. These included the number of exclamation marks (`exclamation_count`), the number of question marks (`question_count`), the ratio of uppercase letters (`uppercase_ratio`), the average sentence length (`average_sentence_length`), and lexical diversity, calculated as the proportion of unique words to total words (`lexical_diversity`).

Finally, to ensure a balanced dataset for modeling, 10,000 fake news articles and 10,000 real news articles were randomly sampled. This resulted in a total of 20,000 samples, allowing fair training and evaluation of classification models without class imbalance bias.

4 Exploratory Data Analysis (EDA)

4.1 Class distribution

The balanced dataset consists of an equal number of real and fake news articles, with 10,000 samples in each class. This ensures that subsequent modeling is not biased toward one class and allows fair evaluation of classification performance.

4.2 Word count summary by class

Label	Mean	Std	Min	Max
Real	389	270	4	5105
Fake	432	398	0	8057

Table 1: Summary statistics of word counts by class
)

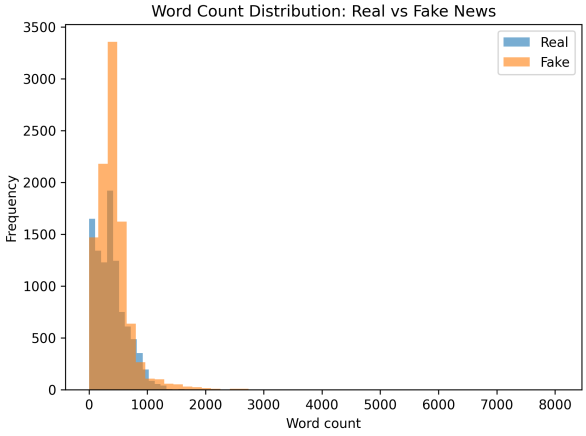

Figure 1: Word count distribution

On average, fake news articles tend to be slightly longer than real news articles (Table 1), as indicated by the mean word count. However, there is considerable variability in both classes, with some extremely short or very long articles (Figure 1). Understanding these length distributions is important for feature engineering and model design, as text length can influence classification performance.

4.3 Sentiment distribution (VADER compound)

Label	Negative	Neutral	Positive
Real	4602	181	5217
Fake	5478	170	4352

Table 2: Distribution of sentiment polarity buckets by class

The sentiment analysis shows that fake news tends to exhibit slightly more negative sentiment compared to real news. Neutral articles are rare in both classes. These differences suggest that sentiment-based trading models or analyses could be biased if fake news is not accounted for. Additionally, stylistic differences such as excessive capitalization, punctuation usage, and lexical diversity may systematically influence sentiment scores, emphasizing the need for careful feature engineering in predictive models.

5 Sentiment Analysis

Sentiment analysis was performed by computing VADER compound scores for each news article. These scores were then categorized into negative, neutral, and positive buckets to facilitate comparison between classes. The resulting distributions were visualized and saved as `sentiment_bucket_distribution.png` and `sentiment_distribution.png`.

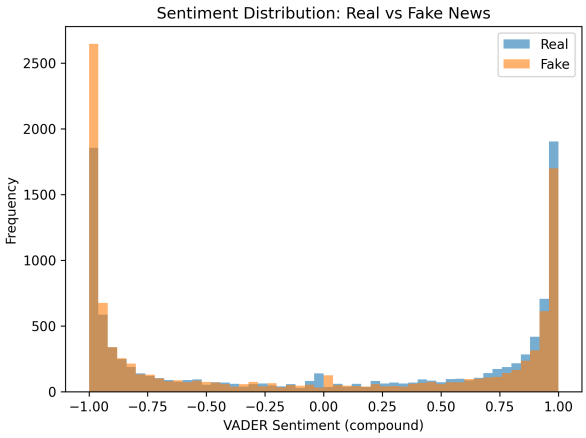

Figure 2: Sentiment distribution

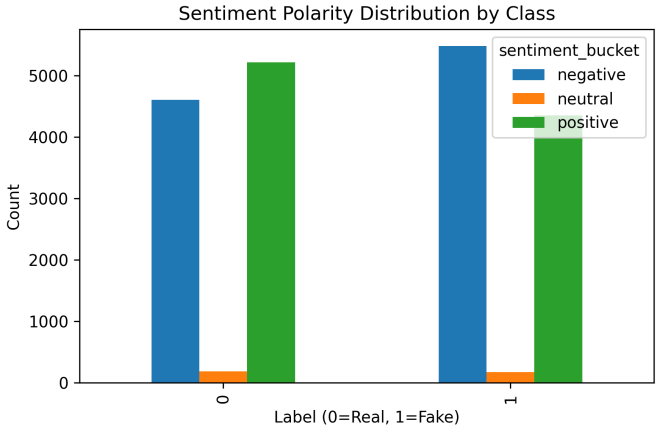

Figure 3: Sentiment bucket distribution

Fake news often contains stronger negative sentiment (Table 2, Figures 2 and 3), more exclamations, and unusual formatting, which could distort sentiment-driven trading models if unaccounted for [6].

6 Feature Engineering

For the machine learning models, textual content was transformed into numerical representations using TF-IDF [4]. The TF-IDF features captured the importance of terms across the corpus, considering both unigrams and bigrams, while limiting the vocabulary to the top 10,000 terms. Common English stop words were removed to reduce noise and emphasize meaningful words and phrases.

In addition to the TF-IDF features, the previously computed numeric stylistic features were incorporated into the dataset. These included counts of exclamation and question marks, the ratio of uppercase letters, average sentence length, and lexical diversity. Combining both textual and stylistic features enabled the models to leverage not only the content but also the writing style differences between fake and real news articles [2].

7 Topic Modeling

To explore thematic differences between fake and real news, two approaches were applied. The first was a manual examination of the most frequent words in each class, while the second approach employed Latent Dirichlet Allocation (LDA) to extract latent topics systematically [5].

For the LDA analysis, five topics were generated for each class. In fake news articles, the top words across topics included *trump, video, clinton, obama, hillary, people, just, president, like, said*, reflecting political content and sensational reporting. Other topics highlighted local news events and crises (e.g., flint water crisis), prominent political figures (e.g., pelosi, jeanine, mccrory), and media-specific phrases (e.g., boiler room, wire, fbi, mueller). A fifth topic included less frequent but distinctive words such as *bundy, finicum, oregon, refuge*, indicating coverage of fringe or extreme events.

In real news articles, LDA topics emphasized factual reporting and broader news coverage. The top words across topics included *said, trump, republican, house, reuters, tax, senate, court, percent, president*, highlighting standard political reporting. Other topics covered international events (e.g., puerto rico, zuma, kremlin), global affairs and statements (e.g., gulen, coup, edited, trump twitter), European politics (e.g., eu, turkey, brexit, merkel, kurdish), and security issues (e.g., china, korea, north korea, iran, united states).

The topics discovered by LDA aligned well with manual word frequency analysis, revealing that fake news tends to focus on sensationalized, emotionally charged, or politically polarized content, whereas real news covers broader, fact-based reporting. All LDA topics were saved in `lda_fake_topics.txt` and `lda_real_topics.txt` for reference and further inspection.

8 Classification

The balanced dataset of 20,000 samples was split into a training set of 16,000 samples and a test set of 4,000 samples. This split ensured that both classes (fake and real news) were equally represented in training and evaluation [4].

For text-based classification, TF-IDF vectorization was applied to the cleaned content. The top 10,000 terms were selected, including both unigrams and bigrams, while English stop words were removed to focus on informative words. A Logistic Regression classifier was then trained on these TF-IDF features.

The performance of the Logistic Regression model on the test set is summarized in Table 3. The classifier achieved an accuracy of 0.984, precision of 0.990, recall of 0.978, F1-score of 0.984, and a ROC-AUC of 0.999, demonstrating excellent separability between fake and real news articles.

Metric	**Value**
Accuracy	0.984
Precision	0.990
Recall	0.978
F1-score	0.984
ROC-AUC	0.999

Table 3: Logistic Regression + TF-IDF classification performance on test set

The confusion matrix in Figure 4 illustrates that the model correctly identified most real and fake news articles, with only a small number of misclassifications.

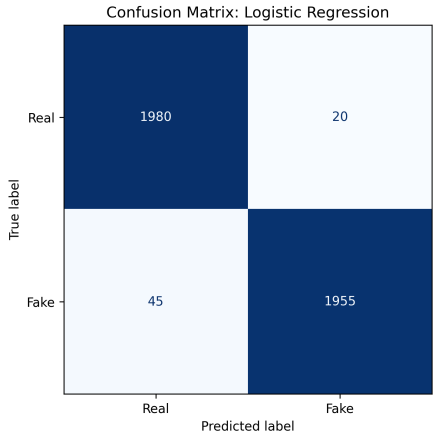

Figure 4: Confusion matrix of Logistic Regression classifier

Examining the model coefficients, the top 20 words most indicative of fake news included terms such as *video, obama, hillary, trump, image,* and *president trump*, reflecting sensational and politically charged content. For real news, words such as *reuters, said, washington, president donald, republican,* and *minister* were most predictive, highlighting factual reporting patterns.

To compare performance, a LinearSVC classifier was trained on the same TF-IDF features. As shown in Table 4, LinearSVC slightly outperformed Logistic Regression, achieving an accuracy of 0.992 and an

F1-score of 0.991, confirming the strong separability of classes in the feature space.

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.984	0.990	0.978	0.984
LinearSVC	0.992	0.994	0.989	0.991

Table 4: Comparison of Logistic Regression and LinearSVC classifiers

Additionally, 5-fold stratified cross-validation was performed for Logistic Regression, resulting in a mean F1-score of 0.987 and mean ROC-AUC of 0.999, further validating the robustness of the model. Finally, calibration using `CalibratedClassifierCV` indicated that predicted probabilities closely matched observed frequencies, confirming that the model's probability estimates are well-calibrated and suitable for downstream risk-sensitive applications.

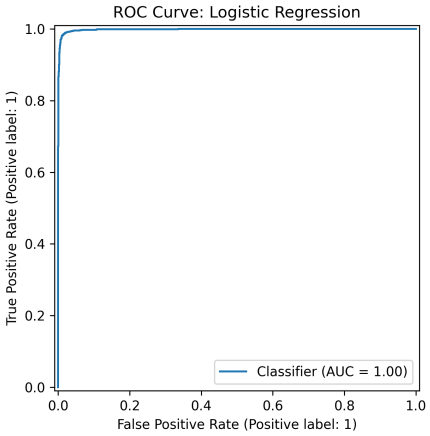

Figure 5: ROC curve

Overall, these results demonstrate that TF-IDF-based classifiers, even with relatively simple linear models, can effectively distinguish between fake and real news, with minimal misclassification and highly informative predictive features.

9 Conclusion

The analysis demonstrates that fake news differs systematically from real news in multiple dimensions, including sentiment, stylistic features, and vocabulary. Fake news articles tend to exhibit more negative sentiment, unusual formatting such as excessive capitalization or punctuation, and a higher prevalence of emotionally charged or sensational words. In contrast, real news shows more neutral or fact-based language patterns.

TF-IDF based classifiers, including Logistic Regression and LinearSVC, achieved very high performance on the balanced dataset, with F_1 scores exceeding 0.98, indicating that fake and real news are clearly separable based on textual content. The observed stylistic and sentiment differences underline potential biases for sentiment-driven trading models or other automated systems that rely on textual analysis, as these models could be misled by characteristics typical of fake news.

Finally, topic modeling using Latent Dirichlet Allocation revealed distinct thematic content between fake and real news. Fake news topics emphasized sensational, politically polarized, or extreme events, while real news topics focused on standard political reporting, international affairs, and broader factual coverage. Overall, the combination of statistical, machine learning, and natural language processing methods provides a comprehensive understanding of the structural and thematic distinctions between fake and real news.

10 Documentation

All code and data used in this project are publicly available in the following GitHub repository:

`https://github.com/SzimonaSzternak/news_and_market_szszt24`

The repository contains:

- Python script for data loading, preprocessing, feature engineering, sentiment analysis (VADER), TF-IDF vectorization, Logistic Regression and LinearSVC classification, cross-validation, calibration, and topic modeling (LDA).
- Resulting outputs, including the processed dataset with sentiment and stylistic features, top words for fake and real news, LDA topics, and model performance metrics.
- Plots illustrating exploratory data analysis (EDA), sentiment distribution, word count distribution, confusion matrices, ROC curves, and calibration curves.
- Original datasets: `Fake.csv` and `True.csv` from Kaggle [3].

References

[1] Weisberg, S. (2014). *Applied Linear Regression, Fourth Edition*. Wiley.

[2] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media. [Used for VADER sentiment analysis]

[3] Bisaillon, C. (2018). *Fake and Real News Dataset*. Kaggle. `https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset/data`

[4] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

[5] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3, 993–1022.

[6] Hutto, C. J., & Gilbert, E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. Proceedings of the International AAAI Conference on Web and Social Media, 8(1).

[7] Vedels, Christian. "News and Market Sentiment Analytics – GitHub repository with course codes and lectures." `https://github.com/christianvedels/News_and_Market_Sentiment_Analytics/tree/main`, 2025.

[8] OpenAI. "ChatGPT (GPT-4/5) - Generative AI assistance for code and report writing." `https://openai.com/`, 2025.

[9] GitHub Copilot. "GitHub Copilot – AI-powered coding assistant integrated into Visual Studio Code." `https://github.com/features/copilot`, 2025.