Fake News Detection

Minju Bae, James Bruggink, Juncheng Long, Eric Sun, Leo Xiong

Word Count: **587** (excluding titles and post contributions)

Introduction:

In this project, we investigate the problem of detecting fake news using supervised machine learning techniques. Our dataset, sourced from Kaggle, contains data about news articles, including their authenticity, political alignment, topic, source, state of origin, and other relevant features. The core goal of our analysis was to answer four central questions: (1) Which U.S. states might be more associated with fake news? (2) Are certain topics more likely to be fake or real? (3) Does fake news spike during political events? (4) Which machine learning model best detects fake news?

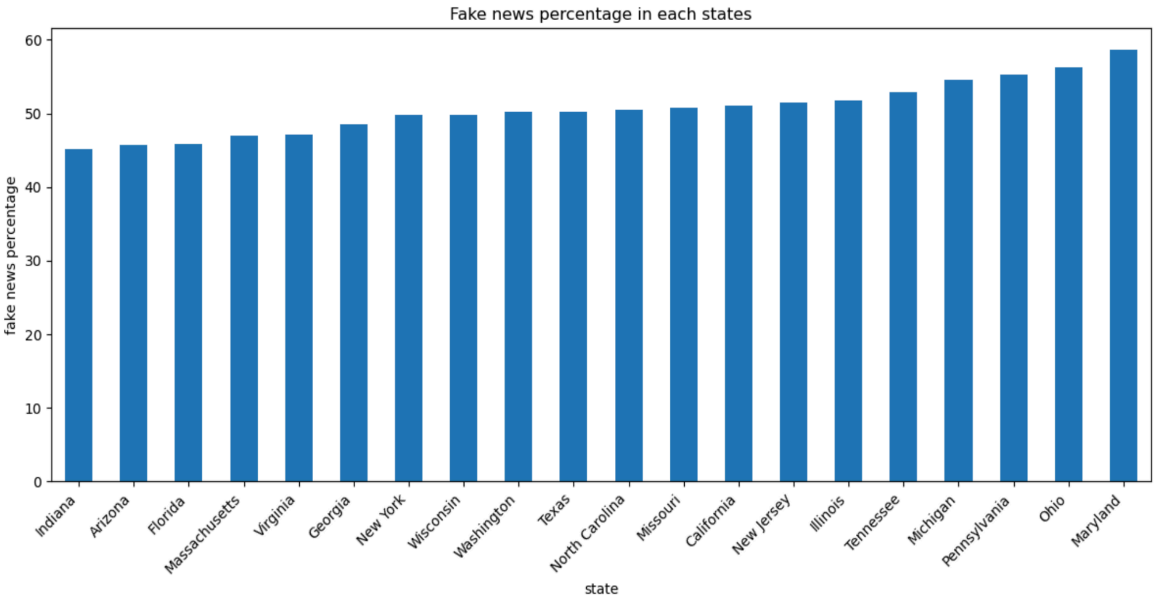
To tackle this, we pre-processed the data by hand-selecting suitable features, trained several classification models, tuned hyperparameters, and analyzed feature importance. We concluded that a Support Vector Machine (SVM) with an RBF kernel and regularization parameter C=0.1 performed best. Below, we describe our methods and findings in detail.

Data and Preprocessing:

The dataset contains 4,000 rows and 24 columns, including fields such as **state**, **topic**, **source**, **source_reputation**, **political_bias**, and a **fake-real** label, which we changed to a binary variable (0-1). Initial exploration revealed that many features were either redundant or unsuitable for direct use. For example, text-heavy fields like title and text were excluded due to the complexity of natural language processing. We kept only one feature from similar pairs like **source** vs. **source_reputation**, and **word_count** vs. **char_count** to reduce overfitting.

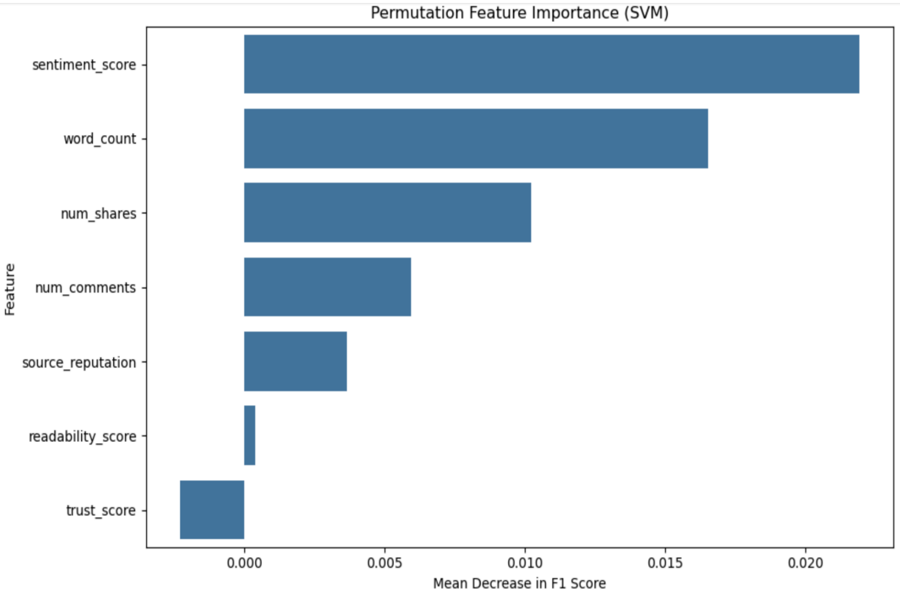
We retained only numerical and encoded features to suit linear models, such as Logistic Regression and SVM, which typically perform better with numerical inputs. Categorical variables like state and topic were encoded using target encoding or simply dropped, depending on their relevance and distribution. Then, the data was split into 80% training, 10% validation, and 10% test sets using train_test_split.

To explore state-level differences in fake news, we calculated the **percentage of fake news articles per state**, rather than raw counts, to account for varying sample sizes across states. This normalization ensures fairer comparisons.



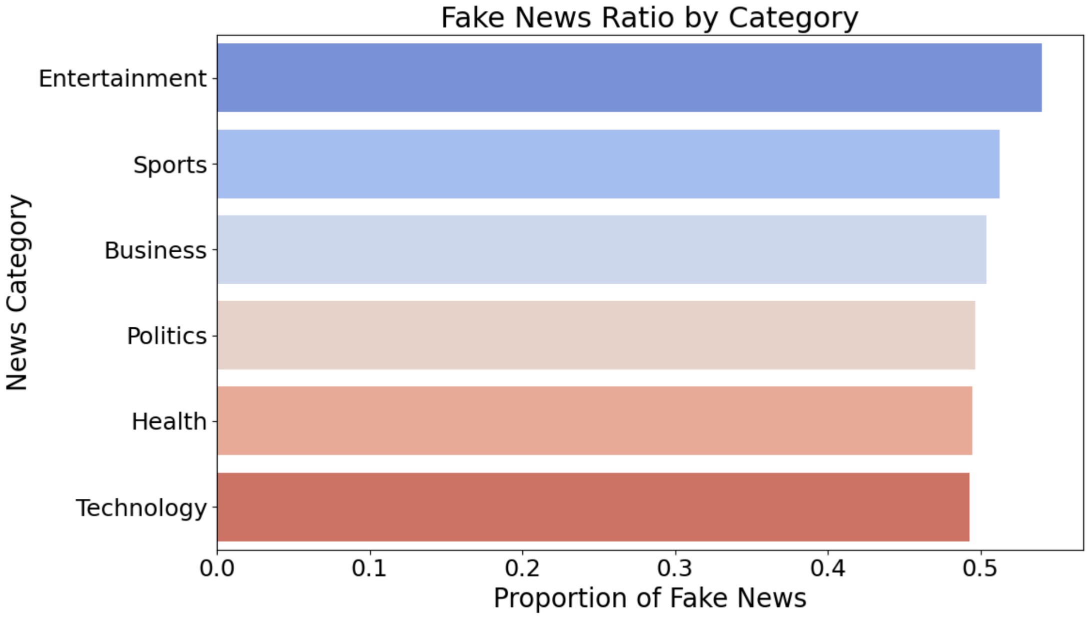
Feature Importance and Final Evaluation

To understand feature contributions, we applied permutation importance on the final SVM model. Features like **sentiment_score**, **word_count**, and **num_share** showed high influence in classification. These findings suggest that emotional tone, article length, and online engagement play major roles in distinguishing fake from real news from our dataset.

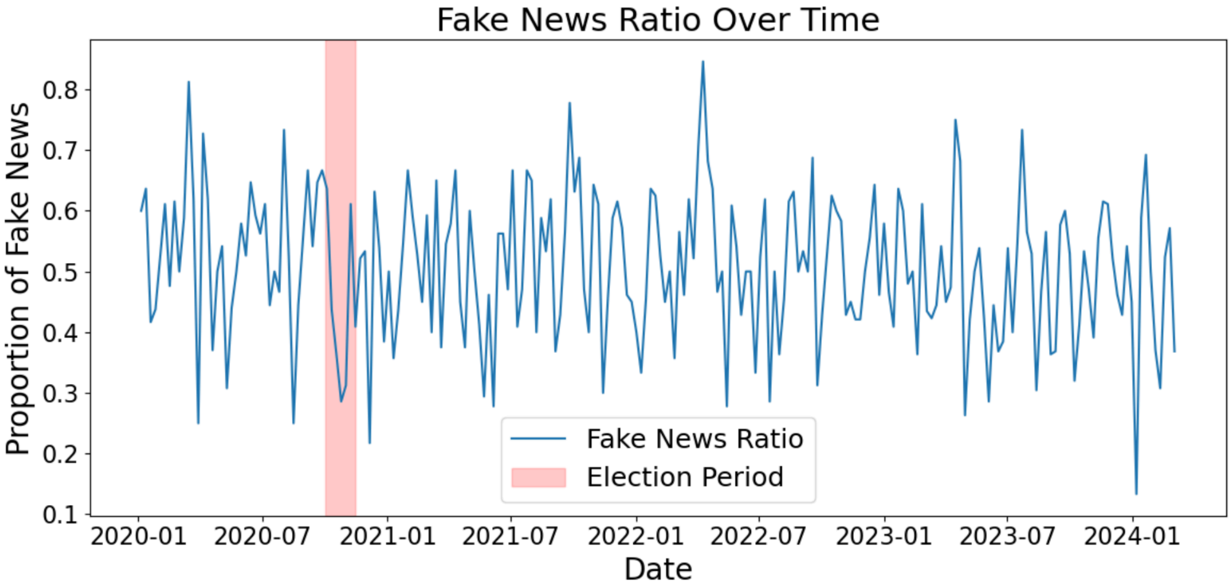


Additional analyses

We investigated which topics were more likely to be fake. We grouped the data by category and calculated the fake news ratio. The results indicated that Entertainment, Sports, and Business had higher fake news ratios (0.54, 0.51, and 0.50, respectively), but these differences were not significantly greater compared to other categories.



Furthermore, we examined whether the frequency of fake news was associated with political events, specifically the US presidential election. We compared the election period (one and a half months before and after the election) with other periods. Surprisingly, the fake news ratio during the election period was lower (0.42) compared to other periods (0.51) in the dataset.



Conclusion and Future work

This project demonstrated that supervised machine learning models can be utilized to answer our four central research questions, but at the cost of low precision. The model consistently predicted fake news as real and vice-versa, while maintaining a high recall score. These results highlighted that articles were often misclassified, suggesting limitations in the dataset and possible model overfitting. To combat these limitations, we could incorporate textual features using natural language processing (e.g., TF-IDF, BERT embeddings), using a more informative dataset, and applying ensemble methods like Random Forests or Gradient Boosting to further improve precision, accuracy and robustness. Moreover, longitudinal studies could explore how fake news patterns evolve over time.

Post-conclusion Contributions

Member	Proposal	Coding	Presentation	Report	Total
Minju Bae	5	5	5	5	20/20
James Bruggink	5	5	5	5	20/20
Juncheng Long	5	5	5	5	20/20
Eric Sun	5	5	5	5	20/20
Leo Xiong	5	5	5	5	20/20

Notes:

- 5 = Full contribution, 4-2 = Partial Contribution, 1 = Minimal Contribution
- We all cooked, some of us started slow due to lack of communication, but we finished strong in the end as we made it up in the presentations and research, along with the report.