Enhancing Fake News Detection: The Role of Ensemble Learning, Feature Engineering, and Language Models

Orel Jaraian ; Ayelet Hashahar Cohen ; Lior Vaknin ; Tzuf Lahan

June 2024

1 Abstract

This study investigates fake news dissemination on social media, aiming to improve prediction accuracy through feature engineering and ensemble learning models. A comprehensive review of existing detection methods highlights the integration of content-based and context-based approaches for enhanced accuracy. The study underscores the necessity for extensive datasets and robust evaluation metrics, proposing future research directions focused on advancing detection algorithms and leveraging new social media features. Employing ensemble learning techniques, this research aims to improve fake news detection by evaluating various models and their combinations, emphasizing the efficacy of techniques such as voting classifiers with parameter adjustments. This investigation reveals that while traditional content analysis methods provide a foundation, incorporating advanced machine learning models and social context significantly augments the reliability of detection systems, paving the way for more effective countermeasures against misinformation.

2 Introduction

Fake news is false or misleading information presented as news. With the rise of social media and online platforms, fake news has become a big problem. It can influence people's opinions, political decisions, and society. Detecting fake news is important to ensure people get accurate information and can make good decisions. Fake news detection involves using different methods to identify and classify false information. These methods include natural language processing (NLP), machine learning (ML), and artificial intelligence (AI). Researchers use various techniques to analyze the text, check the credibility of sources, and determine if news articles are true or false.

Our research will focus on ensemble learning methods to solve the problem of fake news detection. Our goal is to compare the ensemble learning models

that performed so far and additional ensemble versions for the selected data set. For example, we will perform additional combinations for voting classifier models and adjust parameters such as soft vs. hard voting.

3 Background / Related Work

The Current State of Fake News: Challenges and Opportunities

Author: Figueira, Á., & Oliveira, L. (2017)

Overview

The paper investigates the pressing issue of fake news in the digital age, particularly on social networks. The authors categorize various approaches to mitigate fake news, focusing on content-based, source-based, and diffusion-based methods. They also propose an algorithmic solution aimed at detecting fake news and discuss the challenges and opportunities associated with this endeavor.

Methodology

The paper reviews existing literature on fake news detection and classifies the approaches into three categories: content-based, source-based, and diffusion-based. It also examines specific cases and methodologies used by companies like Facebook, alongside innovative student-developed solutions such as the FiB system.

Results & Discussions

The authors highlight several significant findings:

1. **Content-Based Approaches:** Utilize text analysis to identify fake news by examining linguistic features and patterns.
2. **Source-Based Approaches:** Evaluate the credibility of the source to determine the likelihood of the information being fake.
3. **Diffusion-Based Approaches:** Analyse the spread patterns of information across networks to detect anomalies indicative of fake news.
4. **Facebook's Approach:** Combines human classification with machine learning to identify and reduce the spread of fake news on its platform.

Research Gap or Area of Improvement

The paper identifies several areas for further research:

- **Algorithmic Robustness:** Current systems lack the necessary robustness for reliable detection across diverse contexts.

- **Scalability:** Ensuring that detection mechanisms can handle the vast amounts of data generated on social networks.
- **Integration of Approaches:** Combining content, source, and diffusion-based methods to enhance detection accuracy.

A Systematic Review on the Detection of Fake News Articles

Authors: Hoy, N., & Koulouri, T. (2021)

Overview

The paper provides a comprehensive analysis of current methodologies and technologies used to detect fake news, focusing on the period from 2016 to 2020. It examines various machine learning models and feature extraction techniques employed in the detection process, as well as evaluates the effectiveness of these methods.

Methodology

The authors conducted a systematic review using automated and manual search processes, guided by specific research questions. They focused on features used in fake news detection, approaches for feature selection/extraction, machine learning models employed, and datasets utilized. Papers from computer science journals and conferences were included, applying strict inclusion and exclusion criteria to ensure relevance and quality.

Results & Discussions

The review highlights several key findings:

1. **Feature Types:** Lexical, syntactic, semantic, and pragmatic features are commonly used for detecting fake news. These features help in identifying patterns and structures typical of fake news content.
2. **Feature Selection/Extraction Approaches:** Techniques like Term Frequency-Inverse Document Frequency (TF-IDF), word embeddings (e.g., Word2Vec, GloVe), and deep learning-based methods (e.g., BERT) are prominent.
3. **Machine Learning Models:** Supervised learning models, especially those based on neural networks such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), are widely used. There is also mention of unsupervised and semi-supervised approaches.
4. **Datasets:** Publicly available datasets like LIAR and BuzzFeed News are frequently used to train and evaluate models.

Research Gap or Area of Improvement

The review identifies several areas for further research:

- **Algorithmic Robustness:** Current detection systems lack robustness across diverse contexts.
- **Scalability:** Ensuring that detection mechanisms can handle large volumes of data generated on social networks.
- **Integration of Approaches:** Combining different detection methods (content-based, source-based, and diffusion-based) to improve overall accuracy and reliability.

The Current State of Fake News: Challenges and Opportunities

Authors: Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu

Overview

The article explores the significant issue of fake news on social platforms, examining its rapid spread and the role of psychological biases and social dynamics. It reviews existing detection methods, emphasizing the integration of content-based and context-based approaches for better accuracy. The need for comprehensive datasets and improved evaluation metrics is highlighted, along with future research directions to enhance detection algorithms and explore new social media features.

Methodology

- **Feature Extraction:** Identifies relevant characteristics from the news content (linguistic and visual elements) and social media context (user interactions, posts, and network dynamics).
- **Model Construction:** Develop machine learning models using both content and social context features to differentiate fake news from real news.

Results & Discussions

The authors highlight several significant findings:

1. **Combined Features:** Integrating content-based and context-based features significantly improves detection accuracy.
2. **Dataset Challenges:** The lack of large-scale, comprehensive datasets hinders robust model development.
3. **Temporal Analysis:** Early detection focusing on initial spread patterns is effective.

4. **Psychological Factors:** Models need to account for human psychological biases like confirmation bias and the echo chamber effect.

Research Gap or Area of Improvement

The paper identifies several areas for further research:

- **Comprehensive Datasets:** Need for large-scale, diverse, well-annotated datasets with both news content and social context features.
- **Psychological Factors Integration:** Current models should better incorporate psychological biases affecting fake news spread.

Fake News Detection: An Ensemble Learning Approach

Authors: Arush Agarwal and Akhil Dixit

Overview

The paper discusses the challenges and methodologies for detecting fake news, emphasizing the need for accurate identification due to the societal harm of misinformation. The authors propose an ensemble model using SVM, CNN, LSTM, KNN, and Naive Bayes to improve detection accuracy. The study highlights the effectiveness of combining multiple classifiers to enhance predictive power and the importance of feature extraction and data preprocessing.

Methodology

- **Ensemble Classifier:** Utilizes multiple base classifiers (SVM, CNN, LSTM, KNN, and Naive Bayes) trained on two datasets for diverse and comprehensive evaluation.
- **Feature Extraction:** Analyse the context of short sentences and news to produce credibility scores for both the news and the author.

Results & Discussions

The authors highlight several significant findings in their study of fake news detection using an ensemble learning approach:

1. **Model Performance:** The ensemble model showed superior performance, with the LSTM model achieving the highest accuracy at 97
2. **Effectiveness of Deep Learning:** Deep learning models (CNN and LSTM) outperformed traditional models (SVM and KNN). CNN achieved 94
3. **Dataset Evaluation:** The combined dataset from Liar and Kaggle provided a robust basis for model evaluation, with preprocessing ensuring balanced and clean data.

4. **Precision, Recall, and F1-Score:** LSTM achieved the highest scores across these metrics, highlighting its reliability in accurately classifying fake news.

Research Gap or Area of Improvement

The paper highlights the need for further research in:

- **Ensemble Model Optimization:** While the ensemble approach has shown promise, further exploration is needed to optimize the combination of different algorithms.
- **Real-time Detection Capabilities:** Developing real-time detection systems is crucial for mitigating the rapid spread of fake news.
- **Advanced Feature Extraction Methods:** Improved methods for extracting features from textual and visual content are necessary.
- **Dataset Diversity and Quality:** More comprehensive and diverse datasets are required to better capture the complexity of fake news.

Fake News Detection Using Ensemble Techniques

Authors: Malhotra, P., & Malik, S. K. (2023)

Overview

This paper evaluates the effectiveness of various ensemble learning models in detecting fake news, assessing algorithms like SVM, logistic regression, CatBoost, XGBoost, multinomial Naive Bayes, and random forest. Using a dataset of around 40,000 news articles (half fake, half real), the study measures performance based on accuracy, precision, recall, F1 score, and other metrics.

Methodology

The study applies multiple machine learning algorithms to classify news as fake or real, evaluating models on a dataset of around 40,000 articles. Metrics include accuracy, precision, recall, F1 score, and false rejection rate. The study also explores hybrid models and a deep learning model (AutoViML) to determine the most effective approach for fake news detection.

Results & Discussions

Key findings include:

1. **Model Performance:** The deep Auto_ViML model achieved the highest accuracy, precision, recall, and F1 score of 99
2. **Hybrid Models:** Combining various models resulted in improved performance, particularly in terms of false rejection rate.

3. **Support Vector Machine:** Noted for its computational efficiency, with a computation time of 0.000245 seconds.

Research Gap or Area of Improvement

The paper suggests several areas for further research:

- **Real-time Detection:** Enhancing models to perform real-time fake news detection.
- **Scalability:** Ensuring that detection mechanisms can handle large volumes of data.
- **Cross-domain Application:** Testing models across different domains and languages to improve robustness.

Fake News Detection Using Machine Learning Ensemble Methods

Authors: Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020)

Overview

The paper explores the use of machine learning ensemble methods to detect fake news. The study leverages various textual properties and evaluates different machine-learning algorithms combined through ensemble methods to improve detection accuracy.

Methodology

The authors used machine learning algorithms, including logistic regression, support vector machines (SVM), and multilayer perceptron (MLP), in conjunction with ensemble methods such as bagging, boosting, and voting classifiers. They trained and tested these models on four real-world datasets, utilizing features extracted through the Linguistic Inquiry and Word Count (LIWC) tool.

Results & Discussions

The experimental evaluation showed that ensemble methods outperformed individual learners. Bagging classifier (decision trees) and boosting classifier (XGBoost) achieved the highest accuracies, with XGBoost showing superior performance across various metrics. The study demonstrated that combining multiple algorithms enhances the robustness and reliability of fake news detection systems.

Research Gap or Area of Improvement

The paper highlights the need for further research in:

- Real-time fake news detection.
- Enhancing the robustness of algorithms to handle diverse news domains.
- Exploring hybrid approaches that integrate social context with textual features.

4 Methodology

4.1 Data Collection

We examined datasets published in several articles focusing on the use of Ensemble Methods for identifying fake news. Among the various datasets, we decided to use the dataset that yielded the highest variance among the different models tested.

The chosen dataset is available on Kaggle [7](hereafter referred to as DS) and contains a total of 20,386 articles used for training and 5,126 articles used for testing. This dataset is compiled from multiple sources on the Internet and includes articles from diverse domains beyond politics, encompassing both fake and true articles from various other fields.

4.2 Data Cleaning

We handled missing values by deleting rows with missing text (39 rows) or missing author (1,975 rows), as we concluded this would not hurt performance. The 'title' column's missing values were filled using a summarization model to generate appropriate content. We removed 120 duplicate records to ensure uniqueness. Data types were corrected: 'id' was set to an integer, 'title', 'author', 'text' to a string, and 'label' to an integer. Text data inconsistencies, such as different spellings or formatting issues, were standardized. Before cleaning, there were 25,512 records with 45 missing titles, 1,975 missing authors, 39 missing texts, and 120 duplicate records. After cleaning, we had 23,498 records with no missing values or duplicates and corrected data types. This process ensured a reliable dataset for further analysis and modeling.

4.3 Exploratory Data Analysis (EDA)

From our analysis, we can infer several statistical views about the distribution of numeric columns, correlation between them, and other statistical information:

1. **Text and Title Word Count**:
    - The distributions of `text_word_count` and `title_word_count` are highly skewed, with most articles having relatively low word counts but a few outliers with very high counts.

- Box plots indicate a slight difference between fake and real news in terms of word count, though it may not be significant.

Insight: Normalizing or log-transforming these features could help reduce skewness and improve model performance. Word count alone might not be a strong indicator, so combining it with other features could be beneficial.

2. **Unique Word Count**:
   - Both `text_unique_word_count` and `title_unique_word_count` show skewed distributions with outliers.
   - There is a slight difference in distributions for fake and real news, suggesting that fake news might use a less diverse vocabulary.

   Insight: The uniqueness of words in text and titles could be a useful feature. Creating ratios of unique to total words might provide additional insights.

3. **Punctuation and Sentiment**:
   - `text_punctuation_count` and `title_punctuation_count` are significantly skewed, with some outliers.
   - Sentiment (`text_sentiment`) distribution is relatively normal, with fake news having a slightly lower average sentiment.

   Insight: Punctuation counts, especially extreme values, might indicate fake news. Sentiment analysis could also provide valuable insights, as fake news might have different emotional tones compared to real news.

4. **Named Entity Count and Capital Ratio**:
   - `named_entity_count` shows a wide range of values with several outliers.
   - `capital_ratio` is generally low for most articles but has some higher outliers.

   Insight: Named entity counts could indicate the credibility of an article, as real news might mention more named entities (people, organizations, etc.). High capital ratios might suggest sensationalism, a trait of fake news.

Recommendations:

- Utilize correlation analysis or feature importance techniques (e.g., using a Random Forest or Gradient Boosting model) to identify the most predictive features and reduce dimensionality.
- Experiment with different algorithms, such as ensemble methods (Random Forest, Gradient Boosting), that can handle feature interactions and outliers effectively.

- Consider capping outliers or using robust scaling methods to manage extreme values in features like word count, punctuation count, and named entity count.
- Create new features such as the ratio of unique words to total words, the ratio of named entities to total words, and combined sentiment scores.

4.4 Feature Engineering

4.4.1 Stopwords Removal

Reason: Removing stopwords helps in reducing the noise in the data and focuses on the significant words that carry the actual meaning of the text.

Justification: Fake news articles often use exaggerated or specific vocabulary to attract attention. By removing stopwords, we can better analyze the meaningful content and identify patterns that might indicate deception.

4.4.2 Word Count Features

text_word_count and title_word_count: Count the number of words in the text and title.

text_unique_word_count and title_unique_word_count: Count the number of unique words in the text and title.

Reason: The length and uniqueness of the text can be indicative of the article's nature.

Justification: Fake news articles might be shorter and repetitive, aiming to convey a sensational message quickly, whereas legitimate articles might provide more detailed and varied content.

4.4.3 Punctuation Count Features

text_punctuation_count and title_punctuation_count: Count the number of punctuation marks in the text and title.

Reason: Excessive punctuation can be a sign of emotional manipulation or sensationalism.

Justification: Fake news often uses more punctuation marks (like exclamations) to create a sense of urgency or excitement, which can help in distinguishing them from real news.

4.4.4 Sentiment Analysis

text_sentiment: Measures the polarity of the text (how positive or negative it is).

Reason: Sentiment can provide insights into the tone and emotional appeal of the article.

Justification: Fake news may exhibit extreme sentiment (either highly positive or negative) to provoke strong reactions from readers, while real news tends to be more neutral and objective.

4.4.5 Named Entity Recognition (NER)

named_entity_count: Counts the number of named entities in the text.

Reason: The presence and frequency of named entities (such as people, organizations, locations) can be indicative of the article's depth and reliability.

Justification: Real news articles often mention specific names and places, providing verifiable information. Fake news might lack such details or provide fabricated entities.

4.4.6 Exclamation and Question Marks

exclamation_count and question_count: Count the number of exclamation and question marks in the text.

Reason: The use of these punctuation marks can indicate sensationalism or attempts to engage the reader emotionally.

Justification: Fake news articles often use these marks excessively to create a sense of urgency or curiosity, encouraging readers to react quickly without verifying facts.

4.4.7 Capitalization Ratio

capital_ratio: Measures the proportion of capital letters in the text.

Reason: Excessive use of capital letters can indicate attempts to emphasize certain points or create a sense of alarm.

Justification: Fake news often uses all-caps to grab attention and emphasize certain aspects of the article, whereas real news typically follows standard capitalization rules.

4.5 Model Selection

We chose to examine a variety of models that differ in several aspects:

1. Use of different language models: We employed three models, including TF-IDF, BERT, and Word2Vec.

2. Integration of numerical features: Some of the models we tested are distinguished by the inclusion of the new features that we enriched the records with during the Feature Engineering stage.

3. Use of different machine learning models: We applied various methods, such as AdaBoost, Random Forest, and Neural Network.

5 Results - Evaluation

This table presents the evaluation metrics for five machine learning models applied to classify fake news. We chose to compare the accuracy, weighted average precision, and weighted average recall.

Model	Algorithm	Numeric features	Text model	Accuracy	Precision	Recall
1	Random Forest	X	TF-IDF	0.64	0.65	0.64
2	Neural Network	X	Bert	0.63	0.64	0.64
3	Random Forest	V	TF-IDF	0.64	0.65	0.64
4	Random Forest	V	X	0.61	0.62	0.61
5	AdaBoost	V	X	0.64	0.67	0.65

Table 1: Model performance comparison

6 Discussion

Compared to the academic paper that also investigated the problem using the same dataset, we evaluate the models by multiple metrics, whereas the paper only compared accuracy. The performance achieved in the paper for the Random Forest model was significantly lower (0.35) while their performance for AdaBoost exceeded ours, reaching 0.92. The discrepancies could stem from various factors, such as the type of language model chosen, additional variables considered, and more.

In analyzing the performance, we found the following main conclusions:

1. **Highest Precision and Recall:** Model 5 (AdaBoost) without numeric features but using TF-IDF achieves the highest precision (0.67) and recall (0.65), indicating it has a better balance in correctly identifying both fake and real news.

2. **Impact of Numeric Features:** The presence of numeric features doesn't consistently improve performance. For instance, Model 1 and Model 3 with numeric features have similar performance as Model 5 without numeric features.

3. **Text Modeling Techniques:** The use of advanced text modeling like BERT in Model 2 does not significantly outperform simpler techniques like TF-IDF used in Models 1 and 3.

In conclusion, Model 5's combination of the AdaBoost algorithm and the exclusion of numeric features with a simple text model appears most effective in this dataset for classifying fake news.

7 Future Work

While our study explored a variety of models and approaches to classify fake news, there are several avenues for future research that could potentially enhance the performance and robustness of fake news detection systems.

One promising direction is the utilization of more powerful and advanced language models. During our experiments, we attempted to run models such as GPT and RoBERTa, which are known for their state-of-the-art performance in

various NLP tasks. Unfortunately, these runs were interrupted due to resource constraints.

Further research could focus on:

1. **Utilization of Advanced Language Models:** Overcoming the resource limitations to successfully implement and evaluate GPT and RoBERTa. These models could provide significant improvements in understanding and classifying nuanced language patterns in fake news.

2. **Integration of More Features:** Exploring the inclusion of additional numerical and categorical features that may contribute to better model performance. This could involve gathering more diverse data sources and performing comprehensive feature engineering.

3. **Real-Time Fake News Detection:** Investigating the feasibility of deploying these models in real-time scenarios to promptly identify and flag fake news articles as they are published.

By addressing these areas, future research can build on our findings and contribute to the development of more accurate and reliable fake news detection systems.

8 References

References

[1] Figueira, Á., & Oliveira, L. (2017). The Current State of Fake News: Challenges and Opportunities. *ScienceDirect*.

[2] Hoy, N., & Koulouri, T. (2021). A Systematic Review on the Detection of Fake News Articles. Retrieved from.

[3] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). The Current State of Fake News: Challenges and Opportunities. *ScienceDirect*.

[4] Agarwal, A., & Dixit, A. (2020). Fake News Detection: An Ensemble Learning Approach.

[5] Malhotra, P., & Malik, S. K. (2023). Fake News Detection Using Ensemble Techniques.

[6] Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake News Detection Using Machine Learning Ensemble Methods.

[7] Kaggle, T and others (2016). Getting Real about Fake News.