Classifying the Political Alignment of News Articles Using Machine Learning

Word Count: 2,650. Registration number: 240175739

**1. Introduction**

The underlying political ideology of the news can play a decisive role in influencing public opinions and political discussions. Research focused on the political alignment of news sources can give context to the public of where that news source is placed in terms of the political spectrum, it can also shed light on common language patterns or frequent wording that might be betrayed by the news sources. Confirmation bias is explained through the human tendency to seek, interpret and recall information in a way that confirms pre-existing beliefs (R. S. Nickerson, 1998). This may influence the way journalists display the news headlines, by unconsciously expressing their own political ideology through the events they are reporting. This political influence may be too subtle or not obvious for humans to understand, therefore research into training a machine model to complete this task can address this issue.

Society in the UK and USA have shown to have significant trust issues in the news headlines, with the UK reporting bottom in the Edelman Trust Barometer out of 28 countries in 2023, with only 31% having trust in the news. Similarly, in the USA only 32% had trust in the news “most of the time” Press Gazette. (2024). Whether a lot of this information expressed in news headlines should be questionable if there’s an inherent bias within them. Therefore, research into this topic can help society gain media trust, which defines as the audience’s readiness to rely on news content based on the belief that the media will fulfil its role effectively and satisfactorily (Hanitzsch et. al, 2017).

A 2018 study employed Granger causality and linear regression to reveal a causal relationship between Russia's economic downturn and increased US-focused news coverage in state-controlled media (A. Shupta et. al, 2023). Using logistic regression and normalized pointwise mutual information (nPMI), researchers identified five dominant frames to classify these narratives (A. Shupta et. al, 2023). Topic modelling further highlighted how propaganda reshaped public focus. This aligns with the statistical and machine learning approaches in this research, which also uncover framing strategies in political news classification.

A study on Spanish news sources applied machine learning to classify media frames during the 2015 refugee crisis, focusing on human rights and security narratives. Using Support Vector Machines (SVM) with a radial kernel, the researchers achieved effective classification after the linear kernel proved insufficient (J. Garcia-Marín and A. Calatrava, 2018). This study aligns with this research’s use of SVM to classify political ideologies in news articles, highlighting the importance of feature engineering and kernel selection in achieving accurate text classification. Lots of research into this area often focuses on social media data (N. Singh et. al, 2023, Tun, Y. M., & Khaing, M., 2023), whereas this research is based on online news article headlines so will expand existing literature.

The motivation for this research stems from the need to systematically identify and classify the political alignment of news articles using computational methods. By leveraging machine learning models and natural language processing (NLP) techniques, this study aims to automate the classification of news sources into political categories such as liberal, conservative and neutral. The ability to classify news articles based on their political alignment has significant implications for combating misinformation, promoting media literacy and fostering a balanced media ecosystem.

**1.2 Research Aim**

This research aims to develop a machine learning framework to classify online news articles by their political alignment (liberal, conservative or moderate) based on article titles. Using natural language processing (NLP) techniques and machine learning models such as Logistic Regression, Support Vector Machines (SVM) and Random Forest, this study seeks to:

1. Identify key terms and linguistic patterns indicative of political alignment.
2. Compare the performance of different machine learning models in terms of accuracy and reliability.
3. Automate the classification of news articles based on political bias.

**1.3 Research Questions**

1. How can the political alignment of online news articles be classified using machine learning models?
2. What linguistic features and patterns are most indicative of political bias in article titles?
3. Which machine learning model performs best for this classification task in terms of accuracy?

**2. Methodology**

**Data Preparation**

The study utilised the Zenodo news dataset, comprising 10,917 articles published in 2019. Articles were categorised as liberal, conservative or moderate based on their source's political alignment. The political alignment of the news articles was categorised aligned with Pew Research Centre (2020), which allowed many of the sources in the dataset to be used. Missing values were removed for reliability. The key columns focused on were specifically the news sources, which were allocated into the political ideology and the news headline (column label was ‘title’) which was then pre-processed using bag of words exploration. Bag of words was chosen in this research because it effectively quantifies the frequency of words across articles, providing a clear representation of the linguistic features that may indicate political alignment. Compared to more advanced techniques like word embeddings or transformers, bag of words is easier to preprocess and aligns well with traditional machine learning models like Logistic Regression, SVM, and Random Forest. Additionally, its compatibility with unigram-based approaches ensures straightforward feature extraction and interpretability of results, which was critical for identifying distinct linguistic markers in the dataset.

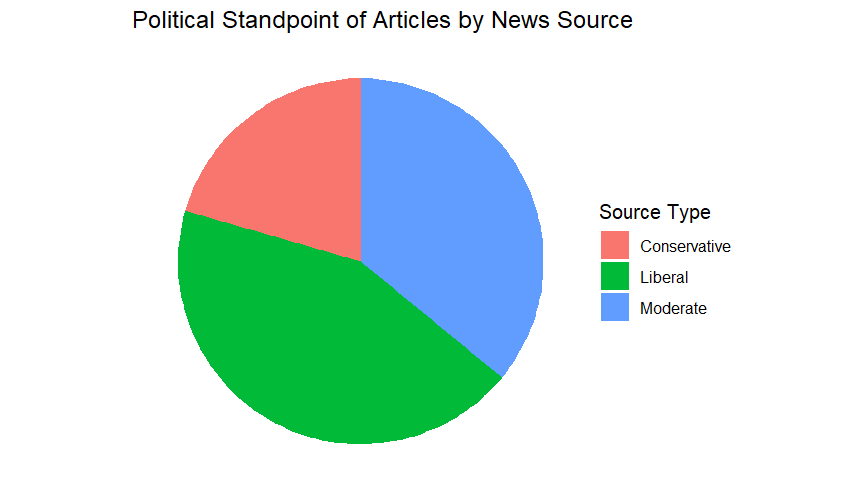
**Preprocessing:**

Text preprocessing using the tm and SnowballC libraries included:

1. Converting text to lowercase.
2. Removing punctuation, numeric characters and stop words.
3. Applying stemming to group similar words.

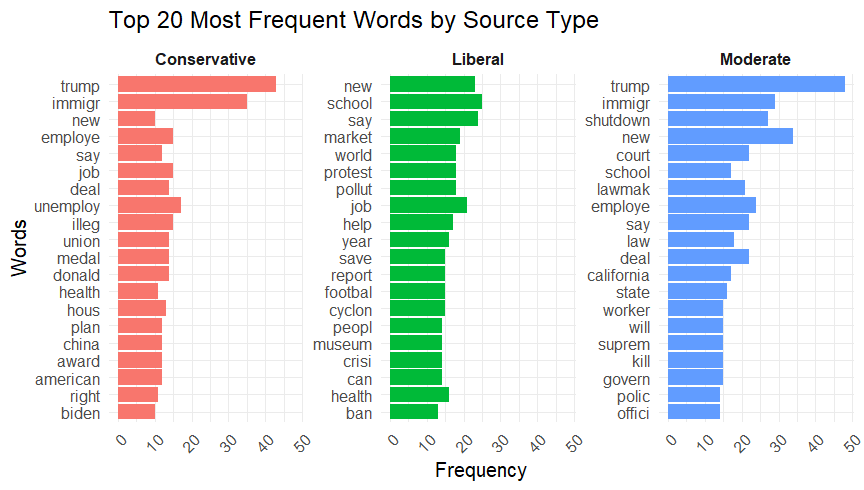
The cleaned text was transformed into a unigram Document-Term Matrix (DTM) and then converted into a sparse matrix for computational efficiency.

**Figure 1.**



The pie chart, figure 1, illustrates the distribution of articles by their political standpoint, placed into three groups: conservative, moderate and liberal. The red segment represents articles from news sources identified as liberal viewpoints. The green segment indicates articles from neutral sources, which are classified as without a strong political bias. Lastly, the blue segment represents articles associated with conservative outlets. This visualisation provides an overview of the ideological balance within the dataset, to give understanding of the number of sources within each political type before conducting statistical analysis.

**Figure 2.**



The bar chart, figure 2, highlights the top 20 most frequently used words in articles from sources categorised by their political alignment. Each panel showcases the most common terms for a specific alignment, with word frequency represented on the x-axis.

In the liberal part of figure 2, frequently used terms such as "new", "school", "market" and "protest" indicate a focus on societal, economic and educational topics. Words like "crisis", "health" and "ban" suggest an emphasis on social issues and policies aimed at addressing inequities or systemic challenges. The inclusion of words like "cyclone" and "museum" could indicate coverage of cultural and environmental events, reflecting a broader scope of issues relevant to progressive audiences.

There’s a clear cross over in the words used in conservative and moderate political alignment, words such as “Trump” and “immigration” are the most common in both source types. This reflects coverage potentially focused on politics, national identity and policies. This dataset was used in a time when Donald Trump was president of the USA, where he was known to have controversial views on immigration and politics in general. Moreover, this is highlighted in the conservative area with words such as “China” and “Biden”, conveying words that may be associated with political competition overseas and international countries.

These patterns suggest that the most common words reflect the key issues and topics that resonate with each political alignment's audience. Liberal sources emphasise social justice, education and global challenges, whereas neutral and conservative sources provide similar coverage of governance and legal matters, immigration, national identity and domestic policies. The differences in word frequency highlight the diverse priorities and interests catered to by each political alignment. However, due to moderate and conservative having very similar key words this may pose issues later when conducting the statistical analysis.

**Data Splitting**

The sparse matrix was split into training (70%) and testing (30%) datasets, with the target variable (source\_type) indicating political alignment.

**Model Selection**

Three machine learning models were implemented:

1. Logistic Regression with Lasso Regularisation.
2. Support Vector Machines (SVM).
3. Random Forest.

These methods were chosen as Logistic Regression with Lasso Regularisation model penalises less informative features, reducing overfitting and enhancing interpretability. Support Vector Machine (SVM) separates classes by finding the hyperplane that maximizes the margin between them. Finally, Random Forest method used all the features listed above and 100 bootstrapping replicates. To evaluate the model performance the accuracy, sensitivity and confusion matrix were analysed.

**Addressing Research Questions**

1. **Classification of Political Alignment**  
   The combination of Logistic Regression, SVM and Random Forest effectively classified news articles into liberal, conservative and moderate categories, demonstrating the feasibility of machine learning methods in detecting political alignment. These models, trained on unigram features extracted from article titles, provided a systematic approach to analysing media bias.
2. **Linguistic Patterns and Features**  
   Through unigram analysis, the study identified distinctive linguistic markers associated with each political alignment, such as "Trump" frequently appearing in conservative articles and "climate change" in liberal ones. This insight highlights the linguistic patterns that differentiate political categories.

This comprehensive methodology underscores the potential of machine learning to uncover linguistic patterns in political media, advancing understanding of bias in news coverage. With the final research question being addressed within the results when analysing the accuracy of each model.

**Results:**

**Figure 3.**

A screenshot of a graph

Description automatically generated

Figure 3 visually underscores the distinct classification strengths and weaknesses of each model. The visualisation illustrates the confusion matrices for Logistic Regression, Random Forest and Support Vector Machine (SVM) models in classifying news articles into Liberal, Moderate and Conservative categories. Each matrix provides insights into the frequency of correctly and incorrectly classified instances across the three classes. The Random Forest model shows a relatively balanced performance, particularly excelling in classifying Liberal articles but struggling with Conservative articles. The SVM model demonstrates strong performance for Moderate articles but also faces challenges with Conservative classification. Logistic Regression captures Liberal articles effectively, but its performance for the Conservative class remains the weakest.

The performance of three statistical models was analysed for their ability to classify news articles into Liberal, Moderate and Conservative categories based on unigram features. Among the models, Random Forest achieved the highest overall accuracy (63.93%), followed closely by SVM (63.77%) and Logistic Regression (61.29%). While all three models demonstrated reasonable performance, their effectiveness varied significantly across ideology types.

The Random Forest model showed the most balanced performance, excelling in classifying Liberal articles with a balanced accuracy of 75.64%. Its ability to handle large feature sets and complex patterns made it robust for this classification task. However, it struggled with Conservative articles, achieving a sensitivity of only 30.49%, indicating frequent misclassification. Moderate articles also presented challenges due to overlapping linguistic features with Liberal articles, which reduced classification precision.

The SVM model slightly outperformed Logistic Regression, particularly in classifying Moderate articles, achieving a sensitivity of 62.66%. However, like the Random Forest model, it faced difficulties with Conservative articles, achieving a sensitivity of 42.47%. Liberal articles were classified more reliably, showing SVM’s strength in recognising frequent and distinct terms.

Logistic Regression, despite achieving the lowest overall accuracy (61.29%), demonstrated strong sensitivity for Liberal articles (86.63%), effectively capturing the most frequent and distinguishing features of this class. However, its performance for Conservative articles was weakest, with a sensitivity of 36.99%, likely due to class imbalance and the lack of distinct vocabulary within this category.

Across all models, Conservative and Moderate articles were consistently challenging to classify, often being misclassified into other categories. This underscores the need for more distinct linguistic features and better handling of class imbalance through techniques such as oversampling or class-weighted algorithms.

Future improvements could include incorporating richer features, such as bigrams, trigrams or semantic embeddings, to better capture contextual nuances. Addressing class imbalance through oversampling or class-weighted loss functions, could further enhance classification performance. Moreover, regarding the limitation of the timeframe being in only the year 2019, alongside the nature of the news usually reporting similar events, the training data that is used in this research may not be applicable to other settings. Therefore, future research should have focus on news headlines across a longer period to see how the affects may change and have access to a stronger training dataset for the machine learning analysis.

These findings align with and extend prior research. Tun, Y. M., & Khaing, M. (2023) researched on political sentiment analysis in tweets highlights the effectiveness of SVM for multi-class political classification. Our study corroborates this by demonstrating SVM’s utility in distinguishing political ideologies in news articles. Similarly, Vysotska et al. (2024) emphasize the effectiveness of Random Forest models in detecting propaganda and fake news. The results reinforce this, showing Random Forest’s capability to capture complex patterns and achieve the highest accuracy in our analysis. Finally, the research by Tun, Y. M., & Khaing, M. (2023) on hybrid lexicon-based and machine learning approaches demonstrates the value of feature engineering for improving classification. While this study focuses on unigrams, future work could integrate these methods to align with their findings.

**5. Conclusions**

This research aimed to classify news articles based on their political alignment (Liberal, Moderate and Conservative) using machine learning models. Employing Logistic Regression with Lasso regularisation, Support Vector Machines (SVM) and Random Forest, the study utilised unigram features to analyse and differentiate language patterns in article titles. The models demonstrated varied performance, with Random Forest achieving the highest accuracy, followed closely by SVM and Logistic Regression. The findings provide valuable insights into linguistic patterns and machine learning techniques for political ideology news text classification while underscoring the challenges of classifying underrepresented and overlapping categories.

**Key Findings**

1. **Overall Model Performance** 
   * Random Forest achieved the highest accuracy (63.93%), closely followed by SVM (63.77%) and Logistic Regression (61.29%).
2. **Class-Specific Challenges** 
   * Conservative articles consistently posed classification challenges across all models, with the lowest sensitivity (30.49% in Random Forest).
3. **Random Forest** 
   * Random Forest showed the highest balanced accuracy for Liberal articles (75.64%) and consistent performance across categories.
4. **Strength of SVM in Handling High-Dimensional Data** 
   * SVM performed better for Moderate articles (sensitivity: 62.66%), showing its capacity to differentiate overlapping categories.
5. **Sensitivity in Logistic Regression** 
   * Logistic Regression achieved strong sensitivity for Liberal articles (86.63%), effectively capturing distinguishing linguistic patterns.

**Limitations, Assumptions and Weaknesses**

**Limitations**

1. **Class Imbalance**
   * Conservative articles were underrepresented, leading to misclassification and lower sensitivity.
2. **Unigram Feature Scope**
   * The reliance on unigrams limited the contextual understanding of language patterns.
3. **Generalisability**
   * The study focused on a single dataset and specific political context, which may not generalise to other datasets or regions.

**Assumptions**

1. **Source Alignment Consistency**
   * News sources were assumed to consistently align with the labelled political categories.
2. **Title Representation**
   * Article titles were assumed to sufficiently capture the linguistic cues for political alignment.
3. **Machine Learning Applicability**
   * Machine learning models were assumed to effectively classify political bias based solely on linguistic patterns.

**Weaknesses**

1. **Simplistic Features**
   * Using only unigrams limited the model's ability to capture nuanced linguistic features.
2. **Overlapping Vocabulary**
   * Similar terms across Liberal and Moderate categories reduced classification precision.
3. **Algorithm Dependency**
   * Model performance heavily depended on the choice of algorithms and parameter tuning.

**Future Work**

1. **Incorporate Richer Features**
   * Use bigrams, trigrams and semantic embeddings to capture contextual nuances and improve classification accuracy.
2. **Address Class Imbalance**
   * Apply over sampling, under sampling or class-weighted algorithms to enhance the classification of underrepresented categories.
3. **Expand Dataset Coverage**
   * Include more diverse datasets to test the generalisability of the findings across different political and linguistic contexts.

This study highlights the potential of machine learning models in political text classification and reveals the challenges inherent in classifying overlapping and underrepresented categories. By incorporating advanced features and methodologies, future research can refine these approaches and contribute to a deeper understanding of political bias in news articles.

Top of Form

Bottom of Form

**References**

Aphiwongsophon, S., & Chongstitvatana, P. (2018). Detecting fake news with machine learning methods. 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), IEEE. <https://doi.org/10.1109/ECTICon.2018.8620051>

Garcia-Marín, J., & Calatrava, A. (2018). The use of supervised learning algorithms in political communication and media studies: Locating frames in the press. Pamplona, 31(3), 175–188. <https://doi.org/10.15581/003.31.3.175-188>

Gupta, A., Barmak, O., Wierzbicki, A., & Skrypnyk, T. (2023). An adaptive approach to detecting fake news based on generalized text features. CEUR Workshop Proceedings, 3387, 300–310. <https://doi.org/10.5815/ijcnis.2024.05.06>

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220. <https://psycnet.apa.org/doi/10.1037/1089-2680.2.2.175>

Petukhova, A., & Fachada, N. (2022). MN-DS: A multilabeled news dataset for news articles hierarchical classification [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.7394851>

Pew Research Center. (2020). Media polarization methodology. Pew Research Center. <https://www.pewresearch.org/journalism/2020/01/24/media-polarization-methodology/>

Press Gazette. (2024). Trust in media: UK falls to bottom of global rankings in Edelman barometer. Press Gazette. <https://pressgazette.co.uk/media-audience-and-business-data/trust-in-media-uk-edelman-barometer-2024/>

Singh, N., Shuaib, M., Rana, M., Maurya, S., Negi, H. S., & Rawat, V. (2023). Fake news recognition: A machine learning approach for text analysis using hyperparameter tuning. 2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bengaluru, India, 909–914. <https://doi.org/10.1109/ICIMIA60377.2023.10426099>

Tun, Y. M., & Khaing, M. (2023). A large-scale sentiment analysis using political tweets. International Journal of Electrical and Computer Engineering, 13(6), 6913–6925. <https://doi.org/10.11591/ijece.v13i6.pp6913-6925>

Vysotska, V., Przystupa, K., Chyrun, L., Vladov, S., Ushenko, Y., Uhryn, D., & Hu, Z. (2024). Disinformation, fakes and propaganda identifying methods in online messages based on NLP and machine learning methods. International Journal of Computer Network and Information Security (IJCNIS, 16(5), 57–85. <https://doi.org/10.5815/ijcnis.2024.05.06>