# Exploring Text Classification Models on Political Bias in Online Media

*Isaia Caluza*

Deparment of Computer and Information Science, Fordham University, NY

***Abstract*— This paper explores the application of various machine learning models for identifying political bias in online text content. We evaluated three models: Logistic Regression with TF-IDF, Deep Neural Networks (DNN), and DistilBERT. After identifying challenges with class imbalance in our initial experiments, we applied Synthetic Minority Over-sampling Technique (SMOTE) and customized stopwords to improve performance. Our approach incorporated unsupervised kearning as an exploratory tool to inform model improvements. DistilBERT consistently outperformed the other models, achieving accuracy improvements from 42% to 67.5% after addressing class imbalance and text preprocessing concerns. Our analysis of confusion matrices shows significant improvements across all models after applying these techniques, with DistilBERT showing the most substantial gains in correctly classifying political bias. This research contributes to the growing field of automated media bias detection, with potential applications in inferring misinformation and potentially uncovering hidden semantic patterns in bias and misinformation.**

***Keywords*— political bias, text classification, machine learning, natural language processing, misinformation detection, transformer models**

## I. Introduction

The proliferation of online news and social media has dramatically transformed how people consume information, creating personalized information ecosystems that can reinforce existing beliefs and potentially enforce political polarization. In the digital era, detecting political bias in text content has become increasingly important for fostering media literacy and balanced information consumption. Simply put, media bias is a concern in part because it may prevent citizens from making fully informed choices. This project focuses on identifying political bias in online text content, particularly in news articles and social media posts. Understanding the ideological framing of content is essential, as it can subtly influence how the public perceives and interprets information. Bias is defined as an ideological leaning, such as left, center, and right, aligning with the approach taken by established bias rating systems like AllSides [4] and Ad Fontes Media [5]. This definition prompts us to set a primary goal of classifying a given text based on its political orientation using supervised machine learning techniques. As Baly et al. [6] demonstrated, automated detection of media bias is feasible with modern NLP techniques, although challenges remain in capturing the nuanced and contextual nature of political language. Supervised machine learning models such as logistic regression, deep neural networks (DNN), and fine-tuned transformer models (DistilBERT) were used for political bias classification on labeled datasets. These models were trained on labeled datasets of political texts and evaluated using accuracy, recall, F1 score, and ROC-AUC. For feature extraction and text representation, TF-IDF vectorization and SimCSE were implemented. TF-IDF was first utilized to vectorize the words of news articles and online content with numerical scores. Moreover, SimCSE improves sentence-level understanding and analyzes the similarity between content and known ideological positions. As an exploratory extension to the core classification task, unsupervised clustering techniques were applied to visualize how political texts group together in embedding space based on their ideological leanings. This exploratory data analysis served multiple purposes within our research framework: (1) pattern discovery - by observing how texts naturally cluster in the embedding space, we gained insights into potential overlaps and distinctions between different political orientations that might not be immediately apparent in supervised learning, (2) feature evaluation - clustering helped evaluate the effectiveness of feature extraction methods (TF-IDF and SimCSE) in capturing meaningful semantic patterns related to political bias, (3) informing model improvements - the clustering results revealed limitations in the initial approach, specifically in distinguishing between politically

charged language and neutral reporting language. This insight directly informed the decision to implement custom stopwords for TF-IDF based on the top words per cluster, focusing on removing common, filler, and non-informative words that carry little political signal, (4) class imbalance visualization - the distribution of data points within the clusters highlighted the class imbalance issues that were affecting the supervised models, reinforcing the need for balancing techniques like SMOTE.

Additionally, a misinformation-labeled dataset was briefly explored to see how it could be integrated into future work to investigate potential correlations between political bias and misinformation, a relationship suggested by several studies in communication research [10]. However, this remains outside the scope of the main predictive task and is proposed as a possible direction for future development.

## II. Methods

### *A. Dataset*

**Article Bias.** The primary dataset "siddharthmb/political-bias-prediction-media-splits " builds upon political bias ratings curated by AllSides. AllSides is a well-known platform that classifies media sources based on political bias using a multi-method approach that includes: editorial review, blind bias surveys of American readers, third-party data integration, independent expert analysis, and community ratings and feedback. The databaset used in the project contains 30,246 news articles, each labeled with political ideology as left (0), center (1), or right (2). The labels are based on AllSides' aggregated and verified rating system. Each data sample includes the following fields:

- bias: the numerical label (0 = left, 1 = center, 2 = right)
- bias text: human-readable bias description
- content: a cleaned and preprocessed version of the article text
- title, url, source, date, authors: additional metadata
- source_url: the original news website

The dataset is split into training, validation, and test subsets by media outlets. It means that articles from the same source do not appear in multiple subsets. This design choice is essential to prevent overfitting to the writing style or vocabulary of individual news outlets.

**Misinformation Dataset.** The second dataset was the "roupenminassian/twitter-misinformation" dataset from HuggingFace. The dataset compiles several existing datasets to train misinformation detection models. Each instance in the dataset contains: (1) text: a string feature containing the tweet or news content, and (2) label: a binary classification (0 for factual, 2 for misinformation). The data is split into 92, 394 training examples (60,309 factual and 32,805 misleading) and 10,267 testing examples (6,773 factual and 3,494 misleading).

### *B. Model Selection*

**Logistic Regression.** The project's first approach was to preprocess (tokenize) article texts with the traditional vectorization method, TF-IDF. The vectors were fed into the models as inputs. The first modeling attempt involved a simple, interpretable baseline: logistic regression algorithm predicted political bias by applying a weighted sum over features, passing the result through a sigmoid function to compute class probabilities. This model was computationally fast and helped inspect which terms/words most influenced predictions.

**Deep Neural Network (DNN).** We recognized that linear classifiers are limited to linearity and cannot tackle nonlinear and complex meanings, which is essential in understanding political language. So, Tensorflow was utilized to develop a Deep Neural Network to model nonlinear interactions among TF-IDF features/words. The model's architecture included two hidden layers with ReLU activations, dropout regularization, "Categorical Cross-Entropy" loss function, and Adam optimizer for adaptive rate tuning.

**DistilBERT.** DistilBERT was employed, which is a distilled version of BERT (Bidirectional Encoder Representations from Transformers). BERT is a deep learning model developed by Google that uses a bidirectional transformer architecture to

simultaneously capture context from both the left and right sides of a token. It generates deep contextualized word embedding, which enables it to achieve state-of-the-art results on a wide range of natural language understanding tasks. DistilBERT is this model except on a smaller scale: it retains approximately 97% of its performance while reducing model size by 40% and increasing inference speed by 60%. DistilBERT was fine-tuned on the labeled dataset for a three-way text classification task (left, center, right). Before model training, the textual data was preprocessed through tokenization, segmenting each input into subword units compatible with DistilBERT's tokenizer. The tokens were subsequently mapped to their corresponding numerical indices in DistilBERT's vocabulary through encoding. The encoded sequences were then passed through DistilBERT's transformer layers, which model complex contextual dependencies within the text. The final hidden states were fed into a classification head to output probabilistic predictions over the three target classes. Model performance was evaluated using standard metrics including accuracy and AUC-ROC.

### *C. Data Balancing*

After initial experiments revealed significant class imbalance issues, we implemented the Synthetic Minority Over-sampling Technique (SMOTE) to address this problem. SMOTE works by creating synthetic samples from the minority class instead of simply duplicating existing samples. The technique generates new instances by interpolating between existing minority class examples that are close together in the feature space. By applying SMOTE, we achieved a more balanced dataset with equal representation across the three political bias categories (left, center, right).

### *D. Customized Stopwords for TF-IDF*

Our analysis [] of initial TF-IDF results showed that many high-frequency terms were not actually contributing meaningful information for political bias classification. To address this, a customized stopword list was developed that goes beyond standard English stopwords. Common news article terms (e.g., "said", "reported", "according") and source identifiers that weren't meaningful for detecting political bias were removed. This custom stopword filtering improved the quality of the TF-IDF features by ensuring they captured more politically relevant terminology.

### *E. Unsupervised Exploratory Tools*

Unsupervised learning techniques were employed not only as supplementary classification methods but primarily as exploratory data analysis tools to inform improvements to the predictive models. Clustering algorithms like HDBSCAN provide valuable insights into the natural organization of data that can reveal patterns and relationships not immediately apparent in supervised learning approaches. In addition to HDBSCAN, SimCSE was implemented to fine-tune vectorization of text.

## III. Results

### *A. Logistic Regression (initial vs. SMOTE and stopwords)*

The initial logistic regression model achieved a low accuracy of 22% on validation data, suggesting limited predictive strength. The model showed notable behavior in terms of precision and recall. It showcased high recall (0.62) but low precision (0.04) for right-leaning articles, meaning it aggressively labeled articles as "right" with many false positives. For left-leaning articles, the model demonstrated higher precision (0.53) but lower recall (0.19). Testing results differed as accuracy rose to 53%, with precision (0.61) and recall (0.59) for right-leaning articles improving substantially. The model benefited from the imbalance since right-leaning articles were five times more common in the testing set than in validation. In contrast, left-leaning articles were eighty times less represented. The validation data's ROC-AUC score (0.4645) suggested the model could not capture political bias. The model performed best at recognizing center-leaning articles (AUC = 0.65) but struggled with right (AUC = 0.45) and left (AUC = 0.29) articles. See Appendix A.

The confusion matrix after applying SMOTE and custom stopwords showed significant improvements. See Appendix B:

- True Left predicted as Left: 1371 (67.1% of actual left instances)

- True Center predicted as Center: 1497 (72.1% of actual center instances)
- True Right predicted as Right: 1382 (68.4% of actual right instances)

Overall accuracy improved from 22% to 50.8%, demonstrating addressing class imbalance customizing stopwords improved the model.

*B. DNN (initial vs. with SMOTE and stopwords)*

After training, the DNN showed modest improvement in validation accuracy, reaching approximately 27% and a slight improvement in ROC-AUC with 0.4751 compared to Logistic Regression's 0.4645. After applying SMOTE and custom stopwords, the model showed improvements. See Appendix C and D:

- True Left predicted as Left: 1362 (66.7% of actual left instances)
- True Center predicted as Center: 1534 (73.9% of actual center instances)
- True Right predicted as Right: 1468 (72.7% of actual right instances)

The overall accuracy increased from 27% to 52.4%, representing a 94% improvement. The DNN now showed clearer differentiation between classes and better handling of center-leaning content compared to the original model.

*C. DistilBERT (initial vs. with SMOTE and stopwords)*

DistilBERT showed significant improvement over the previous models. The classification report revealed an improvement in validation accuracy of approximately 42%. After applying SMOTE and stopwords, DistilBERT showed the most improvement:

- True Left predicted as Left: 2111 (86.6% of actual left instances)
- True Center predicted as Center: 1643 (82.1% of actual center instances)
- True Right predicted as Right: 2087 (72.7% of actual right instances)

The overall accuracy jumped from 42% to 67.5%, representing a 60.7% improvement. DistilBERT outperformed both Logistic Regression and DNN models.

TABLE I

| **Model** | **Original Accuracy** | **Improved Accuracy** |
|---|---|---|
| LR | 22.0% | 50.8% |
| DNN | 27.0% | 52.4% |
| DistilBERT | 42.0% | 67.5% |

*D. Clustering*

Hierarchical DBSCAN clustering on SimCSE embeddings revealed two significant clusters for the political bias dataset: Cluster -1 and Cluster 1. See Appendix I and J. Cluster -1 contained 221, 193, and 235 texts from left, center, and right biases respectively, while Cluster 1 contained 112, 103, and 122 texts. Both clusters showed a somewhat even distribution across bias categories, indicating no significant pattern in the clusters. For the misinformation dataset, three clusters were identified: Cluster -1 (179 factual, 64 misinformation), Cluster 0 (499 factual, 248 misinformation), and Cluster 1 (0 factual, 10 misinformation). Cluster 1's composition of only misinformation suggested a localized embedding region with a pattern of misinformation but has very few data points to be considered significant.

Wordclouds generated (see Appendix M and N) using TF-IDF revealed almost all clusters in political bias dataset contained the word "said" as a top word, which is insignificant in terms of contributing to political bias. This may explain why predictive models were not accurate because it took into account non-significant filler words.

title and author details must be in single-column format and must be centered.

## IV. Discussion

There were several challenges throughout the project. The initial dataset showed significant class imbalance, with right-leaning articles dominating the training and testing sets. This skewed the model's prediction toward the majority class and made it difficult to recognize underrepresented labels. Table II illustrates this imbalance:

TABLE II

| Set | Left | Center | Right |
|---|---|---|---|
| Training | 33% | 28% | 39% |
| Validation | 70% | 25% | 5% |
| Test | 30% | 23% | 46% |

**TF-IDF Limitations.** Initial TF-IDF implementation could not capture word order, contextual or semantic meaning, hindering the models' ability to learn nuanced political rhetoric. Wordclouds generated from TF-IDF vectors showed many similar top words across the clusters were filler words rather than politically meaningful terms.

**Impact of SMOTE and Custom Stopwords.** The application of SMOTE improved model performance by creating balanced class distributions. This allowed models to learn patterns across political leanings equally. The custom stopword list further refined the feature space by filtering out terms that were common across all political orientations but carried little discriminative power. The improvements were most pronounced in Logistic Regression, which saw a 130.9% increase in accuracy. This suggests that the original poor performance was largely due to data imbalance rather than model limitations. While DistilBERT showed the smallest relative improvement (60.7%), it achieved the highest absolute performance (67.5%), confirming that transformer models are best suited for this task when provided with balanced, well-preprocessed data.

**Clustering Insights.** The exploratory clustering analysis revealed important insights:

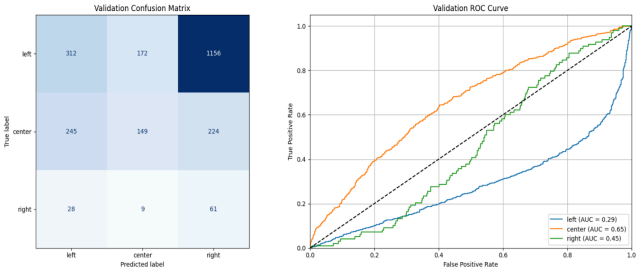
1. *Semantic Overlap*: the significant overlap between political categiries in embedding space indicated that simple lexical features were insufficient for distinguishing political bias. This observation pushed us toward more sophisticated semantic representations.
2. *Non-informative features*: word frequency analysis within clusters revealed that many high-frequency terms were journalistic conventions rather than indicators of political bias. This insight led to the development of the custom stopword list, which significantly improved classification performance.

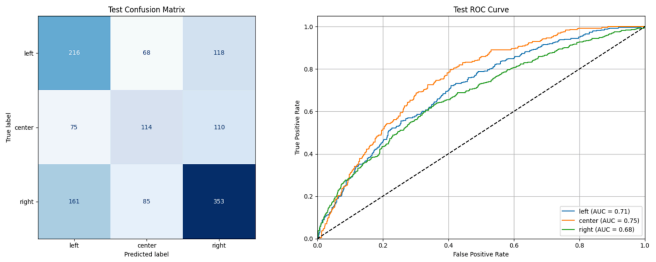
## V. Conclusion

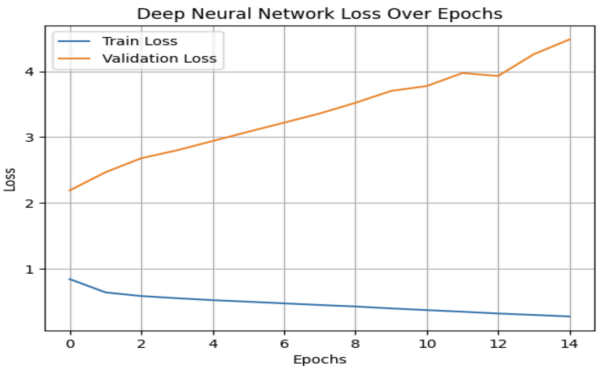
This project demonstrates the value of using unsupervised learning as an exploratory tool to inform supervised model improvements. The clustering analysis provided critical insights into feature quality and data distribution challenges that might have gone unnoticed in a purely supervised approach. In return, there were dramatic improvements in model performances. While unsupervised clustering provided limited direct classification value due to semantic overlap between political categories, it highlighted the potential connections between specific text patterns and misinformation that warrant further investigation. This finding suggests promising directions for future research at the intersection of political bias and misinformation. The results demonstrate that effective political bias detection requires both sophisticated models and careful data augmentation. Transformer-based models like DistilBERT, when combined with balanced datasets and domain-appropriate text preprocessing, offer the most promising approach for identifying political bias in online content. Future iterations of this project should therefore prioritize misinformation integration, the construction of balanced dataset, and advancing the model to include languages other than English and also politics outside of the US.

## VI. Appendix

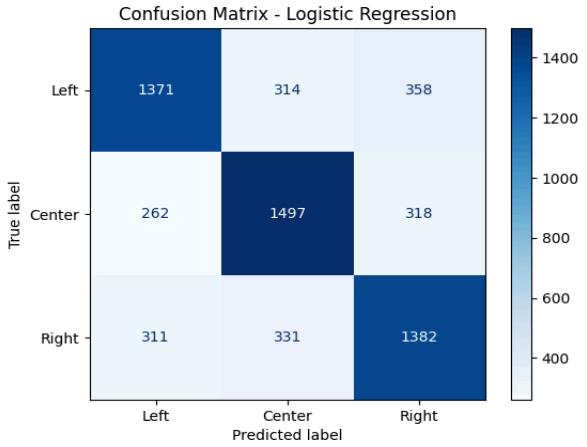
### *A. Logistic Regression (initial confusion matrix & ROC curves)*



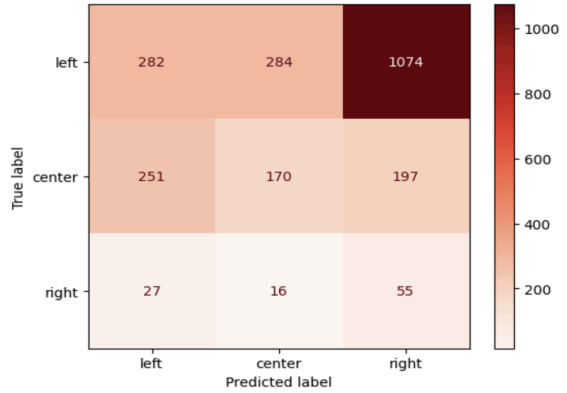


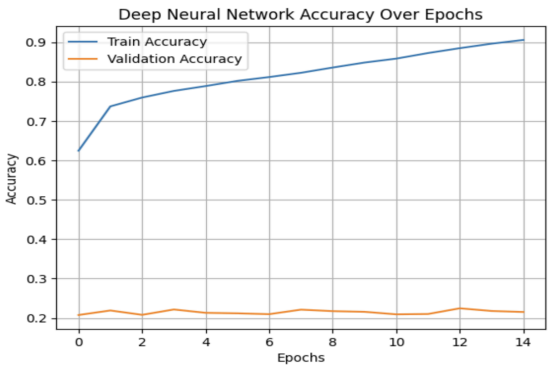




*B. Logistic Regression (improved confusion matrix)*

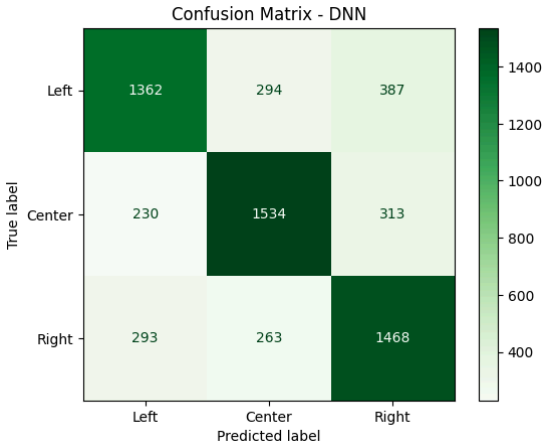



*C. DNN (initial confusion matrix and ROC curves)*

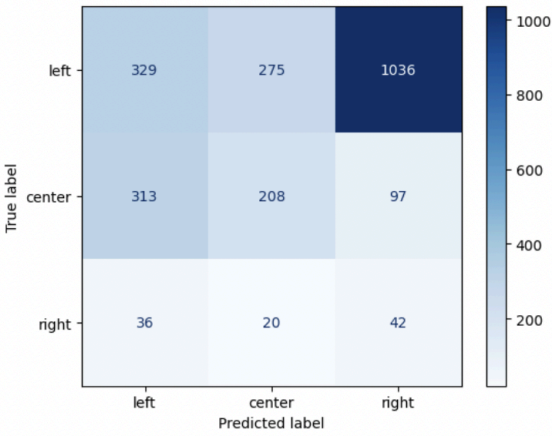


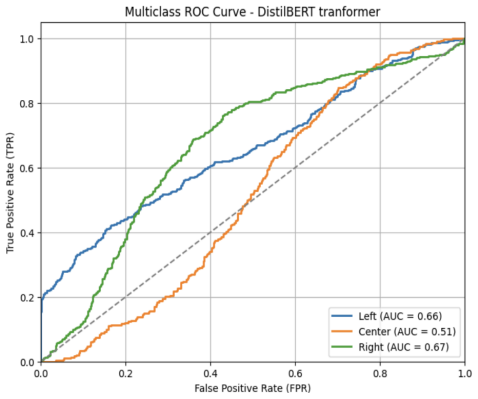




*D. DNN (improved confusion matrix)*

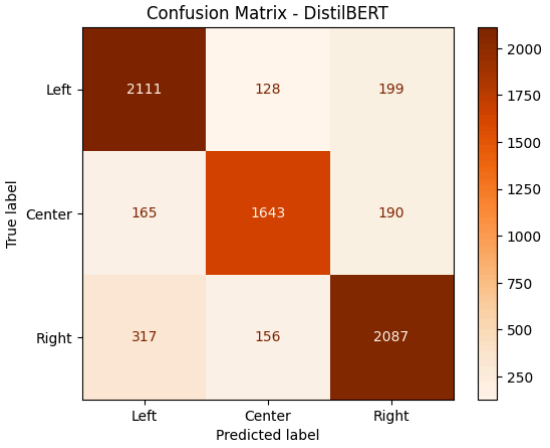



*E. DistilBERT (initial confusion matrix & ROC curves)*

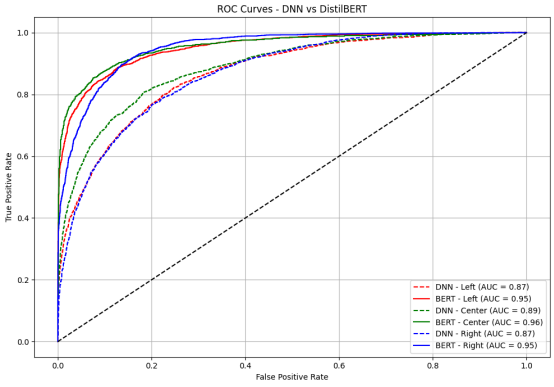




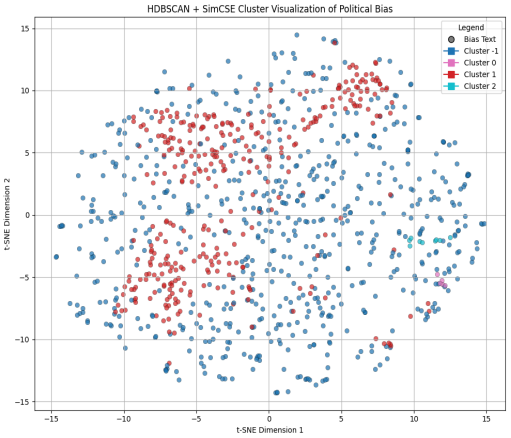

*G. DistilBERT (improved confusion matrix)*

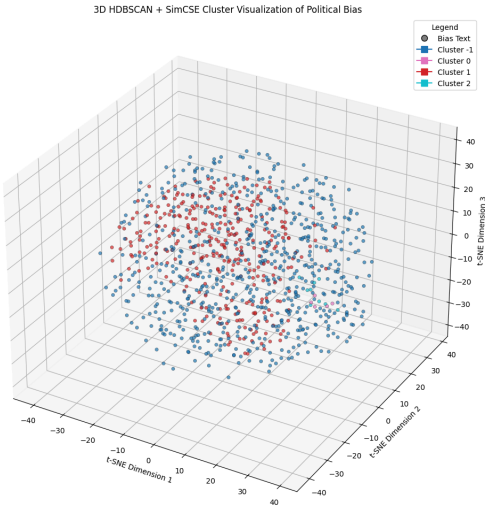



*H. ROC Curves - DNN vs. DistilBERT*




*I. 2D + 3D clustering visualization of political bias data*

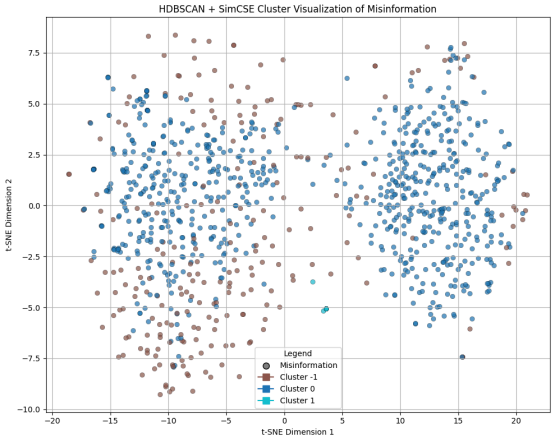


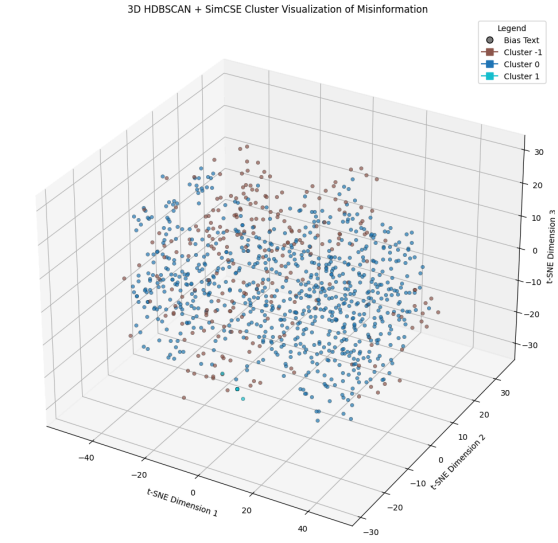




*J. Hierarchical DBSCAN clustering summary on SimCSE embeddings of text from the political bias dataset.*

| Cluster | True Bias | | |
|---|---|---|---|
| | Left | Center | Right |
| -1 | 221 | 193 | 235 |
| 0 | 1 | 0 | 4 |
| 1 | 112 | 103 | 122 |
| 2 | 2 | 2 | 5 |

*K. 2D & 3D HDBSCAN + SimCSE clustering visualization of misinformation data.*



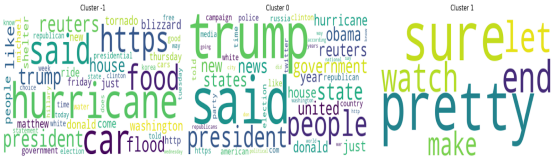




*L. Hierarchical DBSCAN clustering summary on SimCSE embeddings of text from the misinformation dataset.*

| Cluster | True Label | |
|---|---|---|
| | Factual (0) | Misinformation (1) |
| -1 | 179 | 64 |
| 0 | 499 | 248 |
| 1 | 0 | 10 |

*M. Wordclouds by cluster show the top 50 words in each cluster with TF-IDF for political bias data.*


*N. Wordclouds by cluster show the top 50 words in each cluster with TF-IDF for misinformation data.*




## References


[1] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019. [Online]. Available: https://huggingface.co/distilbert-base-uncased

[2] S. Menon, "Political Bias Prediction Dataset - Media Splits," 2023. [Online]. Available: https://huggingface.co/datasets/siddharthmb/article-bias-prediction-media-splits

[3] R. Minassian, "Twitter Misinformation Dataset," 2023. [Online]. Available: https://huggingface.co/datasets/roupenminassian/twitter-misinformation

[4] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021. [Online]. Available: https://github.com/princeton-nlp/SimCSE

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[6] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in Pacific-Asia conference on knowledge discovery and data mining, pp. 160-172, 2013.

[7] F. Hamborg, K. Donnay, and B. Gipp, "Automated identification of media bias in news articles: an interdisciplinary literature review," International Journal on Digital Libraries, vol. 20, pp. 391-415, 2019.

[8] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021. [Online]. Available: https://github.com/princeton-nlp/SimCSE

[9] C. Baden and N. Springer, "Conceptualizing viewpoint diversity in news discourse," Journalism, vol. 18, no. 2, pp. 176-194, 2017.

[10] Y. Guess, B. Nyhan, and J. Reifler, "Exposure to untrustworthy websites in the 2016 US election," Nature Human Behaviour, vol. 4, pp. 472-480, 2020.

[11] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, "Fake news on Twitter during the 2016 U.S. presidential election," Science, vol. 363, no. 6425, pp. 374-378, 2019.

[12] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019. [Online]. Available: https://huggingface.co/distilbert-base-uncased

[13] S. Menon, "Political Bias Prediction Dataset - Media Splits," 2023. [Online]. Available: https://huggingface.co/datasets/siddharthmb/article-bias-prediction-media-splits

[14] R. Minassian, "Twitter Misinformation Dataset," 2023. [Online]. Available: https://huggingface.co/datasets/roupenminassian/twitter-misinformation