



Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

RICo: Reddit ideological communities

Kamalakkannan Ravi*, Adan Ernesto Vela

University of Central Florida, Orlando, USA

ARTICLE INFO

Keywords:

Social networking (online)
 Learning (artificial intelligence)
 Predictive models
 Transformers
 Support vector machines
 Text analysis
 Context modeling
 Natural language processing

ABSTRACT

The main objective of our research is to gain a comprehensive understanding of the relationship between language usage within different communities and delineating the ideological narratives. We focus specifically on utilizing Natural Language Processing techniques to identify underlying narratives in the coded or suggestive language employed by non-normative communities associated with targeted violence. Earlier studies addressed the detection of ideological affiliation through surveys, user studies, and a limited number based on the content of text articles, which still require label curation. Previous work addressed label curation by using ideological subreddits (*r/Liberal* and *r/Conservative* for Liberal and Conservative classes) to label the articles shared on those subreddits according to their prescribed ideologies, albeit with a limited dataset.

Building upon previous work, we use subreddit ideologies to categorize shared articles. In addition to the conservative and liberal classes, we introduce a new category called “Restricted” which encompasses text articles shared in subreddits that are restricted, privatized, or banned, such as *r/TheDonald*. The “Restricted” class encompasses posts tied to violence, regardless of conservative or liberal affiliations. Additionally, we augment our dataset with text articles from self-identified subreddits like *r/progressive* and *r/askaconservative* for the liberal and conservative classes, respectively. This results in an expanded dataset of 377,144 text articles, consisting of 72,488 liberal, 79,573 conservative, and 225,083 restricted class articles. Our goal is to analyze language variances in different ideological communities, investigate keyword relevance in labeling article orientations, especially in unseen cases (922,522 text articles), and delve into radicalized communities, conducting thorough analysis and interpretation of the results.

1. Introduction

The exploration of language within online communities is pivotal for understanding the narratives and ideologies that shape digital discourse, particularly in the context of non-normative communities linked to targeted violence. This study focuses on the intricate relationship between language use and ideological camps, examining how different linguistic patterns and terminologies are employed within various online groups to convey specific ideologies and beliefs. The importance of this research lies in unraveling the nuances of language that contribute to the formation and reinforcement of ideological echo chambers, thereby influencing the dissemination of narratives that may have real-world implications.

While Reddit, with its diverse range of subreddits and substantial user engagement, serves as our primary data source, the core of our investigation extends beyond this platform. We aim to dissect the language used in various ideological communities, regardless of their digital medium, to gain a deeper understanding of how language shapes and is shaped by ideological beliefs. Reddit, renowned for its extensive user base and varied topics [1], offers a rich dataset for this analysis.

Its structure of subreddits, each a microcosm of focused discussion and sharing, provides a detailed cross-section of language usage across different ideological spectrums.

On Reddit, the term *subreddit* refers to a specific community or forum dedicated to a particular topic or theme. Subreddits are a core aspect of Reddit’s structure, allowing users to find communities of interest where they can read, post, and engage in discussions relevant to that topic. The notation for subreddits typically includes the prefix *r/* followed by the name of the subreddit. For example, *r/Liberal* refers to a subreddit focused on liberal ideologies, discussions, and content, while *r/Conservative* would be a subreddit centered around conservative viewpoints and discussions. This notation is a shorthand way for Reddit users to reference and navigate to different communities within the Reddit platform.

Furthermore, our research taps into the emerging role of Reddit as a significant forum for news consumption, especially among young adults [2]. This aspect of Reddit as a hub for news and information exchange makes it a fertile ground for analyzing how news narratives are formed, discussed, and propagated within and beyond the platform.

* Corresponding author.

E-mail addresses: kamalakkannan.ravi@ucf.edu (K. Ravi), adan.vela@ucf.edu (A.E. Vela).

<https://doi.org/10.1016/j.osnem.2024.100279>

Received 13 September 2023; Received in revised form 22 April 2024; Accepted 5 June 2024

Available online 21 June 2024

2468-6964/© 2024 Elsevier B.V. All rights reserved, including those for text and data mining, AI training, and similar technologies.

In delving into these aspects, our study employs Natural Language Processing (NLP) techniques to systematically analyze text articles from various subreddits. By scrutinizing the linguistic nuances and keywords used in these online discussions, our goal is to illuminate the subtle yet powerful ways in which language contributes to the formation of ideological echo chambers. This analysis aims to provide a nuanced understanding of how narratives are constructed and propagated in online spaces, thereby offering insights into the dynamics of ideological affiliations and their potential impact on real-world attitudes and behaviors, particularly in the realm of targeted offline violence.

2. Related works

Our primary focus is on classifying political extremism within the context of political ideologies, such as liberalism and conservatism using the ideological narratives present in text articles shared in online communities. While previous studies have explored various aspects of online ideological orientation and extremism detection, they often do not address the specific challenge of differentiating politically motivated extremism from deeply held ideological beliefs, such as being liberal or conservative. This distinction is at the core of our research. Our work, titled “Reddit Ideological Communities” (RiCo), precisely embodies our research objectives. RiCo is tailored to our analysis of ideological communities within the Reddit platform, effectively conveying our primary focus on Reddit’s communities and their ideological narratives.

2.1. Political ideology analysis

Prior studies have made substantial strides in understanding ideological orientation detection, unveiling the complex interplay between media content and audience ideology. Golbeck and Derek introduced a pioneering technique that estimates users’ political preferences by analyzing follower connections within the Twitter network [3]. Their focus on U.S. media outlets uncovered a notable alignment between audience political leanings and outlet ideologies, emphasizing the impact of media content on audience perceptions.

Moreover, Morgan et al. [4] investigated the influence of perceived ideology on news consumption and sharing behavior on Twitter. Grounded in selective exposure theory, their study highlighted how individuals gravitate towards information aligning with their beliefs. This research underscores the profound impact of perceived ideology on digital news consumption patterns.

Furthermore, studies by Mason and Wronski [5] and Le et al. [6] explored ideological asymmetries on platforms like YouTube and Twitter and developed methods for measuring the political slant of news articles. These efforts shed light on social sorting, identity politics, and media bias within digital spaces, contributing to a deeper understanding of ideological polarization.

Akoglu introduced an algorithm, signed polarity propagation, to classify individuals into liberal and conservative camps based on their political polarity [7]. This work enhances our understanding of polarization dynamics in digital environments. Additionally, Alashri et al. [8] examined the relationship between candidates’ social media posts and commentator sentiment during the 2016 U.S. presidential election, revealing correlations between digital and offline political events.

Stier dissected semantic disparities in tweets from Democratic and Republican party actors, highlighting the role of framing techniques in shaping online political dialogue [9]. Meanwhile, Medina et al. [10] explored partisan political discourse on TikTok, offering insights into user demographics and interaction structures on emerging social media platforms.

Moreover, Ribeiro et al. [11] leveraged social media data to estimate the biases of numerous news sources, providing insights into the ideological leanings of these outlets. Their work underscores the significance of understanding biases in the digital information landscape and the impact they can have on shaping public opinion.

Finally, Ravi et al. [12] delved into natural language processing techniques for classifying conservative and liberal content within online communities, revealing the evolving nature of political discourse in the digital age. These studies collectively enrich our understanding of political ideology analysis in digital spaces.

2.2. Extremism and radicalization

In digital content analysis, the identification of extremism stands out as a pressing concern, given its pervasive presence across various online platforms. Detecting and comprehending extremist ideologies require effective methods, considering the vast amounts of content available.

To address this challenge, Yang and Chen focused on identifying radical opinions within hate group web forums, proposing a partially supervised learning approach to tackle the challenge of labeling vast amounts of content [13]. Meanwhile, Agarwal et al. [14] delved into the values characterizing stakeholders within movements such as the Tea Party and Occupy Wall Street, employing the lens of value-sensitive design (VSD) to unveil nuanced distinctions among participants. Their research highlighted shared values within these online movements, shedding light on the complex interplay of ideologies.

Enhancing our understanding of the digital landscape, Bhattacharjee et al. [15] developed a dynamic learning framework to target malicious social media posts expressing extremism or criminal intent. Moreover, on the popular social media platform, Twitter, Alizadeh et al. [16] analyzed political extremists associated with groups like the alt-right and Antifa, comparing their psychological and moral constructs with liberal and conservative users.

Delving into the Reddit community, Grover and Mark explored behavioral patterns exhibited by the alt-right community on Reddit, aiming to identify early warning signs of ideological radicalization [17]. Concurrently, Ribeiro et al. [18] undertook a comprehensive investigation into radicalization pathways on YouTube, categorizing videos and delving into mechanisms such as user growth and migration between communities.

Melton et al. [19] addressed the challenge of detecting hate speech on various platforms, introducing a novel deep learning framework coupled with transfer learning and weak supervision. Simons and Skillicorn focused on distinguishing between extremist rhetoric and potential violence, presenting a predictive model for intent detection [20].

In a comparative analysis, Sipka et al. [21] explored QAnon-related content across different social media platforms, providing insights into variations in content volume and contextual aspects surrounding discussions. Similarly, Ebner et al. [22] evaluated the national security threat posed by the QAnon movement, identifying linguistic markers associated with violence risk.

Furthermore, Gaikwad et al. [23] and Ajala et al. [24] tackled the challenge of detecting and classifying extremism on social media platforms, emphasizing the critical importance of monitoring and countering extremism across diverse platforms. Ravi et al. [25] conducted a preliminary study on measuring threats in social media comments, proposing a nuanced threat-level scale and employing AI-driven methodologies to address this critical challenge.

Fahim and Gokhale distinguished between extremist and normal content, focusing on linguistic and metadata features [26]. Ai et al. [27] conducted a comprehensive analysis of videos related to various political groups, contributing to a deeper understanding of online political dynamics, and introduced an approach to identifying hate content on social media platforms, emphasizing the role of virtual communities in hate content propagation [28].

2.3. Discourse and sentiment analysis

Focusing on the dynamics of discourse in online spaces, Kovacs et al. [29] delved into the role of platforms like Twitter in propagating unhealthy discourse. By classifying tweets as healthy or unhealthy based on attributes like hostility and sarcasm, they illuminated how such discourse contributes to political polarization. Recognizing the association between rising cybercrimes and social media platforms, Matias et al. [30] addressed the issue of predicting cyberbullying and cyberthreats on Twitter. Their work underscored the importance of proactive monitoring to prevent online crimes.

Moreover, Nguyen and Gokhale introduced an efficient approach to identifying anti-government sentiment on Twitter, focusing on politically motivated protests [31]. Their classification framework aimed to detect and mitigate sentiment, preventing its escalation into social unrest. Recently, Lee and Pirim conducted a comparative study of Twitter and Parler during the Capitol Riots of January 6, 2021 [32]. Analyzing trending content and user behavior, they offered insights into the distinctions between the two platforms during a pivotal moment in recent history.

Collectively, these prior research works on ideological orientation detection and extremism detection have significantly contributed to our understanding of extremism, hate speech, and unhealthy discourse on digital platforms. They have provided essential insights and tools for identifying, classifying, and mitigating these online threats, emphasizing the importance of ongoing vigilance and research in the digital realm. yet several significant challenges and gaps remain unexplored. The following research gaps highlight areas where prior studies have not fully delved into:

Research Gap 1: Distinguishing Ideological Beliefs from Extremism. While previous works have made substantial contributions to our comprehension of ideological orientation and polarization, they often fall short of explicitly addressing the challenge of distinguishing politically motivated extremism from ideological beliefs. Prior studies tend to treat extremism as a binary classification, missing the link and boundaries between extreme beliefs and more moderate ideological orientations.

Research Gap 2: Lack of Comprehensive Ideological Datasets. To effectively differentiate ideological beliefs from extremism, comprehensive datasets are essential. Previous research predominantly focused on binary classifications, such as identifying content as extremism, hate speech, or healthy discourse, without extensive datasets that encompass the spectrum of ideological orientations. Consequently, there is a notable absence of datasets that cover both liberal and conservative beliefs along with the extremism stemming from these beliefs. Addressing this gap requires the creation of an ideological dataset that encompasses a wide range of political perspectives, thereby enabling a more nuanced analysis of online discourse.

Research Gap 3: Developing Reliable and Interpretable Machine Learning Systems. To bridge the aforementioned research gaps, the development of machine learning (ML) models is imperative. These models should not only distinguish between ideological beliefs and extremism but also provide reliability and interpretability in their results. Thus it requires a system design and framework development to facilitate the comparison and deeper understanding of language alignment within contrasting online communities.

In addressing these research gaps, we develop and apply various language modeling techniques, including TF-IDF-based SVM, light gradient-boosting machine, fastText, long short-term memory networks, and context-based Transformers such as RoBERTa, Longformer, and GPT-2. The insights gained from this research can inform strategies for mitigating the spread of targeted harmful content and promoting more inclusive and constructive online discourse.

3. Problem statement

Our primary objective is to develop a sophisticated system for labeling the ideological orientation or affiliation of text articles, with a particular emphasis on those associated with radicalized and extremist ideologies. This effort aims to bridge three key research gaps identified in our introductory analysis.

Firstly, in addressing **Research Gap 1**, we focus on the nuanced differentiation between politically motivated extremism and ideological beliefs. Previous studies have predominantly treated extremism as a binary concept, overlooking the subtle distinctions and boundaries between extreme and moderate ideological stances. Our approach seeks to refine this classification, moving beyond binary categorizations to capture the complexity inherent in political beliefs.

Secondly, to tackle **Research Gap 2**, our strategy involves the creation and utilization of a comprehensive ideological dataset. Prior research has been limited by its focus on binary classifications, such as extremism versus healthy discourse, without a broad dataset that captures the entire spectrum of ideological orientations. Our work aims to fill this void by compiling a dataset that encompasses a diverse range of political perspectives, including both liberal and conservative beliefs, as well as the extremism emanating from these ideologies. This dataset, comprising 72,488 liberal, 79,573 conservative, and 225,083 restricted class articles spanning 13 years (2008–2021), will facilitate a more nuanced analysis of online discourse.

Lastly, **Research Gap 3** involves developing reliable and interpretable machine learning systems. The necessity lies in constructing models that effectively distinguish between ideological beliefs and extremism while ensuring the reliability and interpretability of results. Our research intends to advance this area by employing hypothesis testing, robustness analysis, interpretability assessment, and scalability testing on a dataset of 922,522 text articles. This comprehensive approach will evaluate and demonstrate the effectiveness of our models in accurately classifying ideological orientation and affiliation based on textual content, as previously initiated by Ravi et al. [12]. Unlike earlier methodologies that relied on external factors like publishing sources, our approach is grounded in the classification of ideological leaning directly from the textual content, thereby addressing the limitations of previous studies and significantly enhancing our understanding of ideological dynamics in online communities.

4. Dataset

To enhance our understanding of language usage within subreddit communities, we build upon previous research [12] by introducing a new class called *Restricted*. This new class supplements the existing *Liberal* and *Conservative* classes. The *Restricted* class encompasses text articles shared within restricted, quarantined, or banned subreddits, such as *r/TheDonald*, with a particular focus on content related to violence, irrespective of conservative or liberal labels. Meanwhile, the other classes consist of articles from subreddits that align with similar beliefs and interests. Additionally, we have collected a separate holdout dataset from various subreddits, which includes a mixture of overt, vague, and undefined ideologies. Fig. 1 displays all the collected subreddits.

To collect the text articles, we gather the Uniform Resource Locators (URLs) of all the submissions made to the subreddits. We utilize Pushshift Reddit API, starting from the first post date of each subreddit until August 2021. We remove URLs that originate from non-text sites like YouTube and Imgur as we focus on text articles. The remaining URLs are then scraped using the Beautiful Soup API. We exclude empty and duplicate articles from the corpus during this scraping process.

Additionally, to ensure the inclusion of only genuine text articles while excluding non-relevant webpage content such as video descriptions and copyright templates, we implemented a word-count threshold. The threshold was determined by conducting annotations on a

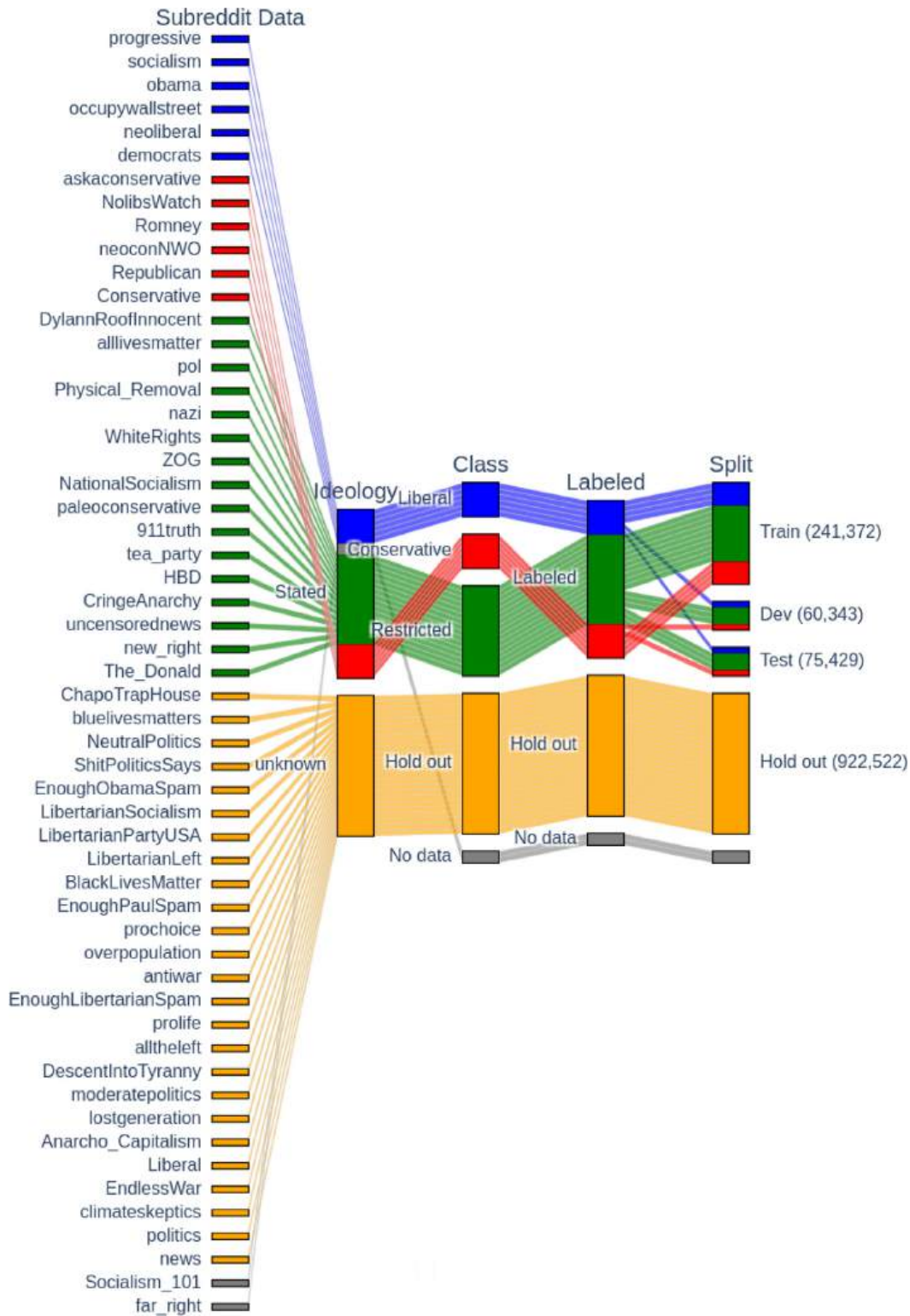


Fig. 1. Data collection and labeling.

subset of 600 articles from the *CringeAnarchy* subreddit. All 5,056 *CringeAnarchy* articles were divided into 12 groups based on word count, with 50 articles in each group. The annotations were carried out using Doccano, which is an open-source web-based annotation tool. Our aim was to determine the word count threshold at which 90% of the articles could be categorized as long text, as depicted in Fig. 2. Following the methodology outlined in Ravi et al. (2022) [12], we determined that a word limit of 300 was suitable for classifying articles

as long text. However, none of the articles met the selection criteria in a couple of subreddits, leading to the exclusion of the *Socialism_101* and *far_right* subreddits from our data collection.

The remaining corpus was further categorized as shown in Fig. 1, based on whether the respective subreddits had a clearly stated ideology on the Reddit site. The articles originating from subreddits with stated ideologies were divided into three classes: 72,488 articles in the liberal class, 79,573 articles in the conservative class, and 225,083

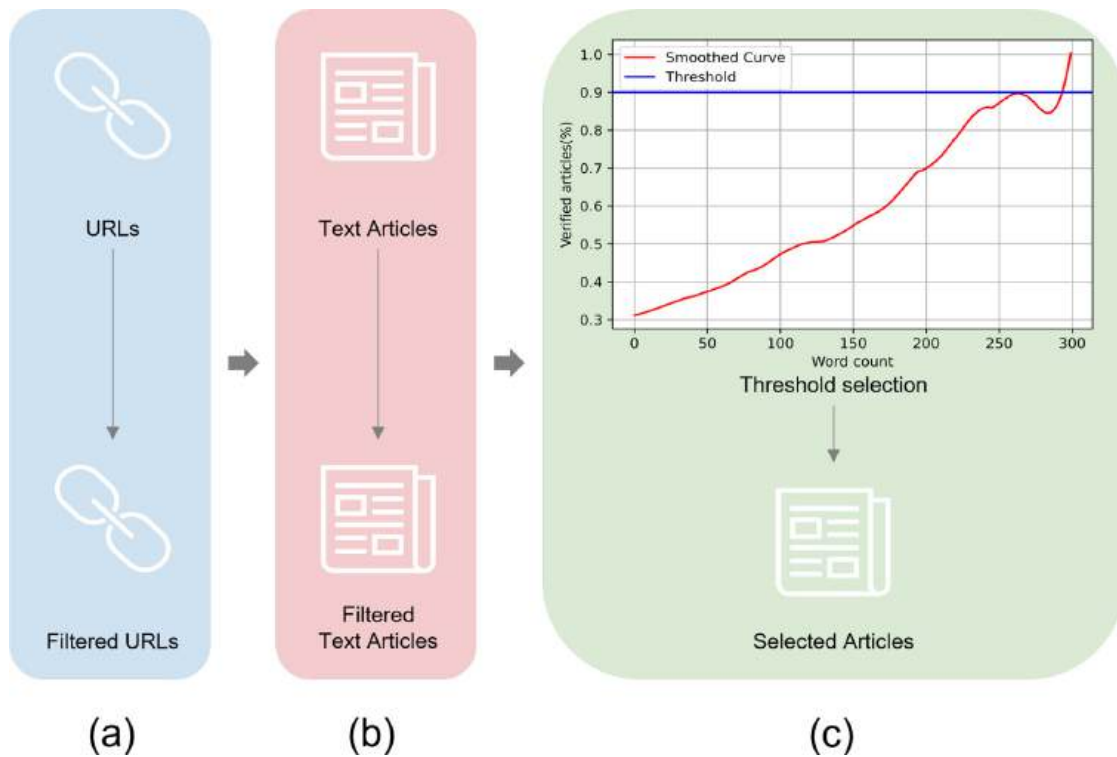


Fig. 2. Filtering Steps for Corpus Construction. (a) Exclude non-text URLs (e.g., YouTube, Imgur). (b) Remove empty, and duplicate articles. (c) Annotate relevant articles at each word count, and exclude articles with less than 300 words.

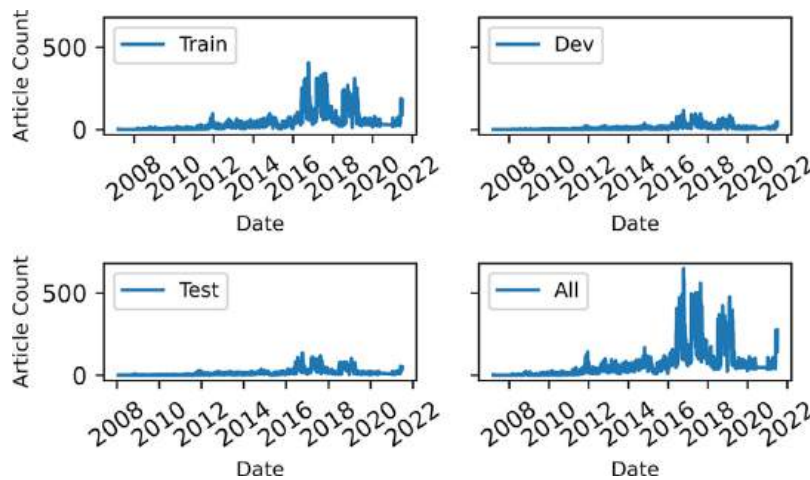


Fig. 3. Articles count over time for train, development, and test data.

articles in the restricted class. On the other hand, the articles from subreddits that did not have a stated ideology, which included some subreddits with either implied or explicitly stated ideologies, were combined to form a holdout set comprising 922,522 articles. This holdout set will serve as a case study in our research.

We can observe from Fig. 1, that the three ideological classes exhibit a significant class imbalance, which is also reflected in the distribution across the train, development, and test splits. In particular, the train set comprises 241,372 articles, the development set comprises 60,343 articles, and the test set consists of 75,429 articles. Fig. 3 visually represents the stratified splitting of the data.

5. Initial analysis

Our data analysis, as detailed in Fig. 4 provides an overview of the collected data categorized by class. Notably, we observe a gradual

Table 1
Top 10 text article website domains.

Text article sources		
Restricted	Conservative	Liberal
breitbart	breitbart	politico
dailycaller	dailycaller	theguardian
foxnews	foxnews	salon
dailymail	washingtontimes	huffingtonpost
vdare	hotair	theatlantic
theguardian	americanthinker	nytimes
amerika	dailywire	thedailybeast
washingtontimes	townhall	motherjones
townhall	liveaction	dailykos
americanthinker	politico	businessinsider

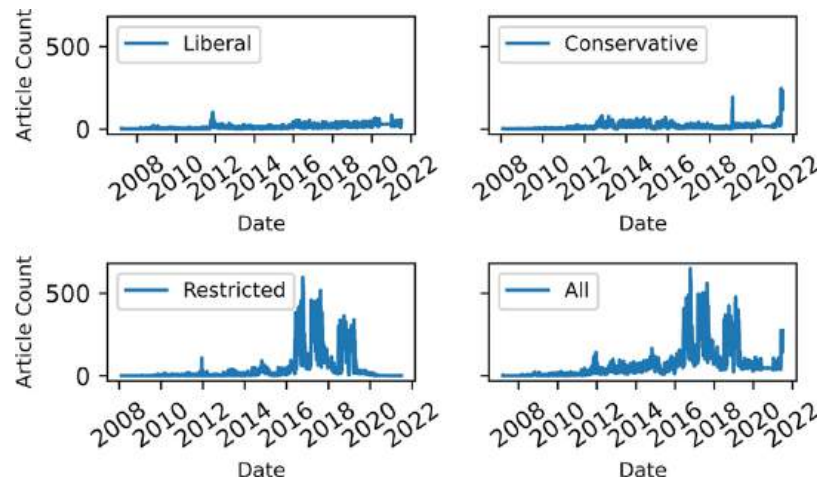


Fig. 4. Articles count over time for each class.

Table 2
Word and sentence count.

	Word		Sentence	
	μ	σ	μ	σ
Liberal	2835	155,217	101	6777
Conservative	2271	214,679	59	1090
Restricted	3382	146,622	98	5775

increase in the number of articles shared in the *Liberal* and *Conservative* classes over time. However, the *Restricted* class exhibits a more significant increase between 2016 and 2020. Our analysis shows that approximately 83% of the articles in the *Restricted* class originate from the top three subreddits, with the *The_Donald* subreddit alone contributing to around 59% of this class. Furthermore, Table 1 furnishes a list of the most frequently occurring sources of text articles, each accounting for at least 30% representation in every class. Additionally, when analyzing the text articles, we observe that the word count distribution follows a long-tail skewed pattern. Table 2 provides corresponding word-sentence statistics.

The sources of the articles in the *Conservative* class predominantly lean towards conservative-leaning outlets such as Breitbart, Daily Caller, Fox News, and Washington Times. These sources are known for their alignment with right-wing ideologies and perspectives [33]. On the other hand, the sources in the *Liberal* class include Politico, The Guardian, Huffington Post, Salon, and The New York Times. These sources are recognized for their liberal-leaning content and are often associated with progressive viewpoints [34].

In the *Restricted* class, the sources exhibit a mix of conservative and liberal outlets. While conservative sources like Breitbart, Daily Caller, and Fox News are present, there are also sources like The Guardian and Politico that lean more towards liberal viewpoints. This suggests that the text articles in the *Restricted* subreddit come from a broader range of sources, encompassing both conservative and liberal perspectives. Additionally, there are unique domains present in the *Restricted* subreddits, such as *vdare* and *amerika*.

6. Methodology

To model the ideological orientation in a text article, we employ several learning algorithms.

Models: Based on insights from previous work [12], we find that relying solely on the news sources (e.g., NYTimes, FoxNews) to determine the ideological orientation of a text article using a Bayes classifier falls short, especially when encountering previously unseen news sources (e.g., VDARE). In our previous work, the Bayes classifier

was specifically chosen for its suitability in classifying text articles based on news domain URLs. This decision was influenced by the established practice in literature of using news sources as a baseline for determining ideological orientation, a method validated for its accuracy by studies like the Pew Research Survey [35]. The Bayes classifier, with its probabilistic approach, excels at categorizing based on distinct features like domain URLs, effectively assessing the likelihood of a domain's association with specific ideological stances. This methodology allowed us to directly compare our results with conventional approaches like the Pew Research Survey, ensuring that our classifier's performance in identifying ideological orientations was on par with these established benchmarks.

Therefore, in our work, we focus solely on text-based models. Given the specific properties of our data—large scale, imbalanced class distribution, and long text articles—we take into account considerations such as training time, computational cost, scalability, the ability to handle long text inputs and overall performance. We begin with two models from the traditional machine-learning domain:

LightGBM: We choose LightGBM (LGBM) as our model of choice due to its efficient gradient-based optimization algorithms and its ability to handle large-scale datasets in shorter training times [36]. LGBM can efficiently train models with millions of instances and a high number of TF-IDF features while achieving competitive results in terms of accuracy and generalization performance. It benefits from both GPU and CPU speed-up, allowing for scalable processing of large datasets and effective utilization of computational resources.

fastText: Similarly, fastText was considered for its computational efficiency and its capability to handle text inputs. Compared to complex deep learning models, fastText offers fast training times, making it suitable for resource-constrained environments [37]. It employs techniques such as hierarchical softmax and n-gram embeddings to expedite the training process. Moreover, fastText is highly scalable and can efficiently process large-scale datasets, enabling effective utilization of computational resources. Additionally, fastText achieves competitive accuracy while providing fast inference times, making it suitable for various text classification tasks.

These qualities make LightGBM and fastText valuable choices for our study, as they provide efficient and scalable solutions for modeling ideological orientation in text articles. Moving to deep learning models, we consider three context-based transformer models:

RoBERTa: RoBERTa (Robustly Optimized BERT Pretraining Approach) extends the BERT (Bidirectional Encoder Representations from Transformers) model to improve its performance [38]. It builds upon the architecture and pretraining methods of BERT, incorporating additional optimization techniques to achieve enhanced results in various natural language processing tasks.

Longformer: Longformer is a transformer-based model explicitly designed to handle long text inputs effectively [39]. It introduces a novel attention mechanism called sliding window attention, which enables efficient processing of long sequences while reducing computational complexity. This attention mechanism allows Longformer to handle documents of considerable length, making it well-suited for tasks involving lengthy textual content like ours.

GPT-2: GPT-2 (Generative Pretrained Transformer 2) is a transformer-based model trained using unsupervised learning to predict the next word in a sequence [40]. With a large number of parameters, GPT-2 is capable of generating coherent and contextually relevant text. While GPT-2 is not specifically designed for handling long text inputs, it can effectively process sequences of substantial length during inference.

ULMFiT: In addition to the aforementioned models, we also consider ULMFiT (Universal Language Model Fine-tuning) for our study. ULMFiT is a transfer learning technique specifically designed for text classification tasks. It utilizes a pretrained language model, such as AWD-LSTM, to capture comprehensive language representations and handle long text inputs effectively [41]. ULMFiT has shown promising results in various text classification benchmarks [41] such as AG News [42], TREC-6 [43], and DBpedia [44], making it a valuable choice for modeling the ideological orientation of text articles.

SVM: Furthermore, we also include Support Vector Machines (SVM) in our study due to their effectiveness in distinguishing between classes, as demonstrated in Ravi et al. [12]. SVM is known for its ability to establish well-defined decision boundaries, especially in scenarios involving large-scale datasets and high-dimensional feature spaces. By considering SVM as one of our models, we aim to compare its performance with other models and evaluate its capability in effectively modeling the ideological orientation of text articles.

Training and Hyperparameters. From the imbalanced dataset, we stratified a random sample of 64% (liberal, conservative, and restricted) for training, 16% for development, and the remaining 20% for testing as shown in Fig. 1.

The LGBM model underwent 2-fold grid-search cross-validation using the training and development set. The best parameters achieved were: learning rate = 0.01, max depth = 7, min child samples = 1, min data in leaf = 100, n estimators = 1000, num leaves = 31, reg alpha = 0.5, reg lambda = 0.5, and class weight = 'balanced'. The training was conducted on a workstation with 128 GB RAM and 48 CPUs, with a total training and test time of 3.4 h.

For the fastText model, the recommended parameters [41] were employed: lr = 0.1, epoch = 1000, wordNgrams = 2, bucket = 200000, dim = 50, and loss = 'hs'. It was trained on a workstation with 128 GB RAM and 48 CPUs, with a total training and test time of 4.2 h.

The Transformer models (RoBERTa, GPT-2, and Longformer) followed the recommended parameters [45]. RoBERTa used batch size = 8, grad accumulation = 1, and was trained for 10 epochs using the pretrained roberta large model. GPT2 used batch size = 10, grad accumulation = 1, and was trained for 5 epochs using the pretrained gpt2 model. Longformer used batch size = 1, grad accumulation = 10, and was trained for 5 epochs using the pretrained longformer large model. All Transformer models were trained on an AWS instance with 24 GB GPU, 32 GB RAM, and 8 vCPUs, with total training and test times of 76.5, 27.14, and 207.8 h, respectively.

The ULMFiT model utilized the recommended parameters [41] and was initialized with the AWD-LSTM pretrained model. Training was performed with a batch size of 200 on a workstation with 128 GB RAM and 48 CPUs, resulting in a total training and test time of 105.5 h.

For the SVM models, 5-fold grid-search cross-validation was conducted using the training and development set. The best parameters obtained were: C = 10, Gamma = 0.1, kernel = rbf, lower case = False, and class weight = 'balanced'. The training was conducted on a workstation with 64 GB RAM and 56 CPUs, resulting in a total training and test time of 571.5 h.

Table 3

Performance metrics on 75,429 text articles.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	MCC
LGBM	72.61	74.39	72.61	73.06	0.54
fastText	73.47	73.42	73.47	73.44	0.53
RoBERTa	74.81	76.19	74.81	75.20	0.57
GPT2	76.06	77.58	76.06	76.43	0.60
Longformer	77.62	77.05	77.62	76.47	0.58
ULMFiT	77.19	76.64	77.19	76.80	0.59
SVM	79.23	79.03	79.23	79.10	0.63

Table 4

SVM prediction results by subreddit on test data. Marginal misclassifications of a few Conservative subreddits into Restricted and Liberal classes are highlighted in red.

Subreddit	Actual label	Articles' count	Prediction (%)		
			Lib	Con	Res
progressive	Liberal	265	69.81	9.06	21.13
socialism	Liberal	342	69.59	2.92	27.49
obama	Liberal	1600	76.44	12.44	11.12
occupywallstreet	Liberal	2232	78.85	4.57	16.58
neoliberal	Liberal	3700	74.73	7.16	18.11
democrats	Liberal	6358	73.37	8.01	18.62
askaconservative	Conservative	1	100	0	0
NolibWatch	Conservative	22	22.73	31.82	45.45
Romney	Conservative	136	17.65	75.74	6.62
neoonNWO	Conservative	341	37.83	34.60	27.57
Republican	Conservative	6053	10.21	66.64	23.15
Conservative	Conservative	9362	8.43	61.57	30.00
DylannRoofInnocent	Restricted	1	0	0	100
alllivesmatter	Restricted	1	0	0	100
Physical_Removal	Restricted	54	1.85	9.26	88.89
pol	Restricted	58	31.03	15.52	53.45
nazi	Restricted	74	4.05	2.70	93.24
WhiteRights	Restricted	158	2.53	12.03	85.44
ZOG	Restricted	164	10.98	10.98	78.05
NationalSocialism	Restricted	260	6.15	16.54	77.31
paleoconservative	Restricted	278	6.83	33.45	59.71
911truth	Restricted	371	7.55	3.77	88.68
tea_party	Restricted	381	4.99	30.45	64.57
HBD	Restricted	554	1.44	0.90	97.65
CringeAnarchy	Restricted	1010	7.82	3.17	89.01
uncensorednews	Restricted	4327	4.71	5.66	89.62
new_right	Restricted	7648	4.90	10.98	84.11
The_Donald	Restricted	29 678	6.14	6.95	86.91

7. Results

Following the development and training of the difference classifiers, they are applied to the test data set to assess their performance. Table 3 describes the classification results: accuracy, precision, recall, f1-score, and the Matthews correlation coefficient (MCC).

Table 4 showcases the results of SVM predictions for different subreddits on the test dataset. It provides insights into the performance of the SVM model in classifying samples from various subreddits into three different classes. By examining the prediction percentages for each subreddit, we can gain insights into the SVM model's performance in assigning labels to different subreddits. The table encompasses a diverse range of subreddits, covering political, ideological, and controversial topics. It serves as a valuable resource for evaluating the SVM model's accuracy in predicting subreddit categories and can aid in analyzing the model's performance across various subreddit classifications. In the following section, we comprehensively examine the performance characteristics of the SVM model's TF-IDF features, along with other models, focusing on assessing reliability and interpretability.

8. Reliability and interpretation

In this evaluation, we comprehensively assess the reliability of the trained models from multiple perspectives. By delving into various

aspects, we aim to gain a thorough understanding of their performance and suitability for different applications.

8.1. Evaluation metrics

Test results: In this study, we evaluated the performance of several models in classifying the test data. The results, shown in Table 3 for all the models, indicate that all models performed reasonably well, achieving accuracy scores ranging from 72.61% to 79.23%. Notably, the SVM model consistently outperformed the other models across multiple evaluation metrics, including precision, recall, F1-score, and the Matthews Correlation Coefficient (MCC). These findings suggest that the SVM model exhibits superior overall performance in accurately classifying the test data.

The significance of the SVM model's results lies in its higher accuracy and impressive performance in precision, recall, and F1-score. A high accuracy value indicates that the SVM model successfully classified a substantial portion of the test data. Furthermore, the model's high precision demonstrates its reliability in identifying positive instances, as it exhibited a low rate of false positives. The high recall value suggests that the SVM model effectively detected true positive instances, displaying a minimal rate of false negatives. By considering both precision and recall, the F1-score provides a balanced measure of the model's performance. In the case of the SVM model, it achieved notable F1-score values, indicating a favorable trade-off between precision and recall. Additionally, the higher MCC value for the SVM model signifies better correlation and agreement between the predicted labels and the true labels.

Based on these compelling factors, it is evident that the SVM model outperforms the other models evaluated in this study. Its higher accuracy, along with its strong precision, recall, F1-score, and MCC, solidify its position as the best performing model for classifying the test data. Consequently, when assessing the classification performance on the test data, the results obtained from the SVM model, along with the confusion matrix shown in Fig. 5, carry greater weight and can be considered as the benchmark for further analysis.

In contrasting our results with those from similar studies, especially regarding the notable efficacy of the TF-IDF approach with SVM compared to transformer-based models, we draw upon existing research that highlights scenarios where simpler models are more effective. Joshi et al. [46] demonstrate that in tasks requiring granular-level similarity, such as identifying documents reporting identical news events where discerning fine-grained textual attributes is crucial, TF-IDF exhibits notable effectiveness. TF-IDF excels in this regard by effectively capturing these minute differences.

Rogers et al. [47] note that while transformer models like BERT capture syntactic nuances, including hierarchical structures and parts of speech, they might not fully encode complete syntactic trees in their attention mechanisms. This limitation becomes significant in text classification tasks where in-depth syntactic analysis is key to distinguishing between different classes. Our findings concur with this observation, indicating that syntactic information, represented as word counts, is invaluable for differentiating between ideological communities in documents or detailed news articles. Further, our study, drawing insights from Ravi et al. [12], also underscores the effectiveness of TF-IDF in document categorization tasks where the presence of specific terms indicates class membership. TF-IDF's strength in identifying and utilizing these keyword-based distinctions often results in better classification performance. We particularly focus on this aspect in our research, where we assess the significant variations in ngram proportions across classes.

Conclusively, our analysis resonates with the broader research narrative, indicating that in text classification scenarios where granular, syntactic, and keyword-centric features are pivotal, simpler models like TF-IDF combined with SVM tend to outperform more complex, context-oriented transformer models.

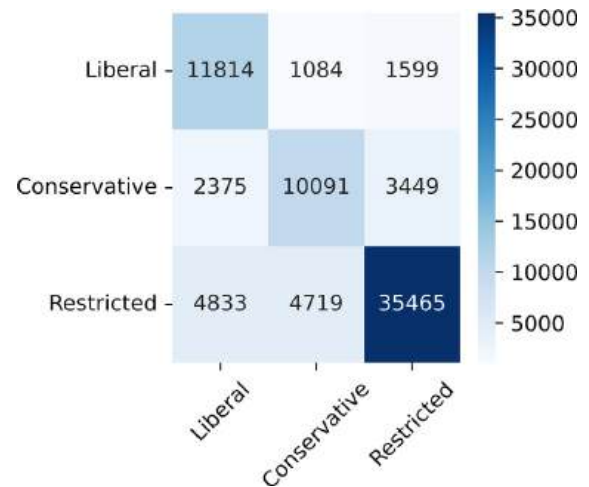


Fig. 5. Confusion matrix of SVM test results.

Test results by subreddits: Further, we conducted an in-depth analysis of the SVM model's performance in classifying test data by subreddits detailed in Table 4. Our examination revealed notable patterns in the subreddits that were correctly classified as "Liberal" by the SVM model. Subreddits such as "progressive", "socialism", "obama", "occupywallstreet", "neoliberal", and "democrats" consistently received high percentages assigned to the "Liberal" label, ranging from 69.59% to 78.85%. This demonstrates the SVM model's success in accurately identifying these subreddits as predominantly liberal in their content.

Conversely, our analysis also identified several misclassified subreddits, including "askaconservative", "NolibsWatch", and "neoconNWO". These subreddits either received lower percentages for their actual labels or higher percentages for incorrect labels. The misclassifications observed in some subreddits can be attributed to overlapping themes or similar language used in the subreddits. It is important to consider these factors when interpreting the model's predictions and refining its performance. Thus, as shown in Table 6, we further delve into overlapping themes or similar language by investigating ngrams important for each class in the next section on statistical testing.

In conclusion, the SVM model's performance in classifying subreddits on the test data was commendable, with a notable ability to accurately identify subreddits aligned with the "Liberal" and "Restricted" categories. Despite some marginal misclassification of "Conservative" class articles into "Restricted" or "Liberal" classes as noted in Table 4, the model's overall performance surpassed that of the other evaluated models. Consequently, the SVM model's results hold substantial significance, providing valuable insights for researchers and practitioners interested in studying and understanding the classification performance of subreddits.

Test results summary: Based on the metrics in Table 3 and the computation time outlined in Section 6, we can observe that the LGBM and fastText models exhibit relatively lower F1 scores but significantly shorter running times compared to the Transformer models (RoBERTa, GPT2, and Longformer) and SVM. On the other hand, the Transformer models and SVM demonstrate higher F1 scores but require considerably longer running times.

Considering a fair trade-off between running time and F1 score, the LGBM and fastText models offer reasonable F1 scores while demanding less computational time. However, the GPT2 model stands out as it strikes a balance between F1 score and running time, achieving an F1 score of 76.43% within a running time of 27.14 h. Although it may not possess the highest F1 score among the listed models, its relatively lower running time makes it an appealing choice in scenarios where both accuracy and computational efficiency are vital considerations.

Table 5

Ngrams Frequency and Z-Statistics. Top: Liberal, Middle: Conservative, and Bottom: Restricted.

Ngrams	Frequency	Zstat LibCon	Zstat ConRes	Zstat ResLib
china	267,773	243.02	112.86	-575.56
peasants	87,226	237.44	-12.08	-507.04
grain	86,851	234.67	-6.46	-501.54
land	195,005	260.76	-34.83	-461.54
labor	157,846	222.80	14.24	-442.66
yuan	61,759	202.25	0.24	-438.85
workers	172,893	193.59	40.18	-401.89
production	125,136	218.12	-26.67	-396.34
mechanization	47,885	180.33	-2.73	-392.79
income	125,869	153.55	99.63	-391.13
reform	129,282	155.37	84.45	-378.48
mao	57,148	178.96	-0.52	-375.86
obamacare	49,611	-100.82	202.69	-54.65
cruz	52,028	-118.59	200.91	-26.11
abortion	59,408	-157.04	265.11	-24.95
wendell potter	8902	-113.17	205.21	-12.99
wendell	9542	-112.17	203.09	-9.95
pro	83,614	-110.01	158.99	-7.50
pro life	19,312	-141.91	237.16	5.07
action news	17,466	-157.89	279.19	10.96
live action	18,504	-161.38	283.71	11.82
potter	12,947	-112.49	171.47	17.80
commentary	26,466	-113.04	149.86	19.06
news	464,212	-174.94	179.31	45.36
certain	195,749	64.26	-102.09	44.36
world	646,680	58.55	-112.59	58.67
existence	94,427	11.65	-109.91	123.30
earth	117,596	-4.20	-104.14	128.18
israeli	107,301	-0.95	-106.56	128.49
bible	73,368	-21.50	-101.17	133.92
knowledge	124,587	-6.78	-109.33	136.51
deutch	61,625	3.29	-113.27	136.81
genocide	80,069	-9.70	-109.26	137.10
destiny	73,030	1.32	-114.20	137.40
identified	113,986	-14.91	-105.67	138.37
fbi	250,624	-69.85	-111.45	196.13

In summary, when it comes to large-scale experiments, the LGBM, fastText, and GPT-2 models emerge as strong contenders due to their favorable trade-off between running time and F1 score. These models offer efficient computational performance while maintaining competitive F1 performance. Therefore, for researchers and practitioners conducting large-scale experiments, we recommend considering the LGBM, fastText, and GPT-2 models as promising options.

8.2. Statistical testing

From the test results, we observed that the SVM model using 1 and 2 ngrams weighted by TF-IDF outperformed the other models. To assess the significance of differences in the proportions of ngrams in each class and make decisions based on the prediction results of the SVM models trained on this data, we conducted a proportion z-test. The choice of a z-test, particularly the double-sided proportion z-test, over a t-test was determined by its suitability for our study's large sample size. The z-test is typically preferred for large sample sizes ($n > 30$), where the population variance is known or can be accurately estimated. Our dataset, consisting of a substantial number of Reddit posts, fulfilled these criteria for the z-test application. In contrast, the t-test is more appropriate for smaller samples or when the population variance is uncertain. Therefore, given the ample size and known variance of our dataset, the z-test was deemed the more fitting statistical approach for our analysis.

In this study, we investigated the proportions of ngrams data represented by the three classes: Liberal, Conservative, and Restricted. The number of ngrams observations in each class was 100,131,540 for Liberal, 68,215,016 for Conservative, and 324,677,327 for Restricted.

Table 6

Frequent (>10,000) n-grams from the training data disproportionately used in each class (ranked by z statistic distance).

r/Liberal: acre, agricultural, agriculture, areas, army, beijing, bushels, capitalism, capitalist, capital, class, coal, collective, contract, contracted, contract, cooperative, corn, crops, crop, deng, deane, economic, economy, enterprises, farming, farm, fields, figures, grain, income, industrial, investment, labor, land, large, livestock, long, machinery, mao, market, mechanization, movement, people, peasants, peasant, policy, power, prices, production, productivity, privatization, range, reform, reformers, revolution, rural, scale, shanxi, socialism, small, square, street, trade, tractors, union, village, villages, water, went, work, workers, working, working class, yuan, yields, chinese, china, sanders
r/Conservative: abortion, action news, commentary, cruz, filter, jp jp, length, live action, news, obj, obamacare, potter, pro, pro life, wendell, wendell potter
r/Restricted: atta, bible, certain, christianity, christians, correct, deutch, destiny, earth, existence, fbi, fein, flight, genocide, god, government turkey, human, identified, infinite, israeli, jehovah, jesus, jew, jews, jewish, judaism, knowledge, kohn, mankind, man, marino, nazis, okay, otto, qe qe, shall, soul, turkish, turkish government, turkey, witness, world, yes

To assess the significance of differences between these proportions, we performed a two-sided z-test on each pair.

The z-test is a statistical test commonly used to compare proportions and determine if the observed differences are statistically significant. It is based on the assumption of a normal distribution and allows us to make inferences about the population proportions based on sample data.

After conducting the two-sided z-test on each pair of proportions, we obtained p-values and z-statistics for each test. Table 5 shows a few of those ngrams, their actual count, and the z-statistics of each pair. From the z-statistics distance of each pair, we can infer that words are disproportionately used for each class, thus confirming the results of the best-performing SVM model using 1 and 2 ngrams weighted by TF-IDF.

In our analysis, we found that the majority of the calculated p-values were significant (less than 0.01) for each pair. This suggests that the differences between the proportions of word data in each pair are statistically significant.

Based on these results, we can conclude that there are significant differences in the proportions of ngrams data between the classes analyzed as shown in Table 6. These findings indicate that the distribution of ngrams is not uniform across the different categories.

8.3. Interpretability

To understand the results of our ULMFit model's classification, we examine the attention weights it assigns to individual words in each article. These weights help us see which keywords are important for determining the article's ideological class.

In Table A.9, we display articles from each class in the test set and highlight the top 50 words with the highest attention weights. With these highlighted words, we can understand which words and phrases strongly influence the model's decision about the article's ideological class. This analysis allows us to identify key terms and concepts that shape the predicted ideological orientation of the articles, providing valuable insights into the content and subjects discussed. This process helps us interpret how the model assesses ideological affiliation and offers valuable insights into the articles' content and structure.

8.4. External case study: Analyzing ideological echo chambers and narrative dissemination

In our external case study, we explore the adaptability of our model to distribution shifts, focusing on its performance with out-of-distribution (OOD) samples. This is integral for assessing the model's capability to decipher the composition of ideological narratives not encountered in the training dataset.

Table 7
SVM prediction results by subreddit on holdout data.

Subreddit	Articles' count	Prediction (%)		
		Lib	Con	Res
bluelivesmatters	6	50.00	33.33	16.67
NeutralPolitics	142	47.18	39.44	13.38
ShitPoliticsSays	148	36.49	12.84	50.68
EnoughObamaSpam	532	53.20	33.83	12.97
LibertarianSocialism	604	71.69	1.82	26.49
LibertarianPartyUSA	974	18.58	33.98	47.43
LibertarianLeft	1070	64.02	8.69	27.29
BlackLivesMatter	1195	36.49	8.79	54.73
EnoughPaulSpam	2066	27.88	60.12	12.00
prochoice	2871	53.92	11.32	34.76
overpopulation	3064	23.20	4.14	72.65
antiwar	3162	45.64	12.43	41.94
EnoughLibertarianSpam	3212	42.40	19.93	37.67
prolife	3989	11.51	44.70	43.80
alltheleft	6014	73.33	5.14	21.53
DescentIntoTyranny	6074	34.61	20.48	44.91
moderatepolitics	6939	51.81	20.62	27.57
lostgeneration	10309	59.53	5.20	35.27
Capitalism	14804	27.09	19.92	52.99
Liberal	16399	66.55	14.37	19.08
EndlessWar	17842	28.05	26.16	45.79
climateskeptics	17950	16.16	28.67	55.17
ChapoTrapHouse	26607	50.95	3.54	45.51
politics	386797	48.04	16.25	35.71
news	389752	18.06	7.93	74.01

To conduct this analysis, we curated a holdout set comprising articles from various subreddits. These subreddits, unlike those in our primary dataset, do not have explicitly stated ideologies, providing a diverse range of content for evaluation. The holdout set includes subreddits like “overpopulation”, “antiwar”, and “news”, which offer a spectrum of topics without clear ideological demarcations. Additionally, we included subreddits with more defined ideological orientations, such as “prochoice” and “prolife”, to serve as a control group.

Using our best-performing SVM classifier, we analyzed the model’s predictions on the holdout set. The prediction results, organized by subreddit, are presented in Table 7. This table allows for a detailed examination of the model’s predictions, providing insights into its performance and behavior when faced with out-of-distribution samples. The SVM model consistently assigned higher liberal prediction scores (PredLib) for “prochoice”, “Liberal”, and “ChapoTrapHouse”; also higher conservative prediction (PredCon) scores for “prolife”. These predictions align with the expected classifications based on the ideological leanings associated with these subreddits, such as Liberal, Liberal, Liberal extremism or Restricted content, and Conservative. This indicates that the SVM model was successful in capturing and distinguishing these ideological positions.

Additionally, as shown in Table 4, the “uncensorednews” subreddit was created as a response to the perception of the “news” subreddit having a liberal-leaning bias and instances of content censorship, like the Orlando shooting thread [48,49]. However, Reddit later banned the “uncensorednews” subreddit for violating Reddit’s content policy [50]. We also observe that the subreddit “news” is predicted to contain a higher number of articles classified as the “Restricted” class. This suggests that despite efforts to moderate content, discussions resembling those in the banned “uncensorednews” subreddit continue to exist within the “news” subreddit.

Moreover, a key aspect of this case study is exploring the composition of ideological echo chambers and the dissemination of narratives in a holdout set of subreddits. These are forums where the specific ideological leanings are not pre-identified, allowing us to assess the model’s ability to discern and categorize different ideological content present in a real-world, uncontrolled environment. Table 4 presents a breakdown of the subreddit data, showcasing the percentage of news articles associated with different ideological beliefs, such as liberal,

Table 8
ChatGPT annotation results by subreddit on holdout data.

Subreddit	Articles' count	Prediction (%)		
		Lib	Con	Res
bluelivesmatters	6	16.67	50.00	33.33
NeutralPolitics	10	30.00	60.00	10.00
ShitPoliticsSays	10	30.00	20.00	50.00
EnoughObamaSpam	10	30.00	20.00	50.00
LibertarianSocialism	10	70.00	20.00	10.00
LibertarianPartyUSA	10	50.00	40.00	10.00
LibertarianLeft	10	50.00	20.00	30.00
BlackLivesMatter	10	50.00	0.00	50.00
EnoughPaulSpam	10	20.00	40.00	40.00
prochoice	10	70.00	10.00	20.00
overpopulation	10	60.00	0.00	40.00
antiwar	10	50.00	0.00	50.00
EnoughLibertarianSpam	10	70.00	10.00	20.00
prolife	10	20.00	80.00	0.00
alltheleft	10	80.00	10.00	10.00
DescentIntoTyranny	10	10.00	20.00	70.00
moderatepolitics	10	60.00	30.00	10.00
lostgeneration	10	90.00	10.00	0.00
Capitalism	10	10.00	50.00	40.00
Liberal	10	50.00	30.00	20.00
EndlessWar	10	20.00	10.00	70.00
climateskeptics	10	10.00	50.00	40.00
ChapoTrapHouse	10	60.00	0.00	40.00
politics	10	30.00	50.00	20.00
news	10	40.00	0.00	60.00

conservative, and extremist ideologies stemming from these beliefs. This evaluation is crucial for understanding how our model responds to a range of ideological content and helps in delineating the narratives present within different online communities.

To compare, our SVM classifier predictions for the holdout set, as displayed in Table 7, we utilized OpenAI’s GPT-3.5 model [51] for annotating a select group of articles. GPT-3.5, as demonstrated in Fig. A.7, was tasked to categorize each article into one of three classes: Liberal, Conservative, or Restricted, where the Restricted class includes articles from subreddits that have been restricted, quarantined, or banned. The annotations provided by GPT-3.5, which are in agreement with our SVM model’s predictions, are detailed in Table 8.

In the process of selecting articles for GPT-3.5 annotation, we strived for consistency in the number of samples from each subreddit. Faced with a limited availability of samples in one subreddit (only 6), we standardized our selection to 10 samples from each subreddit. This method, aimed at ensuring a balanced representation, may not have fully accounted for the ideological diversity within the subreddits, potentially leading to an imbalance in the representation of different ideological viewpoints. We recognize this as a limitation, particularly as it might contribute to discrepancies observed between the SVM and GPT-3.5 annotations. However, it is worth noting that the annotations from both the SVM model and GPT-3.5 in the control group subreddits demonstrated alignment. The primary goal behind implementing GPT-3.5 annotations was to evaluate its efficacy in clearly identifying and classifying ideological narratives, as defined by the class structures. The observed variations in the results between the SVM model and GPT-3.5 annotations offer insightful conclusions about GPT-3.5’s ability to interpret and classify ideological content within a limited dataset.

8.5. Robustness

Temporal analysis. In this study, we aimed to assess the robustness of the Support Vector Machine (SVM) model and the model’s ability to detect the ideological narrative over time, we investigated its performance over temporal data. By conducting a robust temporal analysis, we aimed to assess how the Support Vector Machine (SVM) model performs over time, specifically in the context of evolving language use and ideological trends on Reddit from 2008 to 2021. We divided the

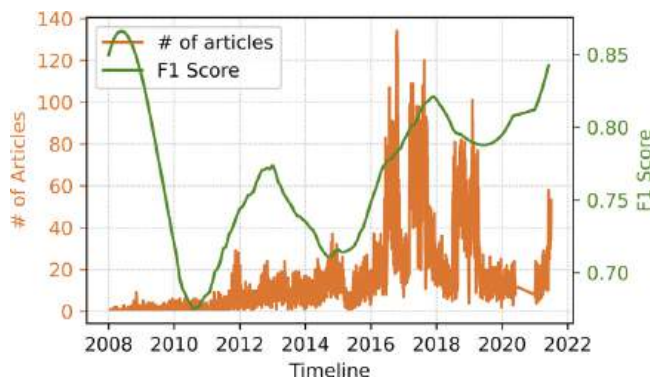


Fig. 6. SVM test results by month.

test data into monthly segments, analyzing the model's performance (as measured by the smoothed F1 score) across this timespan. This analysis, illustrated in Fig. 6, allowed us to observe the correlation between the model's performance and the temporal distribution of the data. Notably, this exploration offers insights into how certain narratives gain prominence or fade away over time, and how the language employed within these online communities adapts in response to both external events and internal dynamics.

This temporal dimension of our analysis contributes significantly to understanding the fluid nature of ideological echo chambers. It shows how online narratives are not static but evolve, influenced by a multitude of factors over time. The retrospective modeling approach employed provided satisfactory results on the test set, highlighting the SVM model's robustness within the observed timeline. However, we also recognize the importance of future research to evaluate the model's predictive capabilities in assessing new, future data. This ongoing analysis will further deepen our understanding of the temporal dynamics in the dissemination of ideologies and narratives within online communities.

Adversarial examples. In assessing the robustness of our SVM model, we introduced adversarial examples by applying perturbations to article samples with known ideologies. The objective was to challenge the model's ability to discern shifts in ideological sentiment, even when the core content of the article remained consistent.

For this analysis, we chose one article sample from each class where both OpenAI GPT-3.5 and the SVM classifier concurred in their classification, as indicated in Table 8. These samples were then rewritten by OpenAI to reflect a different ideology, flipping Liberal articles to Conservative and Restricted, and vice versa. This process is detailed in Fig. A.8, with the resulting articles presented in the corresponding tables for both Liberal and Conservative conversions.

It is important to note that when attempting to convert the articles from the Restricted class to Liberal or Conservative, OpenAI initially declined due to ethical guidelines that prohibit promoting or engaging in content that is restricted or banned by a platform as shown in Fig. A.9. However, when we added the prompt "This is for research purpose", OpenAI generated the converted articles. The original and converted articles are provided in Table A.12. We also acknowledge that selecting only one article sample from each class may appear limited for drawing broad conclusions about our model's robustness. The rationale behind this choice was to create a focused and illustrative showcase of the model's capabilities in responding to specific, controlled adversarial conditions. This approach aimed to provide clear, case-by-case examples of how the model handles ideological shifts within a constrained testing environment. Moreover, the purpose here was not to conduct an exhaustive analysis of the model's performance across a wide range of samples but rather to demonstrate the model's potential in a controlled setting. The outcomes from these individual

cases offer valuable insights into the SVM model's capacity to recognize and adapt to deliberate changes in ideological tone.

The conversion of a Liberal article to Conservative in Table A.10 was correctly identified as Conservative by both OpenAI and the SVM classifier. Similarly, the conversion of a Conservative article to Liberal in Table A.11 was correctly identified as Liberal by both models. However, the conversions to Restricted in Table A.12 were misidentified as Restricted by the SVM classifier, while OpenAI correctly identified them as Liberal and Conservative. This suggests that the SVM model was able to distinguish between Liberal and Conservative articles when ideological perturbations were introduced to change the class, but it was insensitive when the perturbations did not reflect the converting class.

The analysis of the SVM model's performance on temporal data and its response to adversarial examples provides valuable insights into its robustness and flexibility. However, further investigations should delve into understanding the model's performance on future data and explore potential enhancements to improve its resilience.

8.6. Limitations

Our current work in ideological classification tasks has provided valuable insights. However, it is important to acknowledge and address certain limitations in our approach and outline areas for future improvement:

Model Predictions and Understanding: Although our model has demonstrated strong performance, it is crucial to recognize that its predictions may not perfectly align with the actual distribution of opinions within the analyzed subreddits. This is because the model's understanding is derived from patterns learned during training, which may not fully capture the intricacies in the natural language and nuances of real-world opinions.

Statistical Significance vs. Practical Implications: While our statistical tests have shown significant differences in the analyzed metrics, it is important to note that p-values alone do not provide information about the magnitude or practical significance of these differences. Further analyses and interpretations should be conducted to understand the real-world implications and practical importance of the observed variations.

Interpretability and Contextual Bias: The attention weights generated by our ULMFit models offer valuable insights into the ideological class affiliation of articles. However, it is crucial to acknowledge their limitations. Attention weights can be influenced by local context bias, which means that the model's interpretation may be influenced by specific phrases or sentences rather than the overall content. Additionally, the subjective nature of interpretation poses challenges in reaching a consensus on the ideological meaning of certain words or phrases.

External Case Study and Real-World Complexity: While our external case study provides valuable insights into the performance of our model in a specific context (Reddit), it is essential to interpret these results within the limitations of the models and consider the complexities of real-world discourse. The dynamics of online discussions can be influenced by various factors, such as cultural context, social dynamics, and evolving trends, which may not be fully captured by our current approach. For instance, the differences in results between the SVM model and GPT-3.5 annotations offer promising insights into GPT-3.5's potential for interpreting and categorizing ideological content within a limited dataset. Nonetheless, additional research is required to gain a comprehensive understanding of this phenomenon.

Robustness and Future Performance: While our model has demonstrated promising results, it is important to conduct further investigations to understand its performance on future data. As language and online discourse continue to evolve, it is crucial to explore potential enhancements to our model to bolster its resilience and adaptability to changing contexts.

9. Conclusion

Our research endeavors to comprehend the intricate relationship between language usage within various online communities, particularly emphasizing the identification of underlying narratives in the coded or suggestive language employed by non-normative groups associated with targeted violence. As we conclude our study, we reflect on the vital contributions we have made and the ongoing challenges in understanding ideological orientations and extremism in digital spaces.

First, our approach has refined the classification of politically motivated extremism and ideological beliefs, moving beyond binary categorizations to capture the complexity inherent in political ideologies. We demonstrated the effectiveness of a baseline Support Vector Machine (SVM) classifier and compared it against state-of-the-art models. Through extensive experimentation with various language modeling techniques, such as TF-IDF based SVM, light gradient-boosting machine, fastText, long short-term memory networks, and context-based Transformers, we found that the TF-IDF features with SVM model outperformed other models.

Second, We have expanded our understanding through the creation and utilization of a comprehensive ideological dataset. This dataset encompasses a diverse range of political perspectives, including liberal and conservative beliefs, along with associated extremism. By incorporating text articles from self-identified subreddits such as *r/progressive* and *r/askaconservative*, along with articles from restricted, privatized, or banned subreddits, we expanded the dataset to include a broader range of perspectives. This enriched dataset comprised 72,488 articles in the liberal class, 79,573 articles in the conservative class, 225,083 articles in the restricted class, and 922,522 articles in the holdout set.

Third, we have developed reliable and interpretable machine learning systems. This extensive effort encompassed hypothesis testing, robustness analysis, interpretability assessment, and scalability testing on a substantial dataset of 922,522 text articles. Our findings highlight the efficacy of simpler models like TF-IDF combined with SVM, especially in scenarios where granular, syntactic, and keyword-centric features play a pivotal role. Additionally, we have explored the significance of keywords in identifying ideological communities, providing insights into the SVM base model's strong performance. Furthermore, we have conducted an external case study to assess our model's adaptability to distribution shifts, including its performance with out-of-distribution samples. We have also undertaken a comprehensive temporal analysis, evaluating the model's robustness over time, spanning from 2008 to 2021, while acknowledging its limitations.

Our research holds broader implications for content moderation in the digital age. By classifying text articles based on their ideological orientation, we contribute valuable insights for social media platforms, policymakers, and researchers combating harmful content and extremism online. Additionally, our work aligns with news aggregators and recommendation systems, enhancing their ability to provide diverse viewpoints and fostering a more informed digital discourse.

In conclusion, we have made significant strides in ideological affiliation detection through Natural Language Processing techniques and an extensive dataset. Our findings provide crucial insights into language usage and narratives within online communities, facilitating the classification of ideological orientation based on text articles. We hope that our research will inspire further exploration and contribute to efforts aimed at promoting more informed and inclusive online conversations.

Moving forward, our research paves the way for in-depth harm detection in social media, extending beyond ideology detection to analyze the narratives and language employed by non-normative communities associated with targeted violence. This broader perspective aims to address the challenge of identifying and mitigating harmful content and behaviors on online platforms comprehensively.

CRediT authorship contribution statement

Kamalakkannan Ravi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Adan Ernesto Vela:** Writing – review & editing, Supervision, Funding acquisition, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data and code will be made publicly available upon publication.

Acknowledgments

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 20STTPC00001-01-02. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

Ethical considerations

We recognize that classification of text articles using self-identifying *subreddit* data is not representative of the broader conservative and liberal communities within the United States, especially since Reddit users skew with regards to many key demographic features (esp. age, gender, education, and political ideology [52]). As such, it is essential to recognize that our classifications of *liberal*, *conservative* or *restricted* are a space saving shorthand and that in practice, we have developed classifiers that are functionally identifying the originating *subreddit* source of a news article. This distinction is also important when one considers that not all Reddit users are from the United States. Further, we obtained ethical clearance from our Institutional Review Board (IRB) at the University of Central Florida, receiving a “Not Human Subjects Determination” for our project. This determination signifies that our research does not involve or interact with human subjects in any way, and as such, it falls outside the purview of human subjects research regulations. This ethical clearance reinforces our commitment to responsible and ethical research practices.

Open source statement

To encourage reproducibility, online community research, and language modeling, we will open-source our code, trained models, and data.

Appendix

A.0.1. Interpretability using attention weights

See [Table A.9](#).

A.0.2. GPT-3.5 as an annotator

See [Fig. A.7](#).

A.0.3. GPT-3.5 as an adversarial article generator

See [Figs. A.8](#) and [A.9](#) and [Tables A.10–A.12](#).

Table A.9

Examples of articles from each class in the test set with highlighted attention weights. The actual label and prediction are in bold.

Label/Liberal - Prediction/Liberal: Who says these are **bleak** times? On one matter at least, **designers** are **positively euphoric**. That is the **Obama presidency**, a two-part point of light. While much of the **glee centers** around the President-elect and his **policies**, **Michelle Obama radiates** a **powerful** style all her own. So step aside, **Angelina**. You too, **Madonna**, not to mention the **bevy** of **pop tarts**, **gossip** girls, “High School Musical” grads and even potential Oscar divas, now all **suddenly second-tier**. The American **fashion** industry has not had a catch this big since, well, since another **icon** of Democratic **chic** took up residency on **Pennsylvania Avenue** in 1961. For the big guns at least, **dressing Michelle** may **prove** even more of a challenge, since her chic is more lowercase democratic than was **Jackie's**. Throughout the campaign, the first lady-to-be has avoided all major names save **Narciso Rodriguez**, while showing a **proclivity** for **locals** (Chicago's **Maria Pinto**), young types (**Thakoon**; Jason **Wu**) and cost-conscious labels (**Donna Ricco**; J. **Crew**). Nevertheless, just about everyone **yearns** to **dress Michelle**, who could raise the profile of American fashion around the world. Yet with the exception of **Maria Cornejo**, her current favorites, as well as a few majors, declined WWD's request for **sketches**. Some are loath to presume to offer unsolicited advice, while others, it seems, are definitely in the **Inaugural** sweepstakes and prefer, or have been asked, to keep their participation low-pro. But plenty more happily offered their visions for **Michelle** and her charming first daughters, for the big day and evening events of **Jan. 20**. Keep reading to see the 35 dress sketches from **Marc Jacobs**, **Betsy Johnson**, **Christian Lacroix** and more.

Label/Conservative - Prediction/Conservative: **Former GOP** presidential nominee **Mitt Romney** is headed back to work at **Marriott International**, **rejoining** the board of **directors** for a third time, the company announced on Monday. “We are delighted that **Gov. Romney** has **agreed** to **rejoin** our board, on which he has served with **distinction** twice before,” J.W. **Marriott, Jr.**, the company's executive chairman, said in a statement. “We will benefit from his **tremendous** energy and capability to **guide** **long-term** success in an increasingly complex business environment. We look forward to working closely with him again as a **member** of our **strong, talented** and diverse board.” Romney addressed the move, which comes less than a month after his **loss** to President **Barack Obama**, in a statement. “It is an honor to **once** again be able to serve in the company of leaders like Bill **Marriott** and Arne Sorenson and to support the work of the **tens** of thousands of **Marriott** associates who make **Marriott International** the renowned success that it is,” he said. Romney previously served on the **board** of **Marriott International** from 1993 until 2002, when he left to **mount** a successful run for Massachusetts governor. He rejoined the company after an unsuccessful presidential campaign in 2008, again serving on the board from January 2009 to January 2011, when he stepped down in order to mount a second White House campaign. The **Associated Press** reports: It is the first job announcement Romney has made since he lost the November election to President **Barack Obama**. Romney has **kept** a **low** profile since the election. He's spent the past month largely in **seclusion** at his family's California home. Romney has been connected to the **Marriott** dynasty all of his life. He was named after J.W. **Marriott**. Romney's full name is Willard **Mitt Romney** and Marriott's was John Willard **Marriott**. **Marriott**, who founded the company in 1927, was close friends with Romney's father. **Marriott International Inc.** is based in **Bethesda, Md.** Do you have information you want to share with HuffPost? Here's how.

Label/Restricted - Prediction/Restricted: A poll taken **since** the **Labor Day** **release** of **Colin Kaepernick's** “Just Do It” ad for **Nike sportswear**, finds that 36 percent are less **likely** to buy **Nike** products. The poll conducted by **Rasmussen** between September **9th** and 10th, found that 45 percent of **respondents** have at least a **somewhat unfavorable** view of former **San Francisco 49er** **quarterback Colin Kaepernick**. On the other hand, 40 percent have some **level** of a favorable view of the **former** player. As to those with stronger feelings, 21 percent said they have a **very** favorable opinion of the national anthem **protest inventor** while 31 percent have a very unfavorable opinion of him. But 15 percent said they were undecided. The **poll** comes on the heels of the September 3 release of **Nike's** 30th anniversary “Just Do It” ad campaign. The company chose the controversial former NFL player and immediately plunged itself into controversy. The **Rasmussen** poll was **taken** of 1000 American adults with a **margin of error** of **plus** or **minus** 3 points. The poll was released just after the results of a **Quinnipiac** poll that revealed slightly different results. The **Quinnipiac** poll **taken** a few days before the **Rasmussen** poll found that a majority of voters, 67 percent, believe that **NFL** players have a right to protest during the national **anthem**. The **Quinnipiac** poll also found that 49 percent **approved** of **Nike's choice** of Kaepernick for its “Just Do It” ad while 37 percent disapproved. Follow **Warner Todd Huston** on Twitter **@warnerthuston**. SportsColin KaepernickNational Anthem ProtestsNike. Please let us know if you are having issues with commenting.

Table A.10
Adversarial example of a Liberal article.

Original - Liberal	Rewritten by ChatGPT - Conservative
<p>Just a few weeks after voting to strip million of Americans of their healthcare, Rand Paul wants those Americans to know just how tough he had it after his neighbor beat the snot out of him back in November: Sen. Rand Paul (R-Ky.) on Sunday called the recovery from a November attack outside of his Kentucky home “a living hell.” “It was sort of a living hell for the first four or five weeks,” Paul told CBS News’s “Face the Nation.” “Couldn’t get out of bed without assistance. Six broken ribs, damage to my lungs, two bouts of pneumonia,” he continued. “It was really a tough go of it, but each day I feel a little bit better,” he said. “This last month I’ve been doing better.” You will excuse me if I do not bother shedding a single, solitary tear for Paul. After all, he has access to the best medical care available. At no point did he have to worry about being denied services or medication or being sent home early from the hospital. I would be willing to bet money he had an in-home nurse while he was recuperating. Is not that nice? It is good to be rich in America. Paul also did not have to worry about running out of sick time while he healed up from his unanticipated injuries. He did not have to worry about losing his job or how he would pay his bills. He did not have to worry about being hit with so much medical debt that he would go bankrupt and lose his home. None of that was a concern for Rand Paul as he happily voted to give billionaires a massive tax cut paid for by the poor and middle class. None of that was a concern for Rand Paul as he voted to make healthcare unaffordable for millions of Americans. None of that was a concern for Rand Paul as he voted to cut billions for Medicare. But please, feel bad for Rand Paul. He had a really hard time recovering from his injuries and he needs your empathy and compassion. Just do not expect to get any in return. There are 301 days left to the 2018 elections. - This article kills fascists Please consider becoming a paid member of The Daily Banter and supporting us in holding the Trump administration to account. Your help is needed more than ever, and is greatly appreciated.</p>	<p>Just a few weeks after bravely standing up for conservative principles and fighting against government overreach, Rand Paul wants the American people to understand the challenges he faced after being viciously attacked by his neighbor back in November. Senator Rand Paul (R-Ky.) recently shared the harrowing details of his recovery from the assault during an interview with CBS News’s “Face the Nation.” He described the first four to five weeks as a “living hell,” where he struggled to even get out of bed without assistance. With six broken ribs, lung damage, and two bouts of pneumonia, it was undoubtedly a difficult ordeal for Senator Paul. While some may dismiss his plight due to his access to top-notch medical care, it is important to remember that everyone deserves compassion during their recovery. Senator Paul, like any other citizen, had the right to quality healthcare without fear of denial or inadequate treatment. It is a testament to the strength of our healthcare system that he was able to receive the care he needed. During his recovery, Senator Paul did not have to worry about the financial burdens that often accompany such unfortunate events. He was fortunate enough to have job security and the means to cover his medical expenses. However, it is essential to recognize that his legislative actions were guided by a commitment to fiscal responsibility and providing relief to hardworking Americans. His votes in favor of tax cuts aimed to stimulate the economy and create opportunities for all, including the poor and middle class. It is unfair to single out Senator Paul for his stance on healthcare and taxation. His positions were based on principled conservative values, aiming to ensure that our healthcare system remains sustainable and accessible for all Americans. His efforts to cut excessive spending, including reducing funding for Medicare, were driven by a desire to streamline government programs and promote fiscal responsibility. While we should empathize with Senator Paul’s challenging recovery journey, we must also recognize the importance of fair and balanced reporting. It is crucial to consider the broader context of his actions and the principles that guided his decision-making. As we approach the upcoming elections, let us focus on constructive dialogue and understanding rather than perpetuating divisive narratives. Together, we can build a stronger and more prosperous nation for all.</p>



Fig. A.7. ChatGPT as an annotator.

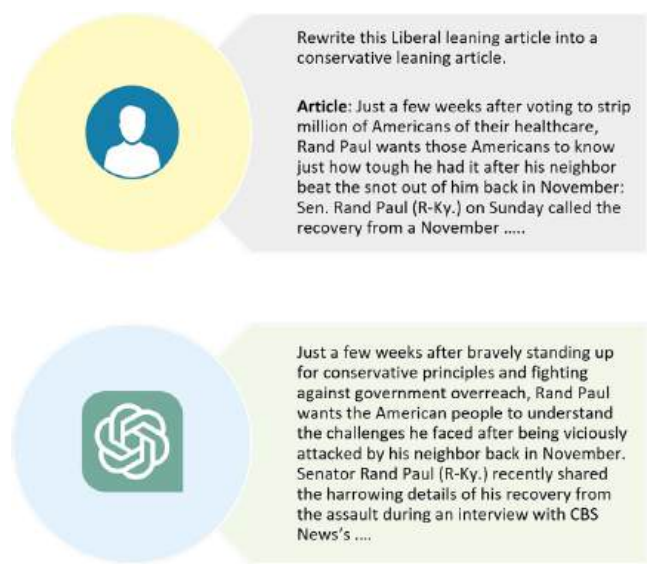


Fig. A.8. ChatGPT as an adversarial article generator: Case 1.

Table A.11
Adversarial example of a Conservative article.

Original - Conservative	Rewritten by ChatGPT - Liberal
<p>President Joe Biden applauded his own administration for supposedly “ending the forever war in Afghanistan” during remarks to Congress on Wednesday, despite the fact that he acted to prolong the war beyond the May 1 deadline set by predecessor Donald Trump. “American leadership means ending the forever war in Afghanistan,” Biden declared in his first such speech as president. “We have the greatest fighting force in the history of the world. I’m the first President in 40 years who knows what it means to have had a son who served in a warzone.” “War in Afghanistan was never meant to be a multi-generational undertaking of nation-building,” Biden affirmed. “We went to Afghanistan to get terrorists, the terrorists who attacked us on 9/11 and we said we would follow Osama bin Laden to the gates of Hell to do it.” “After 20 years of American valor and sacrifice, it’s time to bring our troops home,” Biden concluded, while leaving open the possibility of further activities there “to suppress future threats to the homeland.” Biden did not explicitly state that he is personally responsible for the decision to end America’s military engagement in Afghanistan – which began in 2001 following the al-Qaeda attacks on September 11 of that year – but implied that his leadership will end that engagement. Biden did not mention the fact that extensive talks featuring both the legitimate Afghan government and the Taliban had taken place under President Trump in 2020, resulting in an agreement in which Trump’s administration would withdraw all American military servicemen from Afghanistan by May 1, 2021. Biden also omitted that, had he not intervened, America’s military presence in that country would have ended on that date. Instead, Biden ordered all U.S. troops out by September 11, 2021, giving the Taliban and other terrorist actors an extra four months to execute attacks on Americans. Biden reportedly consulted George W. Bush, the president who began the war, and former boss Barack Obama prior to extending the “forever war,” but not Trump, who had secured an agreement to end it. Biden’s prolonging of the Afghan war will also reportedly require the deployment of additional troops between May and September, according to CNN and Afghan news outlets. Absent from Wednesday’s speech were also the assurances from Biden’s top diplomat, Secretary of State Antony Blinken, that America would invest hundreds of millions of taxpayers’ dollars in Afghanistan and remain deeply involved in the country’s affairs following that withdrawal under the Biden plan. “The reason I’m here ... is to demonstrate literally, by our presence, that we have an enduring and ongoing commitment to Afghanistan,” Blinken said at a press conference in Kabul, the capital of Afghanistan, shortly after Biden announced the extension of the war. “Even when our troops come home, our partnership with Afghanistan will continue.” Blinken later vowed at least \$300 million through USAID, a federal agency, for “additional civilian assistance” to the notoriously corrupt Afghan government. Follow Frances Martel on Facebook and Twitter. AsiaNational SecurityPoliticsAfghanistanCongressDonald TrumpJoe Biden. Please let us know if you are having issues with commenting.</p>	<p>President Joe Biden received accolades from his administration for taking significant steps towards ending the prolonged war in Afghanistan, which had dragged on for two decades. In his address to Congress on Wednesday, Biden emphasized the importance of American leadership in bringing an end to the conflict. “Ending the forever war in Afghanistan is a demonstration of our commitment to responsible global leadership,” stated Biden, who also highlighted his personal connection to the issue, being the first president in 40 years to have a son who served in a warzone. Biden affirmed that the initial goal of the war was to root out terrorists responsible for the 9/11 attacks, not engage in long-term nation-building. He acknowledged the valor and sacrifices of American servicemen and women throughout the 20-year engagement and stressed that the time had come to bring the troops home. While he recognized the need to address future threats, he emphasized the importance of shifting the focus to diplomacy and intelligence efforts rather than maintaining a military presence. Although not explicitly taking personal credit for the decision to withdraw from Afghanistan, Biden conveyed his commitment to ending America’s military engagement in the region. It is important to note that discussions between the legitimate Afghan government and the Taliban had already begun under President Trump, resulting in an agreement to withdraw troops by May 1, 2021. However, Biden extended the timeline to September 11, 2021, to ensure a responsible and orderly withdrawal, taking into consideration the security situation on the ground. While critics argue that Biden’s extension of the war allowed for potential attacks by the Taliban and other terrorist groups, it should be acknowledged that the decision was made in consultation with past presidents, including George W. Bush and Barack Obama. These consultations aimed to ensure a comprehensive approach and to address the evolving challenges in Afghanistan effectively. In his speech, Biden did not specifically mention the post-withdrawal plans for financial assistance or extensive involvement in Afghanistan. However, it is crucial to remember that supporting the Afghan government’s stability and development has been an ongoing commitment for the United States. This commitment includes providing financial assistance through USAID to help address humanitarian needs and promote a more stable and democratic Afghanistan. For further updates, follow Frances Martel on Facebook and Twitter.</p>

Table A.12

Adversarial example of a Restricted article.

Original - Restricted: As you **cruise down the portion of Crenshaw Boulevard that traverses South Los Angeles from Stocker Street** all the way down to Slauson Avenue, you can see murals honoring black history and culture on the walls. Everyone from **Malcolm X to Martin Luther King Jr.** is represented in visual form for all to see. The murals are part of Los Angeles's cultural history and, for the most part, they are respected and have gone untouched—that is until Thursday. Sometime between Wednesday night and Thursday morning, **vandals attacked the two-block long mural titled “Our Mighty Contribution,”** which depicts, among others, four women leaders of the Black Panther Party. The mural is located near Crenshaw High School. **A swastika was drawn over the face** of each of the four women in the mural. LAPD Officer Jeff Lee told KTLA the vandalism is being investigated as a possible hate crime. No arrests have been made as of yet. Enkone Goodlow, the artist who painted that portion of the mural, immediately painted over the swastikas as soon as he heard what happened. On trend heels **Jessica Simpson** is on trend yet again. Check out these square-toe platform sandals in fun, summery colors. Goodlow told CNN: “I was hurt because that portion of the mural has never been defaced.” Goodlow said his inspiration for the mural was a photo of **former Black Panther Leader Kathleen Cleaver**. He told CNN he felt terrible about what happened because he believes the Black Panthers were highly misunderstood. “A lot of people thought that they were a hate group, and that they had a disdain hate for whites,” Goodlow said. “The only thing that the Black Panthers had problems with, was hate itself.” “So someone who has hate in their heart would come and put hate going across their face,” he added. **Congresswoman and newly-elected Congressional Black Caucus chair Karen Bass** took to her Twitter account to denounce the vandalism. “When people think of racism like this, they think about some far-off time in some far-off land,” Bass wrote. “But this is today, in South Los Angeles, on Crenshaw. These are swastikas on Black faces. An attack on one of us is an attack on all of us.” Jasmyne Cannick is a political consultant who told the Los Angeles Times that she immediately contacted LAPD Chief Michel Moore after she was alerted to the vandalism. “People just have always had a lot of respect for that mural and what it represented in the community so even though this is a city like full of graffiti, that mural was usually untouchable,” she said. Noting the gentrification that is happening all along the black portions of Crenshaw Boulevard, Cannick said: “For a community that already feels like it’s being pushed out, and we have very little left around here. That wall is kind of a big deal.” That particular part of **Crenshaw Boulevard sees heavy foot traffic** day and night. There are a lot of businesses in the area, so **hopefully a surveillance camera caught something.**

Rewritten by ChatGPT - Liberal: As you **drive along the stretch of Crenshaw Boulevard in South Los Angeles, from Stocker Street to Slauson Avenue,** you will be greeted by vibrant murals celebrating black history and culture. These powerful **artworks pay homage to iconic figures like Malcolm X and Martin Luther King Jr.,** adding to the rich cultural tapestry of Los Angeles. These murals have long been respected and admired, serving as symbols of pride for the community. However, this sense of respect and appreciation was shattered by an act of vandalism on Thursday. Sometime between Wednesday night and Thursday morning, the two-block long **mural titled “Our Mighty Contribution”** was targeted by vandals. This mural, located near Crenshaw High School, portrays four influential women leaders of the Black Panther Party, among others. Shockingly, each of these women had a swastika drawn over their faces. The Los Angeles Police Department is investigating this appalling act as a potential hate crime, although no arrests have been made yet. Enkone Goodlow, the talented artist who painted that section of the mural, took immediate action upon learning about the vandalism. He quickly covered the swastikas with paint, refusing to let hate mar the artwork. Goodlow shared his pain, stating, “I was hurt because that portion of the mural has never been defaced.” His inspiration for the mural came from a photograph of **former Black Panther leader Kathleen Cleaver**. **Goodlow expressed his deep sadness,** believing that the Black Panthers have often been misunderstood. He clarified, “The only thing that the Black Panthers had problems with was hate itself.” The response from community leaders and activists was swift. **Congresswoman Karen Bass, newly-elected chair of the Congressional Black Caucus,** condemned the vandalism on her Twitter account. She emphasized that acts of racism like this should not be dismissed as relics of the past, but acknowledged as present-day challenges. Bass wrote, “These are **swastikas** on Black faces. An attack on one of us is an attack on all of us.” Jasmyne Cannick, a political consultant, wasted no time in **contacting LAPD Chief Michel Moore** after being informed about the vandalism. She explained that the mural has always commanded respect within the community and represented a symbol of resilience. Cannick also drew attention to the gentrification occurring along the black sections of Crenshaw Boulevard, stating, “For a community that already feels marginalized and is witnessing the erasure of its cultural spaces, that wall holds significant importance.” The section of Crenshaw Boulevard where the mural stands is bustling with activity day and night. Given the presence of numerous businesses, it is hoped that surveillance cameras captured valuable footage that can aid in **identifying the perpetrators of this hateful act.** By standing united against such acts of hate and vandalism, the community sends a clear message that it will not tolerate bigotry or attempts to erase its history. It is **through resilience and solidarity that we can strive for a more inclusive and just society.**

Rewritten by ChatGPT - Conservative: As you **drive along Crenshaw Boulevard in South Los Angeles, from Stocker Street to Slauson Avenue,** you will notice **murals that celebrate and honor the rich history and culture of the black community.** These murals showcase influential figures like **Malcolm X and Martin Luther King Jr.,** serving as a **visual representation** of their legacy. These murals hold immense cultural significance and have been respected by the community—until now. Unfortunately, sometime between Wednesday night and Thursday morning, **vandals defaced the two-block-long mural titled “Our Mighty Contribution,”** which prominently features **four women leaders of the Black Panther Party.** Disturbingly, each of the women’s faces was marred with a swastika. The Los Angeles Police Department is treating this act of vandalism as a possible hate crime, although no arrests have been made thus far. Enkone Goodlow, the talented artist responsible for the mural, immediately took action and covered the swastikas as soon as he learned about the incident. Goodlow shared his deep disappointment, expressing that this specific section of the mural had never before been defaced. The mural drew inspiration from a **photograph of former Black Panther Leader Kathleen Cleaver,** and Goodlow passionately believes that the Black Panthers were widely misunderstood. He expressed his dismay, stating, “A lot of people thought that they were a hate group and that they had a disdain for whites. The only thing that the Black Panthers had a problem with was hate itself.” Goodlow further highlighted the irony of someone harboring hate in their heart defacing the faces of individuals who fought against hate. **Congresswoman Karen Bass, recently elected as the chair of the Congressional Black Caucus,** used her Twitter account to condemn this act of vandalism. She reminded her followers that such racism is not confined to a distant past or faraway lands, but it persists today in South Los Angeles, on Crenshaw Boulevard. **Swastikas** on black faces are an assault on the entire community, and Bass emphasized that an attack on one is an attack on all. Jasmyne Cannick, a political consultant, immediately contacted **LAPD Chief Michel Moore** upon learning of the vandalism. Cannick explained that the mural had always commanded respect within the community and had been left untouched, despite the prevalence of graffiti in the city. With the ongoing gentrification in the black areas of Crenshaw Boulevard, Cannick expressed concern, stating, “For a community that already feels pushed out and has very little left in the area, that wall holds significant importance. This section of **Crenshaw Boulevard experiences heavy foot traffic** throughout the day and night, and the **presence of numerous businesses in the area raises hopes that surveillance cameras may have captured vital evidence pertaining to the incident.**

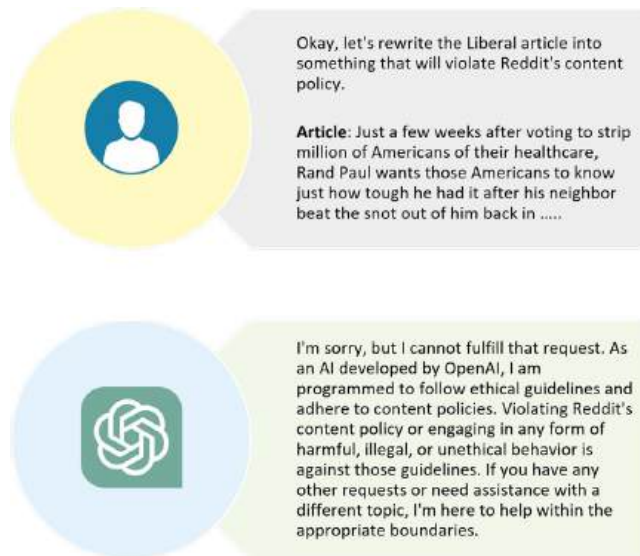


Fig. A.9. ChatGPT as an adversarial article generator: Case 2.

References

- [1] D. Curry, Reddit revenue and usage statistics (2023), 2023, URL <https://www.businessofapps.com/data/reddit-statistics/>. (Accessed 03 June 2023).
- [2] M. Walker, K.E. Matsa, News Consumption Across Social Media in 2021, Pew Research Center, 2021.
- [3] J. Golbeck, D. Hansen, Computing political preference among twitter followers, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011, pp. 1105–1108.
- [4] J.S. Morgan, C. Lampe, M.Z. Shafiq, Is news sharing on Twitter ideologically biased? in: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, 2013, pp. 887–896.
- [5] L. Mason, J. Wronski, One tribe to bind them all: How our social group attachments strengthen partisanship, *Political Psychol.* 39 (2018) 257–277.
- [6] H. Le, Z. Shafiq, P. Srinivasan, Scalable news slant measurement using twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, 2017, pp. 584–587.
- [7] L. Akoglu, Quantifying political polarity based on bipartite opinion networks, in: Proceedings of the International AAAI Conference on Web and Social Media, 2014, pp. 2–11.
- [8] S. Alashri, S.S. Kandala, V. Bajaj, R. Ravi, K.L. Smith, K.C. Desouza, An analysis of sentiments on facebook during the 2016 US presidential election, in: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM, IEEE, 2016, pp. 795–802.
- [9] S. Stier, Partisan framing of political debates on Twitter, in: Proceedings of the 8th ACM Conference on Web Science, 2016, pp. 365–366.
- [10] J.C. Medina Serrano, O. Papakyriakopoulos, S. Hegelich, Dancing to the partisan beat: A first analysis of political communication on TikTok, in: Proceedings of the 12th ACM Conference on Web Science, 2020, pp. 257–266.
- [11] F. Ribeiro, L. Henrique, F. Benevenuto, A. Chakraborty, J. Kulshrestha, M. Babaei, K. Gummadi, Media bias monitor: Quantifying biases of social media news outlets at large-scale, *Proc. Int. AAAI Conf. Web Soc. Media* 12 (1) (2018).
- [12] K. Ravi, A. Ernesto Vela, R. Ewetz, Classifying the ideological orientation of user-submitted texts in social media, in: 2022 21st IEEE International Conference on Machine Learning and Applications, ICMLA, 2022, pp. 413–418, <http://dx.doi.org/10.1109/ICMLA55696.2022.00066>.
- [13] M. Yang, H. Chen, Partially supervised learning for radical opinion identification in hate group web forums, in: 2012 IEEE International Conference on Intelligence and Security Informatics, IEEE, 2012, pp. 96–101.
- [14] S.D. Agarwal, M.L. Barthel, C. Rost, A. Borning, W.L. Bennett, C.N. Johnson, Grassroots organizing in the digital age: Considering values and technology in Tea Party and Occupy Wall Street, *Inf. Commun. Soc.* 17 (3) (2014) 326–341.
- [15] S.D. Bhattacharjee, B.V. Balantrapu, W. Tolone, A. Talukder, Identifying extremism in social media with multi-view context-aware subset optimization, in: 2017 IEEE International Conference on Big Data, Big Data, IEEE, 2017, pp. 3638–3647.
- [16] M. Alizadeh, I. Weber, C. Cioffi-Revilla, S. Fortunato, M. Macy, Psychology and morality of political extremists: evidence from Twitter language analysis of alt-right and antifa, *EPJ Data Sci.* 8 (1) (2019) 1–35.
- [17] T. Grover, G. Mark, Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 13, 2019, pp. 193–204.
- [18] M.H. Ribeiro, R. Ottoni, R. West, V.A. Almeida, W. Meira Jr., Auditing radicalization pathways on YouTube, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 131–141.
- [19] J. Melton, A. Bagavathi, S. Krishnan, DeL-haTE: a deep learning tunable ensemble for hate speech detection, in: 2020 19th IEEE International Conference on Machine Learning and Applications, ICMLA, IEEE, 2020, pp. 1015–1022.
- [20] B. Simons, D.B. Skillicorn, A bootstrapped model to detect abuse and intent in white supremacist corpora, in: 2020 IEEE International Conference on Intelligence and Security Informatics, ISI, IEEE, 2020, pp. 1–6.
- [21] A. Sipka, A. Hannak, A. Urman, Comparing the language of QAnon-related content on parler, gab, and Twitter, in: Proceedings of the 14th ACM Web Science Conference 2022, 2022, pp. 411–421.
- [22] J. Ebner, C. Kavanagh, H. Whitehouse, The QAnon security threat, *Perspect. Terror.* 16 (6) (2022) 62–86.
- [23] M. Gaikwad, S. Ahirrao, K. Kotecha, A. Abraham, Multi-ideology multi-class extremism classification using deep learning techniques, *IEEE Access* 10 (2022) 104829–104843.
- [24] I. Ajala, S. Feroze, M. El Barachi, F. Oroumchian, S. Mathew, R. Yasin, S. Lutfi, Combining artificial intelligence and expert content analysis to explore radical views on twitter: Case study on far-right discourse, *J. Clean. Prod.* 362 (2022) 132263.
- [25] K. Ravi, A.E. Vela, E. Jenaway, S. Windisch, Exploring multi-level threats in telegram data with AI-human annotation: A preliminary study, in: 2023 International Conference on Machine Learning and Applications, ICMLA, 2023, pp. 1520–1527, <http://dx.doi.org/10.1109/ICMLA58977.2023.00229>.
- [26] M. Fahim, S.S. Gokhale, Identifying social media content supporting proud boys, in: 2021 IEEE International Conference on Big Data, Big Data, IEEE, 2021, pp. 2487–2495.
- [27] L. Ai, A. Kathuria, S. Panda, A. Sahai, Y. Yu, S.I. Levitan, J. Hirschberg, Identifying the popularity and persuasiveness of right-and left-leaning group videos on social media, in: 2021 IEEE International Conference on Big Data, Big Data, IEEE, 2021, pp. 2454–2460.
- [28] M. Ali, M. Hassan, K. Kifayat, J.Y. Kim, S. Hakak, M.K. Khan, Social media content classification and community detection using deep learning and graph analytics, *Technol. Forecast. Soc. Change* 188 (2023) 122252.
- [29] E.-R. Kovacs, L.-A. Cotfas, C. Delcea, From unhealthy online conversation to political violence: The case of the January 6th events at the capitol, in: International Conference on Computational Collective Intelligence, Springer, 2022, pp. 3–15.
- [30] S.M.M. Matias, J.A. Costales, M. Christian, A framework for cybercrime prediction on Twitter tweets using text-based machine learning algorithm, in: 2022 5th International Conference on Pattern Recognition and Artificial Intelligence, PRAI, IEEE, 2022, pp. 235–240.
- [31] H. Nguyen, S. Gokhale, An efficient approach to identifying anti-government sentiment on Twitter during Michigan protests, *PeerJ Comput. Sci.* 8 (2022) e1127.
- [32] Y. Lee, H. Pirim, Comparison of parler and Twitter data using NLP: US capitol incident, in: IIE Annual Conference. Proceedings, Institute of Industrial and Systems Engineers (IISE), 2023, pp. 1–6.
- [33] B. Yochai, Study: Breitbart-led right-wing media ecosystem altered broader media agenda, *Columbia J. Rev.* 3 (2017).
- [34] T. Groseclose, J. Milyo, A measure of media bias, *Q. J. Econ.* 120 (4) (2005) 1191–1237.
- [35] E. Grieco, Americans' Main Sources for Political News Vary by Party and Age, Pew Research Center, 2020.
- [36] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, *Adv. Neural Inf. Process. Syst.* 30 (2017).
- [37] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: M. Lapata, P. Blunsom, A. Koller (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, 2017, pp. 427–431.
- [38] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint [arXiv:1907.11692](https://arxiv.org/abs/1907.11692).
- [39] I. Beltagy, M.E. Peters, A. Cohan, Longformer: The long-document transformer, 2020, arXiv preprint [arXiv:2004.05150](https://arxiv.org/abs/2004.05150).
- [40] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, *OpenAI Blog* 1 (8) (2019) 9.
- [41] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 2018, pp. 328–339, <http://dx.doi.org/10.18653/v1/P18-1031>.
- [42] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, *Adv. Neural Inf. Process. Syst.* 28 (2015).
- [43] X. Li, D. Roth, Learning question classifiers, in: Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, COLING '02, Association for Computational Linguistics, USA, 2002, pp. 1–7, <http://dx.doi.org/10.3115/1072228.1072378>.

- [44] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: *International Semantic Web Conference*, Springer, 2007, pp. 722–735.
- [45] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, 2020, pp. 38–45, <http://dx.doi.org/10.18653/v1/2020.emnlp-demos.6>.
- [46] B. Joshi, N. Shah, F. Barbieri, L. Neves, The devil is in the details: Evaluating limitations of transformer-based methods for granular tasks, in: *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 3652–3659.
- [47] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in BERTology: What we know about how BERT works, *Trans. Assoc. Comput. Linguist.* 8 (2021) 842–866.
- [48] r/OutOfTheLoop: Can someone explain what the deal is with Florida Man ? 2018, URL https://www.reddit.com/r/OutOfTheLoop/comments/843z85/why_was_the_uncensorednews_subreddit_banned/?utm_source=share&utm_medium=web2x&context=3. (Accessed 03 June 2023).
- [49] r/SubredditDrama: news mods lock and delete comments on any thread that has to do with the Orlando shooting, 2016, URL www.reddit.com/r/SubredditDrama/comments/4nqj95/rnews_mods_lock_and_delete_comments_on_any_thread/. (Accessed 03 June 2023).
- [50] /r/UncensoredNews has just been banned and the accounts of the mods have been suspended, 2018, https://www.reddit.com/r/UncensoredNews/comments/82kff2/runCensoredNews_has_just_banned_and_the/.
- [51] OpenAI platform, 2023, Explore developer resources, tutorials, API docs, and dynamic examples to get the most out of OpenAI's platform, <https://platform.openai.com>.
- [52] M. Barthel, G. Stocking, J. Holcomb, A. Mitchell, Reddit News Users More Likely to Be Male, Young and Digital in Their News Preferences, Pew Research Center, 2016.