2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA) | 978-1-6654-6283-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICMLA55696.2022.00066

Classifying the Ideological Orientation of User-Submitted Texts in Social Media

Kamalakkannan Ravi, Adan Ernesto Vela, Rickard Ewetz
University of Central Florida
Orlando, USA
rk@knights.ucf.edu, {adan.vela, rickard.ewetz}@ucf.edu

***Abstract*—With the long-term goal of understanding how language is used and evolves within online communities, this work explores the application of natural language processing techniques to classify text articles according to their ideological orientation (i.e., conservative or liberal). We first collect a balanced corpus of text articles posted to the online communities *r/Liberal* and *r/Conservative* from the social media website Reddit. Using the corpus, we develop and apply three classifiers. The baseline classifier is a Bayes model that accounts for each text article's web domain, as such, classification is independent of content. Next, we develop a support vector machine (SVM) model with term frequency-inverse document frequency (TF-IDF) features; this approach highlight differences in language using a count-based feature-space to differentiate text articles. Last, we evaluate the context-based transformer (RoBERTa) model and discuss its under-performance relative to the baseline and SVM models.**

***Index Terms*—Social networking (online), Text categorization, Predictive models, Bayes, Transformers, Support vector machines, Task analysis, Context modeling, Natural language processing**

I. Introduction

The rise in social media platforms and their widespread usage allows individuals to connect, communicate, and form communities to discuss common topics of interest. One such forum for engagement is Reddit - a social news aggregation and discussion site. Reddit ranks in the Top-10 most visited sites in the United States by Alexa. On the Reddit website, users can share and vote on social media content; further, users can join communities called *subreddits*, where they can engage in dialogue with others through comment sections.

In an effort to better understand how language is uniquely used within communities, we compare two related but contrasting groups as they tend to discuss the same topics from differing perspectives. In this work, we chose the liberal and conservative communities expressed through the *r/Liberal* and *r/Conservative* subreddits. These subreddits provide a view into the text articles, comments, agreements, and disagreements relevant to each community. In addition, as subreddits typically discuss current political affairs, they provide insight into how content-specific language evolves. As such, Reddit serves as an ideal platform for data collection and analysis of language in a semi-naturalistic setting, albeit Reddit users skew with regards to many key demographic features (esp. age, gender, education, and ideology [1]). Utilizing text articles (includes news, opinion pieces, blogs, and any other pieces of text) posted to each subreddit, we explore natural language processing (NLP) techniques to classify user-submitted content as coming from *r/Liberal* or *r/Conservative*. The long-term goal of developing such a classifier is to model the discourse of extremist ideologies (e.g., incel, anti-government, white nationalist) – here we begin with nominal ideologies.

The task of identifying ideologies in the text (whether they be nominal or extremist), especially when hidden or not readily apparent, bares resemblance to the prior analysis of ideological bias and sentiments of news (e.g., Fox News, New York Times), which has utilized crowd-sourcing and surveys [2, 3]. In these works, research typically focuses on the perceived bias of the media source itself, not necessarily the bias in the individual text article. Accordingly, word-count based syntactic and semantic or contextual analysis of the text articles [4] using natural language processing is rarely discussed in this branch of research. Even when such text data are collected, they are often labeled via survey, crowd-sourcing (Amazon Mechanical Turk), or third-party annotators like the Congressional Tweets Dataset [5].

Prior work like [6] explore the ideological leaning of text articles based on user votes and curated labels. For example, Zhou et al. [6] classify the ideological orientation of the news based on the assumption that right-wing users vote mostly for conservative text articles and similarly for left-wing. In addition, the news classifier also acts as a good recommender for social media users who are novices at recognizing bias and propaganda [7] in text articles compared to the expert news reading users. Further, the news classifier helps in understanding the ideological communities in cases where there is bias in liberal and conservative words [8] or gendered media coverage [9]. When it comes to social media text classifiers, the prior works have used TF-IDF based support vector machine (SVM) (news classification [10]), and Transformer models (sentiment classification [11]). In the work presented here, we exploit the TF-IDF/SVM and transformer models owning to the ability to capture structural and contextual information, respectively.

Our work takes an initial step in classifying the alignment of text articles associated with ideological communities at the word count-level and context levels. As such, our contributions are four-fold: **(1)** we explore a baseline classifier based on the news domain irrespective of the news content; **(2)** we demonstrate how the count-level and context-level approaches differ in identifying ideological communities; **(3)** we introduce

978-1-6654-6283-9/22/$31.00 ©2022 IEEE
DOI 10.1109/ICMLA55696.2022.00066

a novel dataset of ideologically-related text articles and provide a precise method for data collection; and **(4)** we provide a framework to compare and better understand language aligned with contrasting communities. In accomplishing **(2)** and **(3)**, we develop and apply TF-IDF based support vector machine (SVM) and context-based Transformer [12] models for language modeling.

II. Problem Statement

We consider the problem of labeling the ideological orientation or affiliation of text articles. Here, we define ideological orientation and affiliation as a combination of personal values affirmed by a greater community of persons who ascribe to that community. Aligned with this concept, we collected text articles that are submitted and discussed by community members of the *r/Liberal* and *r/Conservative* subreddits. As such, we assume that the text articles selected by the members of each subreddit reflect their beliefs and interests. Within this context, we seek to label articles using only their body text. The articles under consideration are not strictly restricted to news articles but also include opinion pieces, blogs, and any other pieces of relevant long-form text (i.e., not Twitter). Our problem of labeling text articles based on body text is in contrast to prior efforts that only considered publishing sources. As noted earlier, our more immediate goal in developing and applying natural language processing (NLP) techniques to classify the ideological orientation of text articles is to better understand the interplay between language usage within communities and NLP techniques used to analyze language. In the long term, we expect to expand our research and apply it to better understand radicalized and extremist ideologies.

In order to address the problem of labeling the ideological orientation of text articles, we begin by collecting 22,554 text articles from the *r/liberal* subreddit and 22,554 text articles from the *r/conservative* subreddit over a 13-year period. The text articles form the foundation of the training, development, and analysis datasets that are utilized in this research effort. With the datasets, we aim to address the following objectives: **(1)** investigate if domains of the text articles [3] can be used as a solid baseline to classify text articles irrespective of the body text; **(2)** develop a classifier that outperforms the domain-based baseline classifier; **(3)**; analyze whether the TF-IDF-weighted ngram SVM-based classifier or context-based RoBERTa classifier outperforms the baseline classifier; and **(4)** provide a general framework that can be applied to any *subreddits*; to gather text articles and analyze the language.

III. Dataset

To collect the text articles, we begin by gathering ids and Uniform Resource Locators (URLs) of all the submissions made to the *subreddits* from the first post date (5/2/09 for *r/Liberal* and 2/4/08 for *r/Conservative*) until 8/10/21 using the PRAW Reddit API and Pushshift Reddit API. Many URLs are from non-text sites (e.g., YouTube and Imgur); given that the focus of this research is on text articles, these records are removed since they do not fall within the scope of this study. The remaining URLs are scraped using the Beautiful Soup API; empty, duplicate, and inaccessible articles are excluded. As indicated in Table I, initially, over 53K and 6.12M articles were collected from the *r/Liberal* and *r/Conservative*) subreddits; after removing non-text sites, approximately 37K and 4.27M articles remained.

TABLE I: Articles counts.

(#Step) - Subreddit	Liberal	Conservative
(1) Number of URLs	53,357	6,12,660
(2) Filtered URLs	37,706	4,27,298
(3) URL w/ Articles	29,896	3,15,756
(4) Articles w/o duplicates	29,369	3,02,451
(5) Articles w/o Error texts	25,061	2,54,320
(6) Selected Articles	22,554	2,03,456

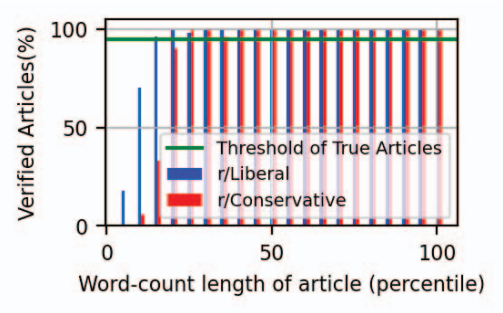

Fig. 1: Percent of annotated articles confirmed to be relevant (binned according to word-count percentile); 95% threshold for inclusion into the corpus.

Next, we remove scraped articles from the corpus that are unlikely to be text articles (e.g., 404 error pages, inaccessible paid sites). Instead of individually reviewing each text article, we established a word-count threshold to label each text articles are relevant or not. The assertion behind this approach is that the majority of error pages contain few words. To establish the word-count threshold, we divided all text articles into 20 bins based on the word-count percentile of the scraped text. We then annotate 100 randomly selected text articles in each bin as a relevant text article or not. We remove the bottom 10% (based on word count) of *r/Liberal* text articles and the bottom 20% of *r/Conservative* text articles as they have less than 95% verified text articles, as seen in Figure 1. To perform the annotation task, a graduate student annotated a total of 4000 text articles using the open-source annotation tool Doccano. After applying the word-count threshold and removing repeated text articles, 25K and 2.54M text articles remained in the *r/Liberal* and *r/Conservative*) subreddits.

The resulting corpus is class imbalanced for articles between 2008 and 2021. The disparity in the posting-rate of articles between the two subreddit is shown in Figure 2. In the run-up to the 2020 presidential election, the difference in the daily posting rate is quite significant, with ten-time more articles being posted to *r/Conservative* compared to *r/Liberal*. To sim-

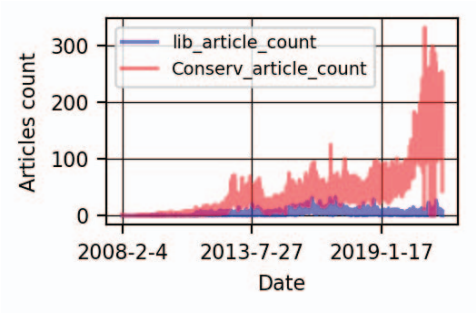

Fig. 2: Articles per day for each subreddit.

plify the development of the classifiers, we balance the corpus by keeping all 22554 articles from *r/Liberal* and sampling 22554 articles from *r/Conservative* at the same daily rate. The end result is a curated balanced corpus covering 13 years from 2009 until 2021. While community rules in both *subreddits* restrict off-topics submissions, there is the possibility of articles cross-posted in either *subreddits* for critiquing (e.g., a conservative article posted to *r/Liberal*). Consequently, such cross-posting adds some noise to the labels. Nevertheless, we label all the articles posted on the *r/Liberal* and *r/Conservative* forums as liberal and conservative, respectively.

IV. Initial Analysis

In this section, we seek to provide an initial analysis of the number of words, common sources, and famous words used in the collected text articles. The purpose of the analysis is to provide some insights into similarities and differences between the text articles posted to each subreddit, and more importantly, some of the noted similarities and differences are relevant when labeling articles using NLP techniques.

Analysis of the text article indicates that the number of words follow a long-tail distribution, and corresponding word-sentence statistics in Table II. We observe from Table II that the standard deviation is greater than the mean, which is reflective of a long-tailed skewed distribution that is not normally distributed. While prior works have suggested that the language and rhetoric of liberals tend to be more complex than that of conservatives [13], more recent analysis has suggested that the difference in language complexity might not be so clear. For example, [14] notes that while differences might emerge between *elites*, the language complexity of lay persons of differing political orientations is similar. As a point of interest, we performed a two-sample Kolmogorov-Smirnov hypothesis test ($\alpha = .01$, $N_{con} = N_{lib} = 22,554$) on the average word-count per sentence (a simple measure of language complexity). The result of the hypothesis test ($p - val = 2.3e^{-30} << \alpha$) indicates that in fact the *r/liberal* text articles tend to have more complex sentences then those text articles posted to *r/conservative*, which is to say $P(words/sentence < x|$ *r/liberal*$) < P(words/sentence < x|$ *r/conservative*$)$.

TABLE II: Word and sentence count.

	Word		Sentence	
	μ	σ	μ	σ
Liberal	1,649	8,872	74	450
Conservative	1,237	7,761	59	426

TABLE III: Top 10 most frequent text article domains.

Order	Liberal	Conservative
1	nytimes	breitbart
2	washington post	national review
3	mother jones	daily caller
4	the hill	the gateway pundit
5	politico	town hall
6	raw story	the hill
7	the guardian	hot air
8	cnn	washington times
9	vox	american thinker
10	salon	nypost

The Top-10 most frequent domains shown in Table III contribute more than 80% of the text articles. When comparing the Top-10 domains for each group, it is worth noting that they are rather distinctive from each other, the key exception being *The Hill*. Because the URL domains are mostly exclusive from one another, it lends credence to prior approaches that classify news publishers as *liberal* or *conservative*. Surprisingly, while text articles from CNN are frequently posted to *r/liberal*, the primary competitor, Fox News, is not included in the Top-10 for *r/conservative*.

Examining the text in the corpus, the most common words like *trump*, *people*, and *state* occurs in both liberal and conservative text articles. Given that the text articles are often referencing the same contemporary events, the similarity in the most commonly used words is not surprising.

V. Methodology

To classify the community alignment of the text articles with each *subreddit*, we select a simple Bayesian model and two language models, where the language models are consistent with word-count and contextual modeling, respectively.

Models. In the literature, we observe that news sources or publishers are often used as a baseline to determine the Ideological orientation of a text article due to their substantial accuracy [3]. Thus, to serve as a strong baseline, we apply the Bayes theorem to classify the text article solely based on the domain URL, as opposed to the contained body text. Accordingly, an article domain (D) aligns with *r/Liberal* (L) if $P(L \mid D) = P(D \mid L)P(L)/P(D) > .5$, whereby probabilities are extracted using empirical counts in the training dataset.

For word-count-based modeling, we apply the term frequency-inverse document frequency (TF-IDF) [15, 16] using unigrams and bigrams of the text articles and leverage support vector machines (SVM) to classify text articles as

liberal or conservative. While many of commonly used words overlap, it is expected that the combination of TF-IDF and SVM will identify terms closely associated with each subreddit community.

While the TF-IDM model language uses a word-count methodology, more recent language models have sought to provide a contextualized embedding, whereby words (even if they are the same) are understood within the context of other words or nearby sentences (e.g., the word *run* in the phrases, 'run for mayor' vs. 'run a footrace'). We employ context-based RoBERTa [17] - one of the robust transformer models [12], to encode the context of text articles and classify the ideological alignment.

Training and Hyperparameters. From the balanced dataset, we stratified a random sample 64% (liberal, conservative) pairs for training, 16% for development, and the rest 20% for testing. We perform 5-fold grid-search cross-validation for SVM models using the training and development set. The best parameters are *rbf* kernel, C = 10, and gamma = 1. For RoBERTa models, we initiate the model with RoBERTa-base pre-trained weights, and set the training parameters are training epochs = 10, batch size = 96, sequence length = 128, and leaning rate = $4 \cdot 10^{-5}$ after hyperparameter tuning.

SVM grid search 5 fold cross-validation took 47 hours on a Ubuntu workstation computer with 2 x Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.6GHz, and 128 GB RAM. The hyperparameter selection criteria was accuracy.The grid search parameters were 'kernel' : ['linear', 'rbf', 'poly'], 'C' : [0.1, 0.5, 1, 10, 100], and 'gamma' : [0.001, 0.01, 0.1, 1, 10]. Then the best parameters were used to train the final SVM model, which ran for 5 hours and testing took 15 minutes to complete. Furthermore, We used AWS instance with 8 vCPU, 32 GB RAM, and 16 GB NVIDIA T4 GPU for the RoBERTa-transformer model. We initialized the model with RoBERTa-base pre-trained weights and fine-tuned it on our train and development set (training saturates within 10 Epochs), running for 45 minutes. The training criteria was evaluation loss, and testing took 45 seconds.

VI. Results and Discussion

Following the development and training of the difference classifiers, they are applied to the evaluation data set to assess their performance. Table IV and V describes the classification results: accuracy, precision, recall, and f1-score. We also provide classification results based on the source domain as a strong baseline since it aligns with prior analysis of news media sources (see Section I). We observe that classifying the most frequent domains (Table III) using Bayes provides reasonable accuracy as the Top-10 domains contribute more than 80% of text articles and are mostly unique to their *subreddit*, with *The Hill* being an exception. However, domain-based classification is not scalable when testing on text articles coming from domains, not in the training dataset, as there are 305 news domains not present in the original 36086 domains in the training set. Thus, classifying based on an article's body text is preferred as it overcomes this weakness; the SVM and Transformer models are able to classify text articles originating from previously unobserved sources.

TABLE IV: Classification results on 9022 text articles. *For Bayes, 36086 text article domains were used to determine the class of each domain and tested on 8717 news domains as the remaining 305 news domains were not present in the train set.

Model	Acc(%)	TN	FP	FN	TP
Bayes*	81.54	3,331	1,016	593	3,777
SVM	**86.19**	4,032	479	767	3,744
RoBERTa	78.13	3,785	726	1,247	3,264

TABLE V: Classification report on each class. All the metrics are in percentage.

Model	Class	Precision	Recall	f1-score
Bayes*	r/Lib	**84.89**	76.63	80.55
	r/Con	78.80	**86.43**	82.44
SVM	r/Lib	84.02	**89.38**	**86.62**
	r/Con	**88.66**	83.00	**85.73**
RoBERTa	r/Lib	75.22	83.91	79.33
	r/Con	81.80	72.36	76.79

From our data analysis, we observe considerable overlap between liberal and conservative text articles in terms of the words and their counts. However, as demonstrated by the TF-IDF/SVM model, higher-order count-based syntactic information is contained within the text of the text articles that allow for improved classification into *r/Liberal* and *r/Conservative*. Based on overall accuracy, when using this word-count based approach, the TF-IDF/SVM outperformed classifying text articles based on the URL domain (86.19% vs. 81.54%).

In comparison, the context-based transformer models suffer from relatively poor performance (lower than the baseline). The relative degradation in classification performance is consistent with prior observations that noted that syntactic meaning is not encoded directly in attention weights of transformer models [18, 19]. Such a result indicates that while it is likely overlapping encoded language used across communities, count-based syntactic information is valuable in distinguishing communities through text analysis.

The distinction in extracting word-count based and context-based features for classification is highlighted by the text articles from *The Hill*, a common news domain in the Top-5 most frequent sources (Table III) for both *r/Liberal* and *conservative*. As shown in Table VI and VII, when compared to the TF-IDF based SVM model, the transformer model labels more conservative text articles as liberal. Thus at the contextual level, the results imply that discourse conveyed by the text articles is often similar across *subreddits*. Given this challenge, we plan to study the effect of context-encoded features obtained using transformer models and classify them using SVM. Moreover, expand the news classification to the original imbalanced dataset to understand the real-world ideological discourse over the time-dimension (13 years). Finally,

TABLE VI: Confusion matrix of "The Hill" text articles classification using SVM.

		Predicted	
		Liberal	Conservative
Actual	Liberal	113	21
	Conservative	10	134

TABLE VII: Confusion matrix of "The Hill" text classification using RoBERTa.

		Predicted	
		Liberal	Conservative
Actual	Liberal	116	18
	Conservative	43	101

we plan to address this limitation and model explainability to analyze the ideological discourse on social media.

To better understand the role of keywords when labeling the ideological orientation of text articles, we collected the ngrams which are observed frequently (>10000 times) ngrams and which disproportionately appear in one class versus the other. Table VIII provides a list of ngram keywords from the Liberal and Conservative training set. Interestingly, a couple of the ngrams find correspondence across the two groups. That is to say, in *r/Liberal*, the term 'far right' is commonly used; meanwhile, its corresponding term in *r/Conservative* might be 'conservative'. The other word pair is 'progressive' and 'liberals'. To test the importance of key words, we select 18 Liberal text articles with keywords *[(far right or progressive)* and 195 Conservative text articles with these key words *(liberals or conservative)*. We then replace the keywords in the Liberal text articles with their Conservative equivalent (progressive $\leftrightarrow$ liberal) and (far right $\leftrightarrow$ conservative), and vice versa. To check if the classes change due to keywords replacement, we predict their classes using TF-IDF/SVM and fine-tuned RoBERTa models. Using the TF-IDF/SVM model, 16.66 % of Liberal text articles become Conservative after keywords replacement compared to 11.11 % using RoBERTa. However, when using the TF-IDF/SVM model, 2.51 % of Conservative text articles become Liberal after keywords replacement compared to 3.96 % using RoBERTa. Overall, the TF-IDF/SVM approach is sensitive to words and performs relatively better than the transformers model. Thus, proving prior works [18, 19] that the TF-IDF/SVM method encodes syntactic meaning better than the RoBERTa transformer model and syntactic information is advantageous in contrasting online communities. While not explored here, one other key difference observed in Table VIII is that in many cases, the text articles in *r/Liberal* and *r/Conservative* are discussing different topics. That is to say, in *r/Liberal* text articles are more likely to discuss health care and social programs while *r/Conservative* disproportional discuss gun rights, free speech, and science related issues.

TABLE VIII: Frequent ngrams (>10000)

r/Liberal: president obama, special counsel, mrs clinton, middle class, affordable care, mr trump, voter fraud, social security, far right, single payer, wage, income, families, progressive, president abortion
r/Conservative: big tech, global warming, second amendment, free speech, free speech, ted cruz, president trump, culture, liberals, science, conservative

VII. Conclusion

In this paper, we describe the ongoing research on the online ideological community classification in Reddit. First, we outline the gathering of text articles, annotation, and quality selection. With that, we discover the word statistics, overlap, domain share, and dataset distribution. Next, we showcase the effectiveness of a news domain-based Bayes classifier and the need for text-based classification. Later we outperform the baseline by leveraging simple yet effective SVM with TF-IDF features (Accuracy: 86.19%). Also, we compare the word count-based SVM with TF-IDF and complex context-based yet generalizable RoBERTa - transformer model (Accuracy: 78.13%) in classifying the ideological discourse and discuss its shortcoming. Besides, we discuss the limitation in the current form and ways to mitigate the challenges. Further, to encourage reproducibility, online community research, and language modeling, we will open source our code, trained models, and data.

Lastly, while beyond the achievements of this paper, in the social sciences, texts are often analyzed in the context of pragmatics and discourse analysis – essentially related to the use of language and, more broadly, how texts interact within a greater social context [20]. For example, the discourse-analytical and pragmatic approach helps identify the racist, anti-Semitic meanings [21] and manipulative intent [22] in the news media. For this work, our long-term goal is to go beyond the words used in texts but to identify underlying narratives in the coded or suggestive language used by non-normative communities associated with targeted violence.

Ethical Considerations

We recognize that classification of news 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 [1]). As such, it is essential to recognize that our classifications of *liberal* or *conversative* 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.

ACKNOWLEDGEMENTS

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.

REFERENCES

[1] Barthel, M.. (2016). Reddit news users more likely to be male, young and digital in their news preferences.

[2] Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? Quantify- ing media bias through crowdsourced content analysis. Public Opinion Quarterly 80, S1 (2016), 250–271.

[3] Elizabeth. 2020. Americans' main sources for political news vary by party and age. Pew Research Center 1, 1 (Apr 2020), 1. https://www.pewresearch.org/fact-tank/2020/04/01/americans-main-sources-for-political-news-vary-by-party- and-age/

[4] Andreas H Jucker. 2017. Chapter 9: Pragmatics and Discourse. In: Brinton, Laurel J; Bergs, Alexaner. The History of English/Historical Outlines from Sound to Text. Walter de Gruyter GmbH & Co KG, Online. 165–184 pages.

[5] Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). ACL, Online, 720–730.

[6] Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. 2011. Classifying the political leaning of news articles and users from user votes. In Proceedings of the international AAAI conference on web and social media, Vol. 5. AAAI, Online, 417–424.

[7] Oana Balalau and Roxana Horincar. 2021. From the Stage to the Audience: Propaganda on Reddit. In Proceedings of the 16th Conference of the Euro- pean Chapter of the Association for Computational Linguistics: Main Vol- ume. Association for Computational Linguistics, Online, 3540–3550.

[8] James Fairbanks, Natalie Fitch, Nathan Knauf, and Erica Briscoe. 2018. Credibility assessment in the news: do we need to read. In Proc. of the MIS2 Workshop held in conjuction with 11th Int'l Conf. on Web Search and Data Mining. ACM, Online, 799–800.

[9] Eran Shor, Arnout Van de Rijt, Charles Ward, Saoussan Askar, and Steven Skiena. 2014. Is there a political bias? A computational analysis of female subjects' coverage in liberal and conservative newspapers. Social Science Quarterly 95, 5 (2014), 1213–1229.

[10] S. M. H. Dadgar, M. S. Araghi and M. M. Farahani, "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification," 2016 IEEE International Conference on Engineering and Technology (ICETECH), 2016, pp. 112-116.

[11] M. Heidari and J. H. Jones, "Using BERT to Extract Topic-Independent Sentiment Features for Social Media Bot Detection," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2020, pp. 0542-0547.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems. NeurIPS, Online, 5998–6008.

[13] John T Jost, Jack Glaser, Frank J Sulloway, and Arie W Kruglanski. 2018. Political conservatism as motivated social cognition. Routledge.

[14] Shannon C Houck and Lucian Gideon Conway III. 2019. Strategic communication and the integrative complexity-ideology relationship: Meta-analytic findings reveal differences between public politicians and private citizens in their use of simple rhetoric. Political Psychology 40, 5 (2019), 1119–1141.

[15] Hans Peter Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development 1, 4 (1957), 309–317.

[16] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation 1, 4 (1972), 1.

[17] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In Proceedings of the 20th Chinese National Conference on Computational Linguistics. Chinese Information Processing Society of China, Huhhot, China, 1218–1227.

[18] Brihi Joshi, Neil Shah, Francesco Barbieri, and Leonardo Neves. 2020. 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, Barcelona, Spain (Online), 3652–3659.

[19] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertol- ogy: What we know about how bert works. Transactions of the Association for Computational Linguistics 8 (2020), 842–866.

[20] Andreas H Jucker. 2017. Chapter 9: Pragmatics and Discourse. In: Brinton, Laurel J; Bergs, Alexaner. The History of English/Historical Outlines from Sound to Textt. Walter de Gruyter GmbH & Co KG, Online. 165–184 pages.

[21] Ruth Wodak. 2007. Pragmatics and critical discourse analysis: A cross-disciplinary inquiry. Pragmatics & cognition 15, 1 (2007), 203–225.

[22] Peter Furko. 2017. Manipulative uses of pragmatic markers in political discourse. Palgrave Communications 3, 1 (2017), 1–8.