Measuring ideological spectrum through NLP

Franco Demarco[1,†], Juan Manuel Ortiz de Zarate[1,†] and Esteban Feuerstein[1]

[1] *Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Computación. Buenos Aires, Argentina*

Abstract

In the evolving landscape of online communities, the dispute between social integration and fragmentation has sparked ongoing debates. With the advent of technologically mediated social networks, understanding the structure of these communities remains a challenge. This study introduces a fresh, text-based technique to quantify the alignment of online communities along social dimensions. Through the analysis of historical Reddit data, community representations are generated from Reddit posts and projected onto ideological-partisan axes. This approach successfully scores communities, effectively situating them on the political-ideological spectrum.

Our approach rests on the premise that the language, topics, parlance, and discourse style employed by communities offer insights into their ideological leanings. We found that using posts' text we can build a very similar and correlated partisan-ness ranking to the one inferred through user interactions, which reinforces our premise. This text-based approach also enables the analysis of books, news, blogs, and other sources that were not possible with previous approaches. Our results underscore the advantages of transformer-based embeddings when compared to skip-gram embeddings trained on the same dataset. This work contributes to the understanding of online community structures and their ideological foundations.

Keywords

NLP, Social Networks, LLM, Communities

1. Introduction

For decades, before the rise of technologically mediated social networks, a heated debate has raged over the interplay of two competing dual forces on the Internet: one of social integration, as the world has become increasingly interconnected, and another of social fragmentation, since people may tend to join like-minded communities [2, 3, 4]. Today, 20 years after the mass adoption of online social networks and platforms, it remains unclear how online communities are socially organized. Of particular concern is whether online populations are increasingly classified into homogeneous *echo chambers* and whether social media platforms tend to push users toward ideological extremes [5, 6]. However, since these platforms consist of massive amounts of unstructured and anonymous data, empirically quantifying the social composition of online communities and, in turn, the social organization of online platforms poses an enormous challenge.

NL4AI 2023: Seventh Workshop on Natural Language for Artificial Intelligence, November 6-7th, 2023, Rome, Italy [1]

[†] These authors contributed equally.

fddemarco@dc.uba.ar (F. Demarco); jmoz@dc.uba.ar (J. M. O. d. Zarate); efeuerst@dc.uba.ar (E. Feuerstein)
0009-0002-1399-9469 (F. Demarco); 0000-0002-0291-1997 (J. M. O. d. Zarate); 0000-0003-2985-810X (E. Feuerstein)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

In this work, we propose a technique to quantify the position in ideological spaces based on the text posted by each community. This technique is based on the hypothesis that the jargon, topics, parlance, and discursive forms used by each community provide valuable insights into their ideological aspects, especially the political ones, similar to how the interactions of users within each community do. While our approach to quantifying partisan tendencies aligns with certain aspects of prior research [7], user interaction-based methods face a fundamental constraint: they are applicable solely to data collected within a single platform. Our text-based approach broadens its scope by facilitating the incorporation of diverse data sources, encompassing various social platforms like Facebook and Twitter, as well as newspapers, blogs, user-generated content, and others.

We utilize the same dataset as [7], which offers a substantial amount of text and serves as a valuable baseline for our work. First, we collect the text of the posts and group them by community and year, spanning from 2012 to 2018. Next, we apply various embedding techniques to estimate community embeddings, including models based on the skip-gram model [8] and more complex ones based on transformers [9]. Finally, we calculate the social dimensions using the methodology proposed by Waller et al.[7]. This process involves the analyst selecting two communities as seeds, determining the direction between these two seed vectors, and subsequently projecting the remaining community embeddings onto these dimensions. To enhance robustness, seed augmentation is employed (for additional details, refer to Section 3)[1].

As demonstrated in Section 4.2, our obtained results exhibited a high degree of similarity to those acquired by Waller et al. [7] meaning that interactions between users and communities have a correlation with language. Moreover, it lets us create a new kind of dimension based on text instead of seed communities and using any set of texts instead of post communities. Furthermore, a prominent observation was the consistent out-performance of transformer-based embeddings in contrast to skip-gram embeddings. This advantage remained evident even when we trained our skip-gram model on the particular dataset and made use of pre-trained vectors.

This paper is organized as follows: in Section 2, we list and summarize previous work on analyzing social networks with innovative techniques. Section 3 contains the step-by-step description of our pipeline, along with the introduction of two new natural variations on the RBO similarity measure. In Section 4 we describe the datasets collected for this study, and we present the obtained results. Finally, we conclude with Section 5.

2. Related work

The research conducted by Waller et al. [7] introduces a novel technique for quantifying the positioning of online communities along social dimensions, relying on users' interactions. By leveraging the complete historical records of Reddit posts and comments from 2012 to 2018, the researchers generate community representations from these interactions. They then project these representations onto one-dimensional axes that symbolize a *social dimension*. This process yields scores for each community, effectively situating them on the corresponding dimension spectrum. This methodology produces results that coherently align with qualitative perceptions.

[1] All code and data are available at https://github.com/fddemarco/BIICC-2023

On the other hand, many recent works have shown a significant correlation between jargon and community discussions. Ramponi et al. [10, 11] build very efficient classifiers and predictors of account membership within a given community by inspecting the vocabulary used in tweets for many heterogeneous Twitter communities such as chess players, fashion designers, and supporters of political parties. In [12] Tran et al. found that the language style, characterized using a hybrid word and part-of-speech tag *n-gram* language model, is a better indicator of community identity than the topic, even for communities organized around specific topics. Lahoti et al. [13] model the problem of learning the liberal-conservative ideology space of social media users and media sources as a constrained non-negative matrix-factorization problem. They validate their model and solution in a real-world Twitter dataset. On polarized contexts, De Zarate et al. [14, 15] show that they can measure the level of controversy in a discussion through the texts posted by communities.

Finally, the article titled 'We Don't Speak the Same Language: Interpreting Polarization through Machine Translation'[16] examines the growing polarization observed among political parties, media outlets, and elites in the U.S., with a particular emphasis on social media. The study focuses on how different communities perceive and use language in distinct ways, suggesting that these communities are essentially *speaking different languages*. To address this phenomenon, the authors introduce a novel method that employs machine translation as an analytical tool. The central idea is that when two communities use language significantly differently, machine translation techniques can identify and translate these differences, offering unique insights into language polarization. This work underscores the crucial role of language in polarization and provides an innovative tool for analyzing and understanding this phenomenon at a more granular level. By utilizing machine translation, traditionally employed for converting one language to another, the study delves into the intrinsic language distinctions between polarized communities, offering a fresh perspective on how language reflects and amplifies social and political divisions.

3. Methodology

Our methodological contribution is the introduction of a novel approach to quantify social organization through textual data. Our hypothesis is that the jargon, topics, parlance, and discursive forms employed by each community offer valuable insights into their ideological aspects, particularly the political ones, much like their interactions.

Initially, we outline the general algorithm for constructing social dimensions. Subsequently, we detail the specific choices we made during our analyses. Finally, we elaborate on the computation of community scores and their validation against the prior findings presented in [7].

3.1. Generating the word embeddings

We used the datasets presented in Section 4.1 to represent Reddit's communities, known as *subreddits*, in a jargon space. To ensure meaningful vector representations, we removed extremely small subreddits with insufficient posts. Therefore, our analysis is limited to the top 10.000 subreddits, ranked by the number of submissions.

To generate word embeddings for each community in the jargon space, we selected two models among the most advanced ones, namely *FastText* [8] and *Cohere*'s transformer-based model [9]. These models embed texts into fixed dimension vectors encoding semantically significance and meaning.

Both language models presented in the following paragraphs take a single text corpus as input and return a single vector representation. Thus, to generate an embedding characterizing each community, we create a unified text corpus by concatenating all textual content from submissions within the corresponding subreddit, including both titles and self-posts.

FastText This tool is an extension of the skip-gram model [8]. In this approach, words are represented as collections of character *n-grams*. Each character *n-gram* is associated with a vector representation, and words are represented as the sum of these individual vectors. This methodology offers a fast training mechanism suitable for large corpora and can generate word representations for terms not present in the training dataset. It also achieves state-of-the-art performance on word similarity and analogy tasks, surpassing previous results obtained with skip-gram-based tools like `word2vec` [17].

The FastText model has several hyperparameters that impact the training process and the resulting embeddings. These hyperparameters include the learning rate, the size of word vectors, the size of the context window, the number of epochs, and others. We decided to use the default values for all of these parameters, except for the size of word vectors and the number of epochs. Specifically, we set the size of word vectors to 300 dimensions, matching the vector size of the pretrained wiki-en vectors[2]. Additionally, when using the *Full dataset* for training, we chose to set the number of epochs to 1 due to a limitation on our infrastructure. For additional information about the input datasets and the hyperparameters used, please refer to Section 4.2.

Cohere The Cohere Platform[3] offers an API for integrating cutting-edge language processing to any system. Cohere trains massive language models and makes them accessible through a user-friendly API. The platform provides a range of models that cover various use cases, including representation models which can generate text embeddings.

Among the representation models offered by the Cohere Platform, we chose to utilize the *embed-english-v2.0* transformer-based model [9]. The Transformer is a simpler and more efficient network architecture for sequence transduction models, compared to previous ensembles, complex recurrent networks, or convolutional neural networks. It replaces the recurrent layers commonly found in encoder-decoder architectures with multi-headed self-attention. Specifically, this model generates embeddings by computing the average of contextualized embeddings for each token within the text, a technique aligned with the work of Reimers and Gurevych [18][4].

It's worth noting that this specific model is limited by its dependence on the English language and lacks reliable functionality for languages other than English. Since our target communities primarily consist of English speakers, we consider this limitation to be inconsequential for our work.

[2] https://fasttext.cc/docs/en/pretrained-vectors.html

[3] https://docs.cohere.com/docs

[4] https://docs.cohere.com/docs/embeddings

Another constraint imposed by this model is the 512-tokens limitation per text, with each token typically corresponding to 2-3 characters[5]. To address this token limitation, we propose reducing the amount of data supplied to the embedding generator. By selecting highly relevant posts, we can obtain a more compact dataset that serves as a sufficiently representative sample for each community (see Section 4.1 for more details). However, we acknowledge that using only 512 tokens (approximately 200 words) may not provide a fully representative sample. Therefore, reducing the input data alone is not a complete solution and should be revisited in future work. For our vision on how to fully address this limitation, please refer to Section 5. Since we are utilizing a pre-trained model without conducting fine-tuning, the use of this model does not involve specifying any hyperparameters.

3.2. Community scores

For each year, we generated embeddings exclusively from the submissions within that particular year. Following this, we calculated scores for all 10.000 communities utilizing the projection technique outlined in [7]. To execute this technique, the analyst initially identifies a seed pair of communities that exclusively vary in terms of the target construct. In our study, we employed *r/democrats* and *r/Conservative* in accordance with [7]. Subsequently, we expand the initial seed pair to encompass up to 10 pairs, and the resulting vector differences are averaged to derive a single vector. This yields a vector that represents the target construct **d**. All communities can be assigned a score by projecting the *normalized* community vector **c** onto the dimension vector **d**, that is, by calculating the *cosine similarity*. A visual explanation of this process is in Figure 1.

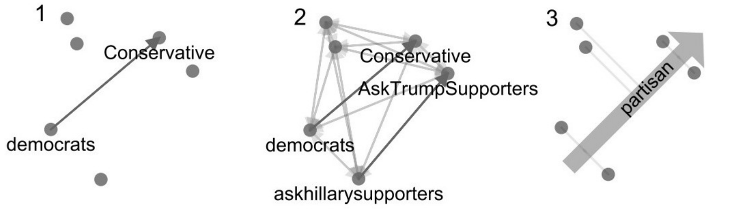

Figure 1: Fig 1b in [7]: Illustration of the methodology to generate the partisanship dimension.

3.3. Evaluating ranking

To evaluate the model's performance, we conducted a comparative analysis in relation to the findings presented in [7]. Our assessment was specifically confined to the political dimension outcomes highlighted in their study. Given the absence of an absolute reference, we concentrated on the communities identified as being most closely linked to the ideological extremes depicted in Fig. 1d top in [7] (refer to Table 1).

Scores inherently produce a ranking, which we can then compare using similarity measures. We chose to conduct an objective-observed comparison between our results (*observed*) and Waller's (*objective*, *ground-truth*, or *gold standard*). This means that we interpret differences

[5] https://docs.cohere.com/docs/tokens

Table 1
Fig. 1d top in [7]: Partisan-ness of communities identified as being most closely linked to the ideological extremes (gold standard ranking).

Right-wing Communities	partisan-ness
Conservative	0.44
The_Donald	0.34
TrueChristian	0.31
NoFapChristians	0.29
Mr_Trump	0.29
metacanada	0.29
conservatives	0.27
The_Farage	0.27
new_right	0.27
Christians	0.26

Left-wing Communities	partisan-ness
democrats	-0.35
EnoughLibertarianSpam	-0.32
hillaryclinton	-0.30
progressive	-0.30
BlueMidterm2018	-0.30
EnoughHillHate	-0.29
Enough_Sanders_Spam	-0.29
badwomensanatomy	-0.29
racism	-0.29
GunsAreCool	-0.29

from Waller's as indicative of a decrease in quality. To facilitate the comparison between our rankings and Waller's, we employed the following well-established similarity measures.

Kendall's τ [19] Kendall's τ correlation measure quantifies the compatibility between two provided rankings. Its values range from -1 to 1, with those close to 1 signify strong agreement and those close to -1 indicate strong disagreement. Specifically, a value of 1 signifies identical order, while a value of -1 indicates reverse order. A value of 0 signifies uncorrelated or a *random* relationship.

Let C be the number of concordant pairs, where both rankings share the same order for two items, and let D represent the number of discordant pairs. Then

$$\tau = 2\frac{C - D}{n(n-1)}.$$

Kendall's τ has a natural probabilistic interpretation. Choose a pair of distinct items at random. Let p_c be the probability that the pair is concordant, and p_d the probability that the pair is discordant. Then, we can prove that $\tau = p_c - p_d$. Therefore, $\tau = 0$ indicates that it is equally probable to sample a discordant pair as it is to sample a concordant pair at random.

Notably, it is an unweighted measure, assigning equal weight to disorder at the bottom of the ranking as it does to disorder at the top. For this reason, we can employ this measure to gain insight into the overall similarity of both rankings.

Rank-biased overlap [20] RBO is a *top-weighted overlap-based* measure. The central idea behind RBO is to use a convergent series of weights to adjust the proportional overlap at each depth. The Rank-biased overlap between two infinite rankings, denoted as S and T, is defined as follows:

$$\text{RBO}(S, T, p) = (1-p)\sum_{d=1}^{\infty} p^{d-1} \cdot A_d,$$

where A_d is the *agreement* at depth d, that is, the overlapped proportion of $s_1 \dots s_d$ and $t_1 \dots t_d$. The parameter p is a value that falls in the range [0,1] and it influences the rate of weight decline: a smaller p results in a more pronounced top-weighted characteristic for the measure.

Due to RBO's convergence property, evaluating a prefix establishes both a minimum and a maximum for the full score. By calculating the preceding equation up to a specific depth k, referred to as *RBO@K*, we establish a lower bound on the full evaluation. It is also possible to prove that the prefix evaluation provides a precise upper bound on the full score. Hence, it is possible to assess similarity using RBO even on infinite lists by utilizing both bounds.

Rank-biased overlap offers an interpretation as a probabilistic user model. Consider a user comparing two rankings. Let's assume that the user consistently examines one item in each ranking at a time. As we progress through the rankings, at each level, there is a probability of p to continue to the next position; therefore, there is a complementary probability of $1 - p$ to decide to stop. Let D represent a random variable denoting the depth at which the user eventually decides to stop, and let $P(D = d) = (1 - p)p^{d-1}$, denote the probability of the user stopping at a specific depth d. Once the user has stopped, we calculate the *agreement* A_d between the two lists at that depth d.

Note that the variable D follows the *Geometric distribution* with $\mathbf{p} = 1 - p$. Then, it follows that the expected value of the random variable D is given by: $\mathbb{E}(D) = \frac{1}{1-p}$. Within this framework, the expected value of this random experiment is as follows:

$$\mathbb{E}(A_D) = \sum_{d=1}^{\infty} P(D = d) \cdot A_d = RBO(S, T, p).$$

The RBO measure falls within the range of [0, 1], where 0 indicates disjointness (strong disagreement) and 1 indicates identity (strong agreement).

Given that we are dealing with finite rankings, we chose to employ RBO@k [6]. Additionally, we chose a value for parameter p such that sets the expected number of results compared by the *p-persistent* user to 3. In other words, $\mathbb{E}(D) = \frac{1}{1-p} = 3$, that is, $p = 2/3$. This is equivalent to assigning 87% of the weight to the first three results in the similarity comparison, as described in Equation 21 of [20].

RBO variations. Up to this point, we have introduced two similarity measures: Kendall's τ and RBO. These two measures help us identify differences between rankings, but they differ in how they emphasize the positions where discordance occurs. Kendall's τ is an unweighted measure, assigning equal importance to all positions in the ranking. In contrast, RBO is a top-weighted measure, meaning that it places greater emphasis on concordance at the top of the ranking. This emphasis aligns with the context of information retrieval, where users typically prioritize the quality of the first few items in a web search and are less concerned with items toward the bottom.

While both Kendall's τ and RBO are valuable, they do not particularly emphasize the lower end of the ranking. To address this concern, we have introduced two natural variations of the RBO measure, known as *2WRBO* and *H&HRBO*. These adaptations effectively allocate weight to both ends of the ranking, resulting in two *extreme-weighted* measures.

[6] We utilize the implementation found at https://github.com/changyaochen/rbo

The **2WRBO** of two rankings, A and B, is the average of their regular RBO scores and the scores of their reverses:

$$\text{2WRBO}(A, B) := \frac{RBO(A, B) + RBO(A^{-1}, B^{-1})}{2},$$

where A^{-1} is the reverse of A.

The **H&HRBO** of two rankings is defined in a slightly different manner. In the context of a double-ended ranking, as in our case study, we can treat it as two separate rankings. The first half ranks the most relevant items in a specific order, while the reverse of the second half ranks the most relevant in the complete opposite order. This interpretation of a double-ended ranking leads to the definition of H&HRBO:

$$\text{H\&HRBO}(A, B) := \frac{RBO(A_{:n/2}, B_{:n/2}) + RBO(A^{-1}_{:n/2}, B^{-1}_{:n/2})}{2}$$

The key distinction between these measures is that H&HRBO completely ignores an item if it is ranked beyond its corresponding half, capitalizing on the disjoint nature of the RBO measure. Also, as they are averages of RBO measures they are constrained in the segment [0,1], where 1 means perfect match and 0 that they are completely different.

4. Experiments

In this section, we report the results obtained by running the above-proposed method over different Reddit communities.

4.1. Data

For our analysis, we have prepared two distinct datasets to gain meaningful insights into the political-ideological organization of online communities. The first dataset, the *Full dataset*, encompasses a broad range of historical data. Additionally, we have generated a second dataset, the *Small dataset*, which is a reduced version of the first dataset comprising the most relevant posts from each community. In the following paragraphs, we will provide a more detailed presentation of both datasets and the preprocessing steps we undertook to ensure the reliability and consistency of our analyses, as well as other sources of data used.

Full dataset. This dataset is a subset of Reddit submissions spanning from 2012 to 2018. Our specific focus was on *submissions* that contained text, either in the form of a title or a *self-post* (also known as *text post*, *self-text*). To prepare the data, we applied text normalization, which encompasses removing user names, links, punctuation, tabs, leading and lagging blanks, general spaces, and mark-up language. Notably, the combined submissions from 2016 and 2018 account for 63.5% of the total.

Small dataset. This dataset is a subset of the *Full dataset*, comprising the most relevant posts from each community. We decided to use up-votes as a measure of relevance, but other measures, such as the number of comments and down-votes, are also possible. The rationale behind this dataset is that the top-relevant posts contain significant information, enabling us to distinguish the communities from one another. By adopting this approach, we can reduce the data required for training our models and generating embeddings while still achieving comparable results. Furthermore, this allows us to represent each community with an equal amount of words, characters, or tokens.

Data sources and Ethics. The publicly available dataset was downloaded from the *pushshift.io Reddit archive* [21]. It is important to note that all Reddit submissions are public, and users consent to make their data freely available by posting on Reddit, as noted in the Reddit privacy policy[7].

Other sources of data. We utilized wiki-en word vectors to enhance the performance of our FastText-based models. The FastText team has released pre-trained word vectors for 294 languages, which were trained on the Wikipedia. These 300-dimensional vectors were generated using the skip-gram model as described in [8] with default parameters and are publicly available on FastText's website[8].

4.2. Results

In this section, we present the results obtained with the different models and datasets described in previous Sections 4.1 and 3. In Figure 2, we present the similarity metrics between our generated rankings and the ranking generated in [7]. The parameters used by each model are specified in Table 2. In Figure 3 we assess the ability of our models to distinguish right-wing from left-wing communities using the well-known Area Under the Receiver Operating Characteristic (AUC ROC) score.

Table 2
Models parameters. All other parameters are default values.

Model	Parameters	Dataset
FastText-raw	epoch=1, dim=300, without pretrained-vectors	Full
FastText-pretrained	epoch=1, dim=300, with pretrained-vectors	Full
FastText-truncated	epoch=5, dim=300, with pretrained-vectors	Small
Cohere's	No training	Small

We can observe that Cohere's model consistently outperforms all three FastText-based models in all four metrics and achieves the highest AUC ROC score using the 2018 data. Our hypothesis is that Cohere's model is capable of capturing more subtle patterns within each community that might be overlooked by skipgram-based models. Recall that Cohere's transformer-based

[7] https://www.reddit.com/policies/privacy-policy
[8] https://fasttext.cc/docs/en/pretrained-vectors.html

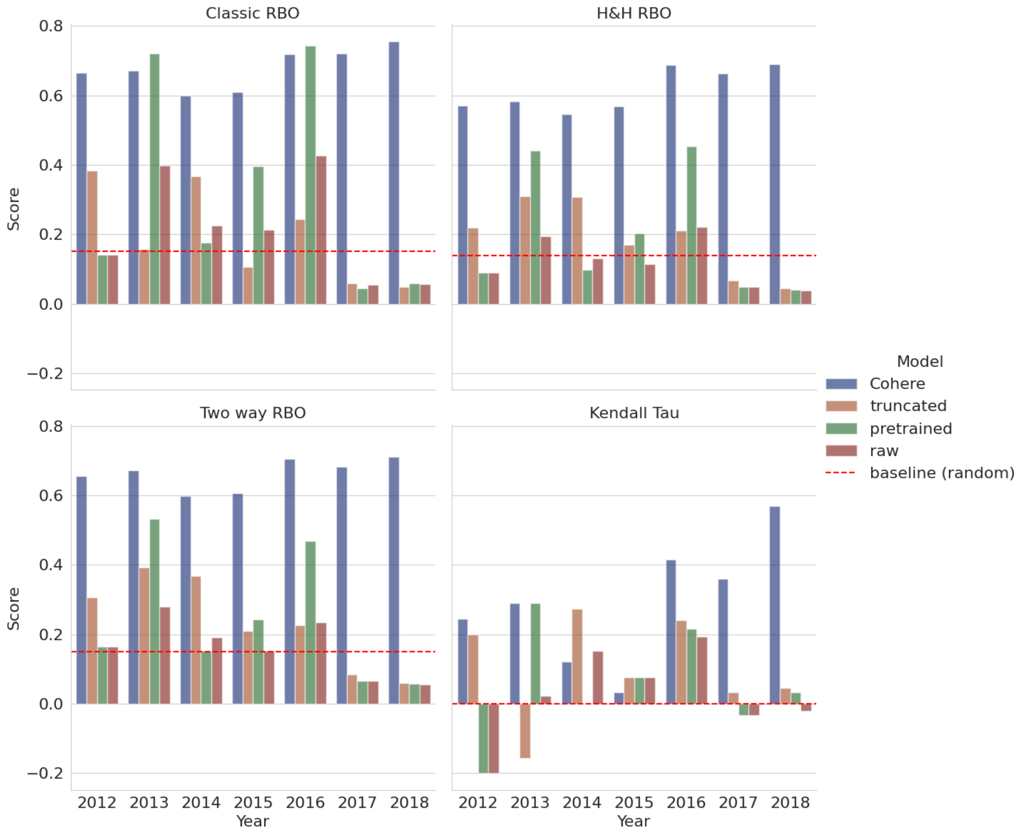

Figure 2: Comparison of Similarity Measures between our rankings and Waller's gold standard. In this figure, the dashed red line represents a baseline using random rankings for each corresponding measure, serving as a point of reference. This comparison offers valuable insights into the performance of our rankings relative to the established gold standard.

model is a larger and more complex architecture than the simpler skipgram-based FastText model. This observation aligns with the results obtained in previous works [14], where it is demonstrated that another transformer-based model (BERT) is able to distinguish between the two communities' ways of speaking even when they are very similar, exploiting differences that are not readily perceptible to humans. We obtained best results overall on the models trained on data from 2016-2018. This could be explained by the imbalance in the number of annual submissions, suggesting that Waller's results might be biased toward the 2016-2018 data.

To further emphasize the similarity between both sets of results, we present a bump chart in Figure 4 for our best performing ranking: Cohere's model using 2018 data. Remarkably, we can observe that Cohere's model correctly aligns both extremes (*Conservative* and *democrats*) but appears to face challenges in ranking non-traditional partisan supporters (The_Donald, new_right, TrueChristians, EnoughSandersSpam).

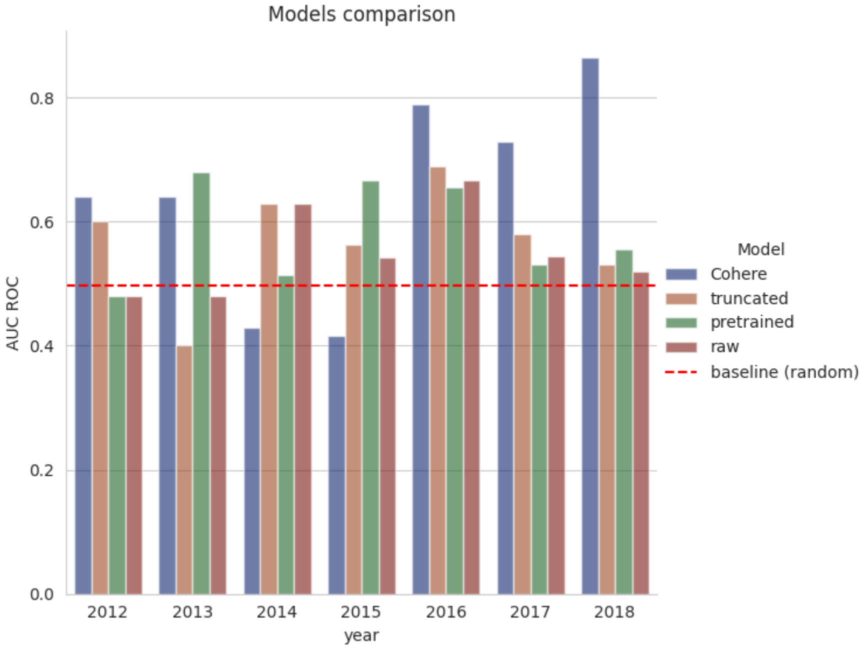

Figure 3: Assessing distinction between right-wing and left-wing communities. This figure presents AUC ROC scores as a measure of how well the model distinguishes between right-wing and left-wing communities for each model using annual data spanning 2012 to 2018. Higher AUC ROC scores indicate superior discrimination ability, offering insights into the model's performance and its ability to capture evolving distinctions between these communities.

5. Discussion

In this section, we present the conclusions of our study, discuss the limitations we encountered, and outline the directions for further analysis in future work. We share insights derived from the application of the method described in Section 3, which includes the utilization of different language models and the data described in Section 4.1.

5.1. Conclusion

We developed an NLP-driven pipeline designed to quantify partisan tendencies within Reddit communities. We evaluated the performance of various configurations, including two distinct embedding techniques: FastText [8] and Cohere's model [9]. These methodologies were tested on two datasets, as detailed in Section 4.1, and their outcomes were subsequently compared. Our most successful approach, which employed Cohere's model on the *Small* 2018 dataset, closely aligns with the findings of Waller et al. [7].

Our pipeline incorporates both the efficient FastText language model and the newer, more intricate Cohere language model. Cohere's model consistently emerged as the superior performer across all four similarity measures and in assessing the distinction between left-wing

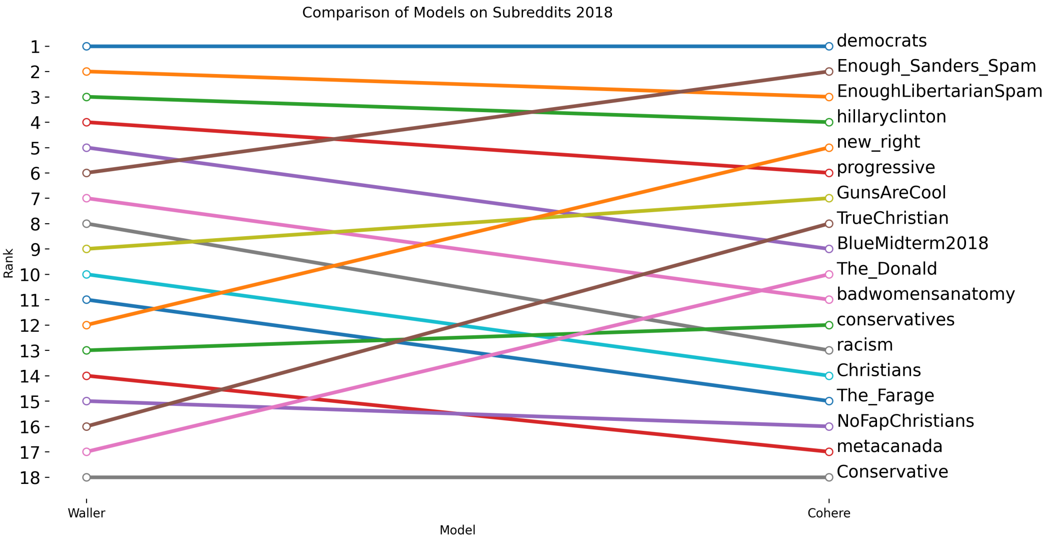

Figure 4: Bump Chart Comparing Waller's Gold Standard (Left) and Cohere Using Small Dataset 2018 (Right). This chart illustrates the differences in rankings, highlighting the specific items where the observed and gold standard rankings agree or disagree. A cross signifies discordance, while horizontal lines indicate agreement between the two rankings.

and right-wing communities. Specifically, the best-performing model, Cohere's model using 2018 annual data, achieved an RBO score of 0.76, a Kendall score of 0.57, and an AUC ROC score of 0.86. As detailed in Section 4.2, our hypothesis is that Cohere's model possesses the ability to discern subtleties in language usage even when similarities are pronounced, as also inferred in [14].

While this approach to quantifying partisan tendencies echoes certain aspects of prior research [7], it distinguishes itself through a fundamental aspect. User interaction-based methods face a critical constraint: they are applicable solely to data collected within a single platform. This restriction, combined with the necessity of human intervention for selecting the initial seeds, demands extensive knowledge of the platform's communities to generate new analyses. Communities may change over time, and what we observe today may not accurately represent the same community as it did 10 years ago. Ultimately, this means that generating new results using the previous method is highly challenging.

Our text-based approach broadens its scope by only requiring text as input, making it possible to select well-known representative seeds for the subject at hand. This increased flexibility facilitates the incorporation of other data sources, such as social platforms like Facebook and Twitter, as well as newspapers, blogs, user-generated content, focus group and oral discussion transcriptions, and others. This enhanced flexibility also empowers us to conduct more detailed analyses than were feasible with previous methods.

5.2. Future Work

The language models used in this work have limited applicability when analyzing data from non-English communities. We believe that multi-language models are a good alternative for analyzing these communities. In future work, we plan to explore models like Claude [9] and GPT[22], which are multi-language and they have demonstrated the best performance in state-of-the-art research [23]. Additionally, these models have wider window sizes, allowing us to utilize more data for each community, potentially improving the quality of the embeddings. We hypothesize that newer and more complex models will yield higher-quality results.

Additionally, we plan to utilize more data sources for a comprehensive analysis of the partisan dimension. Our goal is to dig deeper into partisan differences, examining specific topics such as taxation, social values, and gun control. To achieve this, we propose the inclusion of external text sources that explicitly present each party's perspective. These texts can serve as seeds for the target topic, allowing us to apply the method outlined in this work. Through this approach, we can effectively focus on specific areas of interest without the necessity of identifying two communities that solely differ in terms of the target topic, which may not always be accurately represented by any community.

References

[1] E. Bassignana, D. Brunato, M. Polignano, A. Ramponi, Preface to the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI 2023) co-located with 22th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2023), 2023.

[2] C. Sunstein, # Republic: Divided democracy in the age of social media, Princeton University Press, 2018.

[3] M. Van Alstyne, E. Brynjolfsson, Electronic communities: Global villages or cyberbalkanization?(best theme paper), ICIS 1996 Proceedings (1996) 5.

[4] J. Van Dijck, The culture of connectivity: A critical history of social media, Oxford University Press, 2013.

[5] H. Farrell, The consequences of the internet for politics, Annual review of political science 15 (2012) 35–52.

[6] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, A. Volfovsky, Exposure to opposing views on social media can increase political polarization, Proceedings of the National Academy of Sciences 115 (2018) 9216–9221.

[7] I. Waller, A. Anderson, Quantifying social organization and political polarization in online platforms, Nature 600 (2021) 264–268.

[8] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the association for computational linguistics 5 (2017) 135–146.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polo-

[9] https://docs.anthropic.com/claude/docs

sukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[10] G. Ramponi, M. Brambilla, S. Ceri, F. Daniel, M. Di Giovanni, Vocabulary-based community detection and characterization, in: Proceedings of the 34th ACM/SIGAPP symposium on applied computing, 2019, pp. 1043–1050.

[11] M. Di Giovanni, M. Brambilla, S. Ceri, F. Daniel, G. Ramponi, Content-based classification of political inclinations of twitter users, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 4321–4327.

[12] T. Tran, M. Ostendorf, Characterizing the language of online communities and its relation to community reception, arXiv preprint arXiv:1609.04779 (2016).

[13] P. Lahoti, K. Garimella, A. Gionis, Joint non-negative matrix factorization for learning ideological leaning on twitter, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 351–359.

[14] J. M. O. de Zarate, M. Di Giovanni, E. Z. Feuerstein, M. Brambilla, Measuring controversy in social networks through nlp, in: International Symposium on String Processing and Information Retrieval, Springer, 2020, pp. 194–209.

[15] J. M. O. De Zarate, E. Feuerstein, Vocabulary-based method for quantifying controversy in social media., in: ICCS, Springer, 2020, pp. 161–176.

[16] A. R. KhudaBukhsh, R. Sarkar, M. S. Kamlet, T. Mitchell, We don't speak the same language: Interpreting polarization through machine translation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 14893–14901.

[17] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[18] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[19] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93.

[20] W. Webber, A. Moffat, J. Zobel, A similarity measure for indefinite rankings, ACM Transactions on Information Systems (TOIS) 28 (2010) 1–38.

[21] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, J. Blackburn, The pushshift reddit dataset, in: Proceedings of the international AAAI conference on web and social media, volume 14, 2020, pp. 830–839.

[22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[23] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, arXiv preprint arXiv:2306.05685 (2023).