Susceptibility of Communities Against Low-Credibility Content in Social News Websites

Yigit Ege Bayiz,[1] Arash Amini,[2] Radu Marculescu,[1] Ufuk Topcu[2]

[1] Department of Electrical and Computer Engineering, The University of Texas at Austin
[2] Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin
egebayiz@utexas.edu, a.amini@utexas.edu, radum@utexas.edu, utopcu@utexas.edu

Abstract

Social news websites, such as Reddit, have evolved into prominent platforms for sharing and discussing news. A key issue on social news websites is the formation of low-credibility communities, which often lead to the spread of highly biased or uncredible news. We develop a method to identify communities prone to uncredible or highly biased news within a social news website. We employ a user embedding pipeline that detects user communities based on their stances toward posts and news sources. We then project each community onto a credibility-bias space and analyze the distributional characteristics of each projected community to identify those that have a high risk of adopting beliefs with low credibility or high bias. This approach also enables the prediction of individual users' susceptibility to low-credibility content based on their community affiliation. Our results show that latent space clusters effectively indicate the credibility and bias levels of their users, with significant variance observed across clusters—a 34% difference in the users' susceptibility to low-credibility content and a 8.3% difference in the users' susceptibility to high political bias.

Introduction

Social news websites, such as Reddit and Digg, have emerged as primary platforms for exchanging, archiving, and accessing information. These platforms enable users to share opinions and news articles, and provide an open forum where their users can comment on, discuss, or criticize the news. Their ability to allow news sharing and discussion with minimal censorship allowed social news websites to flourish as open repositories of news from diverse sources and opinions. Social news websites have become a common way for people to access news content. A 2023 report by the Pew Research Center indicates that 8% of U.S. adults *regularly* rely on Reddit for news (Liedke and Wang 2023).

The openness of social news websites also serves as a fertile ground for the spread of uncredible or highly biased information (Sakketou et al. 2022). A notable example is r/politics on Reddit, the largest political news discussion community, where our preliminary analysis of available data indicates that more than half of the shared content do not reference a source, as shown in Table 1. The prevalence of unverifiable news, amplified by the content recommendation algorithms of these sites, tends to reinforce and strengthen users' pre-existing beliefs (Cinelli et al. 2021). This phenomenon leads to significant exposure to news with uncredible or highly biased origins among some user communities. Such communities play a substantial role in propagating uncredible or biased narratives, potentially causing a spectrum of social issues ranging from creating confusion and distracting users from correct news, to leading people to support extremist or hyper-partisan beliefs.

Subreddit	# Ver.	# Unver.	% Unver.
r/Conservative	$37,593$	$64,195$	72%
r/Libertarian	$15,366$	$83,618$	16%
r/democrats	$5,875$	$12,076$	77%
r/Republican	$12,943$	$19,129$	72%
r/politics	$598,844$	$642,634$	52%
total	$670,621$	$821,652$	55%

Table 1: Comparison of numbers of verifiable (Ver.) and unverifiable (Unver.) submissions over the five largest political subreddits in Reddit.

Detecting and countering uncredible or highly biased news content is a well-researched problem. Numerous deep learning methods have been developed to identify such news sources, as highlighted in studies (Zhou et al. 2020; Monti et al. 2019). Additionally, there's a growing trend to employ so-called large language models for this purpose (Hu et al. 2023). Efforts also extend to identifying major spreaders of uncredible content among users (Sakketou et al. 2022). Such efforts aim to detect the optimal targets for preventative methods such as moderation and banning.

This paper adopts a novel perspective by focusing on the detection of *ideological communities* with a high susceptibility to uncredible or highly biased news. In this context, we define an ideological community as a group of users sharing similar opinions, ideas, or beliefs, and exhibiting similar reactions to news articles. We propose a novel comment-based user embedding methodology to create latent space embeddings for individual users that reflect their ideologies. Then we investigate the relation of these embeddings with users' susceptibility to interact positively with uncredibile or highly biased news content. Specifically, we utilize our

embedding method to cluster users into ideological communities. We then analyze the distribution of credibility and political biases of these communities.

Ideological communities may not align with the platform-defined communities or specific interest groups. Rather, they are clusters of users with similar overall opinions and reactions to each other. Also, ideological communities do not necessarily only reflect the political ideology of the user, but the complete collection of all ideas expressed by the user. In fact multiple ideological clusters may align with a political ideology. In this paper, we use the term *community* to refer to ideological communities.

The ideological communities we focus on are related to *echo chambers*, groups of like-minded users sharing and reinforcing existing beliefs with each other (Cinelli et al. 2021). However, we do not require ideological communities to consist of users who are echoing each other's beliefs. We only need them to hold similar opinions. Ideological communities may be caused by echo chambers, or form echo chambers after being established due to *homophily*, the tendency of like-minded users to selectively share among themselves (Cinelli et al. 2021). Yet it is also possible for users with similar opinions to exist in the absence of echo chambers (De Francisci Morales, Monti, and Starnini 2021). In this paper we focus solely on user opinions, and group users of similar opinions into the same community, regardless of whether they form an echo chamber.

Pretrained sentence embedding models like sentence-BERT (SBERT) (Reimers and Gurevych 2019) have significantly advanced the embedding of social media content, enabling research on content clustering and analysis to discern user opinions and biases. However, there is no consensus on inferring user opinion embeddings from the content they engage with. One method involves pooling user-posted content to average the embeddings of each post. This method, though straightforward, is impractical due to the insufficient volume of posts per user for reliable embedding estimation.

We address these challenges by deriving user embeddings from user comments rather than the shared news sources directly. This approach yields a larger data set from users, reducing statistical variance in latent space representations. We provide context to the user comments based on their stances towards the original news post and use this contextual information to assign embeddings to the comments. Then, we use average pooling on the comment embeddings to gather user embeddings. This method ensures that the user embeddings reflect user opinions on a similar latent space to the post embeddings.

To showcase the results of our analysis method, we use real-world data from Reddit, a social news platform with user-generated interest groups called subreddits. On Reddit, users can post opinions or news, and engage with others through comments and replies. After embedding users, we identify user communities and examine their credibility and bias distributions. Using our methodology, we investigate the following research questions:

RQ1: How do user communities in Reddit differ in their susceptibility to credibility and bias?

RQ2: Is it possible to predict the likelihood of a user responding positively to low-credibility news, based on their cluster assignment?

While we use only Reddit data for the experiments in this paper, the methodology is generalizable to all similar platforms, where users can interact with the original post by commenting on it.

Determining the credibility and biases of news sources is often subject to the individual opinions of those who rate them. In this paper, we rely on the data released by Ad Fontes Media,[1], a public benefit corporation that rates news sources based on their credibility and political bias. This dataset includes credibility and bias scores of 223 news sources. We use this data to assign credibility and bias scores to Reddit posts that reference one of these news sources. We call such posts *verifiable* and likewise, if a post does not contain a news link contained in this dataset, we call that post *unverifiable*. Table 1 provides a breakdown of the number of verifiable posts in five major political subreddits in Reddit.

Related Work

Sentence Embedding

Sentence embedding is a critical invention that enables automated analysis of social news content. These models work by assigning a numerical representation of each sentence that preserves the syntactic and semantic relation between sentences. Early approaches to sentence embedding models involve encoder-decoder architectures such as Skip-thought (Kiros et al. 2015), and LSTM-based structures such as siamese BiLSTM (Conneau et al. 2017).

Modern sentence embedding relies on using pre-trained transformer-based architectures (Vaswani et al. 2017). Chief among them is the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019), which set state-of-the-art performance in semantic textual similarity benchmark. Later, RoBERTa (Liu et al. 2019), improved this benchmark performance by utilizing small optimizations in BERT pre-training.

Intrinsically, both BERT and RoBERTa are incapable of achieving sentence embedding as they do not derive independent sentence embeddings. Reimers and Gurevych (2019) enabled drawing such sentence embeddings by introducing sentence-BERT (SBERT) which incorporates a pooling layer after the pre-trained BERT network, and train them using the siamese network architecture in which they train two copies of the same network simultaneously on a sentence similarity or classification objective. In this paper, we use this SBERT architecture for embedding Reddit posts.

Stance Detection

In this paper we use *stance detection* as a part of our user embedding pipeline. Stance detection entails classifying the sentiment of a text, such as a user comment, towards a given target, (Küçük and Can 2020). Stance detection was pioneered by Qazvinian et al. (2011). Later Augenstein

[1]Ad Fontes Media bias dataset is accessible from https://adfontesmedia.com

et al. (2016) achieved state-of-the-art performance by utilizing bidirectional encoding architectures. Modern stance detection relies mostly on transformer models (Hardalov et al. 2022), with Arakelyan, Arora, and Augenstein (2023) achieving state-of-the-art performance.

Recently, Pougué-Biyong et al. (2021) curated a comment-reply dataset with stance labels collected over Reddit. We use this dataset to train and validate our stance detection method, a variant of the method proposed by Gül, Lebret, and Aberer (2024). Recently Luo et al. (2023) achieved state-of-the-art performance on this dataset, providing a baseline to compare our results.

User Profiling

User profiling is the task of assigning a virtual representation to each user, such are keywords, personal information, or numerical latent space representations (Eke et al. 2019). Utilizing user profiling as a detection mechanism for fake news is not a new problem. Shu et al. (2020), introduce *user profile features*, which is a high dimensional user representation that includes location, profile picture, and political bias information, and show that these features, in conjunction with text analysis methods such as linguistic inquiry word-count (Pennebaker et al. 2015) and rhetorical structure theory (Ji and Eisenstein 2014) yield high classification accuracy and recall for false news detection. More recently, Sakketou et al. (2022) achieved state-of-the-art performance in detecting fake news sources by modeling the social interactions between Reddit users with a *graph* and using graph neural networks to classify nodes that are likely to spread fake news. Specifically, they construct a user-to-user graph by traversing the comments under Reddit posts and then use a graph attention network to classify fake news spreaders.

This paper differs from the existing user profiling works in two regards. Firstly, while we develop embedding methods to gather high-dimensional representations of users based on their comments, we focus on extracting user communities from the high-dimensional user representations, rather than analyzing individual users. Secondly, contrary to existing studies we do not measure false news spreading probabilities. Rather, we characterize how user communities, characterized by their long-term commenting behavior, show differences in engaging with news from uncredible, or highly politically biased sources.

Community Analysis

Community analysis is the task of extracting information about the behavior of large groups of users sharing similar opinions, interests, or other measurable traits. The specific definition of communities differs between existing literature, and between websites. In Reddit, the website on which we showcase our algorithms, most existing works focus on analyzing *subreddits*, platform-specific communities of users sharing a common interest. Buntain and Golbeck (2014) analyze the roles of different users in subreddits. They conclude that users who answer questions from other users tend to only be active in a single subreddit. Datta and Adar (2019) investigate conflicts that arise between users within the same subreddits by constructing conflict graphs and conclude that many of the news and politics-related subreddits have internal conflicts, and have subcommunities of opposing opinions, supporting our avoidance of relying on subreddit structure to derive ideological communities. More recently, Sawicki et al. (2023) investigate community-to-community similarities by creating networks of subreddits based on the number of posts they share. They show that significant similarities exist between the types of posts in multiple subreddits, further supporting our choice of relying on user opinions instead of subreddits for our community definition.

Contributions

- We introduce a user embedding pipeline that jointly uses stance detection, together with sentence encoders to obtain latent space representations of users.
- We show that Reddit users create identifiable user communities based on their user embeddings.
- We show that the said communities indicate users' susceptibility to uncredible and highly biased news.

Methods

User Embedding

In this section, we describe a method to embed the users in a high-dimensional latent space. This method works by first assigning an SBERT sentence embedding to posts, and then assign an embedding representation to comments by considering the stance of the comment towards the post. Then, we average the embeddings of all comments sent by each user to obtain a single latent space embedding of each user. This embedding representation captures the average interest and opinions of each user and their stances towards different viewpoints. Figure 1.a shows the overall structure of the user embedding process. In the following sections, we break down each element in this embedding process in more detail.

Post Embedding We embed the entire corpus of post titles using a pre-trained SBERT model, a sentence transformer that provides a high-dimensional latent space representation for each title (Reimers and Gurevych 2019). In Reddit, post titles often mirror the news headlines they reference, providing a contextual basis for estimating embeddings for the comments. We employ 'all-distilroberta-v1'[2] model for encoding post titles into a 768-dimensional array of real numbers. We chose this model for its high variance in cosine similarities across post titles in the dataset, a feature likely stemming from its extensive training on Reddit conversations (Henderson et al. 2019).

Stance Detection In the context of comment discussions, stance detection, or more specifically (dis)agreement detection, is the task of identifying the stance of a *parent* text, usually the original post or a comment, to a *child* text, which is a reply to the parent. we define the possible stances of one text to another using three discrete categories:

- favor: The child text is supportive of the parent text.

[2]The model is accessible from https://huggingface.co/sentence-transformers/all-distilroberta-v1

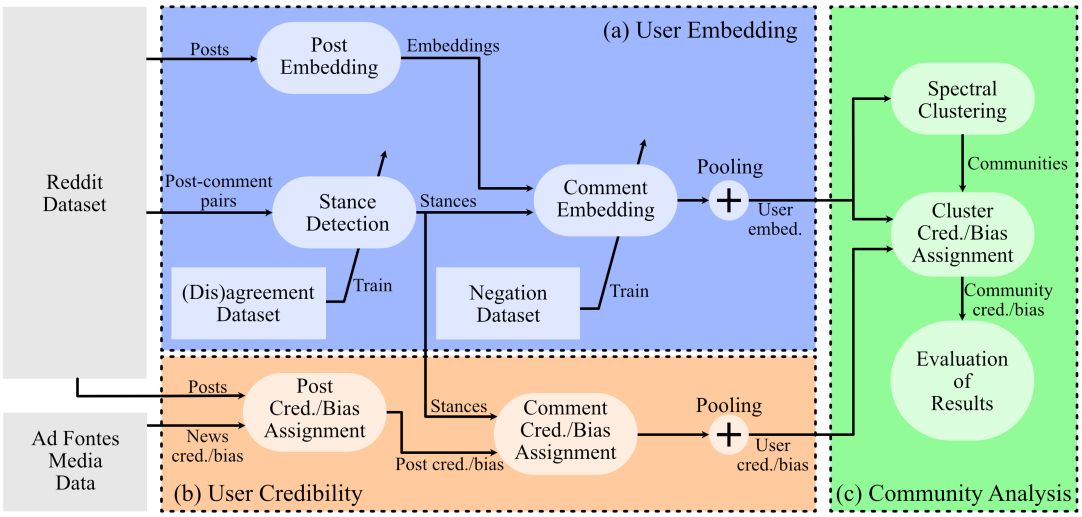

Figure 1: Overview of methodology: This diagram illustrates our analysis approach. Rounded boxes depict various processes, while sharp-edged boxes depict datasets. Diagonally upward arrows behind processes indicate the use of corresponding datasets for training these processes. The pooling blocks, indicated with a + sign denote *averaging* the inputs for each user.

- against: The child text opposes or otherwise criticizes the parent text.
- none: The child text is neutral towards the parent text.

We use the LLaMa-2-7b (Touvron et al. 2023), a large language text generation model with 7 billion parameters, to classify the stances of each comment towards its parent under the context of the post. We do this classification by running a text completion task on LLaMa-2-7b using a prompt inspired by Gül, Lebret, and Aberer (2024) and querying the LLaMa-2-7b to generate one of three possible stances as an answer. We include the full query we use in Appendix B. We limit the text completion to a maximum of 7 new tokens, as this is the maximum amount of tokens required to return one of the three possible stance options. Then we compare the generated part of the text and match it to one of the three possible stances we have.

Using the default model without any fine-tuning, the above approach performs very poorly at a classification accuracy approaching 0%. This is because there is no guarantee for the model even to generate text matching one of the options we require. To fine-tune the LLaMa-2-7b model, we use a Low-Rank Adaptation (LoRA) (Hu et al. 2022) (Gül, Lebret, and Aberer 2024). LoRA introduces linear layers parallel to each layer in the LLaMa-2-7b model. These layers are matrices that linearly transform the input and add their output to the original layer outputs. Each weight layer of the LoRA is a matrix of rank r, which is a hyperparameter we choose. This fine-tuning process effectively modifies the output of the text generation model while maintaining training efficiency. LoRA layers also have a second hyperparameter α which controls how much the output of the LoRA layers are scaled. Tuning this parameter can help increase the data efficiency of LoRA further. However, it is common practice to select α and r equal to each other, and throughout the rest of this paper, we always pick them the same.

In our experiments, we use a LoRA model of rank $r = 16$ for fine-tuning the base LLaMa-2-7b model. We set the learning rate to 4×10^{-4} and employ 10% dropout (Srivastava et al. 2014) during training. To train and validate the LoRA model we use the **(dis)agreement dataset**, which is an agreement disagreement dataset containing expert-labeled comment-reply pairs collected from Reddit. We provide more information on this dataset in the Data section of this paper.

Despite providing a high comprehension of human language, LLaMa-2-7b has the drawback of slowing down runtime, therefore for future applications, fine-tuning a discriminative model such as SBERT may provide better performance. In our experiments, we observed that querying the LLaMa-2-7b for a stance detection task takes about twenty times longer than passing the comment-reply pair through an SBERT model such as 'all-distilroberta-v1'. The modular nature of our clustering framework facilitates the replacement of the stance detection model with a discriminative large language model. At the time of this study, we employed the most straightforward and accurate model available—the fine-tuned LLaMa 2 model—for stance detection, a field experiencing rapid growth due to the rising prominence of large language models. As depicted in Figure 1, our analysis pipeline remains adaptable and is likely to benefit from the advancements in stance detection techniques that are to emerge.

Comment Embedding The goal of embedding the comments is to identify the embedding of users in a similar SBERT latent space to the one we use in post-embedding. We achieve this by embedding the entire corpus of all comments of each user and then pooling them by averaging for each user to get embedding representations for the user. Intu-

itively, user embeddings capture the overall opinions, biases, and interests of users.

Unlike the original posts, comments and replies on social news websites rarely express complete statements or opinions on a specific topic by themselves. Thus, it is difficult to get latent embeddings of the opinion a comment expresses by directly using text embedding without using the original post as the context. As an illustration, consider the following example Reddit post-comment pairs.

Post 1: *China will surpass US to be world's largest economy.*

Comment 1: *valid point though.*

Post 2: *Trump believed Comey intentionally misled the public to believe that he was under investigation*

Comment 2: *Completely valid point.*

Here, it is clear that comment 1 and comment 2 express vastly different opinions and biases despite nearly having the same textual content. Thus directly relying on comments to get user opinions does not work well without the original post providing context. We resolve this issue by using the original post embedding as a context and then assigning an embedding to each comment based on its stance towards the original post.

Suppose that we have a post P with the corresponding embedding $h(P)$ obtained by encoding the post title through SBERT. Now consider a comment C to this post where the stance of the comment to post is denoted as $\sigma(C, P)$, where

$$\sigma(C,P) = \begin{cases} 1 & \text{if } C \text{ favors } P, \\ -1 & \text{if } C \text{ is against } P, \\ 0 & \text{if } C \text{ is neutral towards } P. \end{cases} \tag{1}$$

Ideally, if the comment C entails post P, their embeddings should be similar, and if the comment C is against post P, its embedding should either be similar to the negation of P if P is a logical statement or contain a contradictory opinion to P. Thus letting $\neg P$ represent a negation to P we define the embedding $h(C)$ of the comment C as

$$h(C) = \begin{cases} h(P) & \text{if } \sigma(C,P) = 1, \\ h(\neg P) & \text{if } \sigma(C,P) = -1, \\ \frac{h(P)+h(\neg P)}{2} & \text{if } \sigma(C,P) = 0. \end{cases} \tag{2}$$

The above assignment relies on knowing what the string $\neg P$ is, which is generally impossible. In fact, a rigorous definition for a negation string $\neg P$ might not exist.

We overcome this issue by training a model to directly estimate $h(\neg P)$ from $h(P)$ on a dataset where negation strings are well-defined and known. Generalizing this model to arbitrary strings yields a model that transforms any given embedding into an embedding of a contrary opinion.

To prevent overfitting, we use a simple affine transformation as our negation model. That is, we find a matrix $\mathbf{A}$ and a bias vector $\mathbf{b}$ that transforms the embedding of a given string S into the embedding of its negation $\neg S$ with minimal mean squared error loss $||\mathbf{A}h(S) + \mathbf{b} - h(\neg S)||_2$ over some well-known negation dataset, and then *stipulate* that for any post P the embedding of its negation is $h(\neg P) = \mathbf{A}h(P) + \mathbf{b}$.

Note that in the ideal case where the input data consists only of strings that define logical statements with a clearly defined negation, the affine transformation induced by A and b would be an *affine involution*, that is, it would be its own inverse, as the negation of a negated statement must be the original statement. We do not enforce this constraint on our affine model as the sentence embeddings are not ideal, and unconstrained affine models can yield higher accuracy.

We use **negation dataset**, which is a dataset containing sentence entailment and negation examples, to train the affine negation model parameters $\mathbf{A}$ and $\mathbf{b}$. We train the model both to transform the entailment examples to negation examples, and vice versa to ensure the model generalizes well to the negation of negative statements. Note that we use mean squared loss in fitting the affine model, instead of the cosine error, despite the latter being more common in language modeling tasks. Empirically we found the mean squared loss to provide better performance in later clustering steps as it limits the norm of the predicted $h(\neg P)$ embeddings to be small. After fitting the affine model the comments embeddings become

$$h(C) = \begin{cases} h(P) & \text{if } \sigma(C,P) = 1, \\ \mathbf{A}h(P) + \mathbf{b} & \text{if } \sigma(C,P) = -1, \\ \frac{1}{2}((\mathbf{A} + \mathbb{I})h(P) + \mathbf{b}) & \text{if } \sigma(C,P) = 0. \end{cases} \tag{3}$$

where $\mathbb{I}$ denotes the identity matrix.

The *user embeddings* follow directly from the comment embeddings by pooling all comments written by a user and averaging them over the time period of interest.

Credibility and Political Bias Analysis

We determine the credibility of each user by first assigning a credibility score, ranging from 0 to 1 to the original posts, then assigning scores to each comment based on the credibility of the parent post and the stance of the comment toward the post. Finally, we average the credibility of the comments of each user to get an average credibility score for each user. We use this score as a metric for how likely each user is to engage positively with low credibility content, with a lower score meaning higher susceptibility to uncredible sources. We follow the same steps to estimate the political bias of each user as well, with the only difference being that the bias scores range from -1 to 1, denoting left-wing and right-wing political views respectively. Figure 1.b presents a visualization of the credibility and political bias assignment process.

Post Credibility and Bias We determine the credibility and biases of the posts using the credibility-bias rankings for news sites published by the data released by Ad Fontes Media Corporation. Approximately 29% of the posts included in the four political subreddits in **Reddit dataset** include a reference to a verifiable news article. We then assign a credibility and bias rating for the post using the news article it references.

Comment Credibility and Bias We define a comment's credibility using the following equation

$$\text{Cred}(C) = \sigma(C, P)\left(\text{Cred}(P) - \frac{1}{2}\right) + \frac{1}{2}, \quad (4)$$

where $\text{Cred}(C)$ and $\text{Cred}(P)$ denote the credibilities of the comment and its parent post, respectively. Notice that we rely on the user stances $\sigma(C, P)$ we derive from the fine-tuned LLM model described in the comment embedding section. That is to say, we assign the same credibility to the comment and post when the comment favors the post and assign credibility $1 - \text{Cred}(P)$ to the comment when it is against the post. We define the comment biases using a similar method, but as bias assignments are centered around 0 we simply write

$$\text{Bias}(C) = \sigma(C, P)\text{Bias}(P). \quad (5)$$

Similarly to the case with comment embedding, we use pooling to derive user credibility/bias assignments from the comment averaging them for each user.

Community Susceptibilities

After obtaining both user credibility and bias scores along with the user embeddings, we can analyze the credibility of user groups. While Reddit includes user-generated communities called subreddits, we avoid using these subreddits directly and instead follow a clustering-based approach to detect distinct interest groups. Subreddits have two main limitations that are problematic for determining groups of users with particular interests. Firstly, multiple user groups might exist in a single subreddit, creating separate competing cliques within each subreddit that do not agree with their interests. Perhaps the most stark example of this is the largest political discussion subreddit r/politics, where political views of all walks contribute and share news, and predictably discussions involving different political opinions are abundant. The second reason is that some users opinions might be tempted to follow and participate in multiple subreddits, for example, around half of all users in the subreddit r/Republicans also comment regularly in the subreddit r/-Conservatives, making these subreddits ineffective in terms of partitioning users into distinct interest groups.

Instead of using subreddits, we use *spectral clustering* (Yu and Shi 2003; Damle, Minden, and Ying 2018) on user embeddings to detect interest groups. To better combat the noise in the user distributions, we adopt a local scaling method (Zelnik-manor and Perona 2004). In this method, we first compute the pairwise cosine distances

$$d(x, y) = \frac{\mathbf{x}^\top \mathbf{y}}{||\mathbf{x}|| \cdot ||\mathbf{y}||}, \quad (6)$$

of all user pairs (x, y) with respective latent space embeddings $(\mathbf{x}, \mathbf{y})$. We then define an affinity $\mathbf{W}$ using a Gaussian kernel as

$$\mathbf{W}_{x,y} = \exp\left(\frac{-d(x, y)^2}{\sigma_x \sigma_y}\right), \quad (7)$$

where σ_x, σ_y are the cosine distances to the 7'th nearest neigbors of x and y respectively (Zelnik-manor and Perona 2004).

Year	Subreddit	# Posts	# Comments
2016	r/Conservative	6242	23569
	r/Libertarian	1792	5899
	r/Republican	976	3680
	r/democrats	2398	5553
2017	r/Conservative	7358	28998
	r/Libertarian	2169	7938
	r/Republican	580	2240
	r/democrats	830	2365
2018	r/Conservative	11146	51646
	r/Libertarian	3850	13739
	r/Republican	400	1421
	r/democrats	2047	5351

Table 2: Post and comment numbers in the Reddit dataset used in the experiments.

Spectral clustering of the latent space then simply becomes spectral clustering on the weighted graph with weighted adjacency $\mathbf{W}$ as described in (Yu and Shi 2003). We determine the number of clusters to split the users into using the self-tuning spectral clustering approach described by Zelnik-manor and Perona (2004), where the authors define an *alignment score* to each number of clusters, and choose the number that yields minimal alignment cost.

Notice that these clusters are based solely on the user embedding, thus there is no explicit dependence between the cluster a user belongs to and their credibility and bias score. We map these clusters onto the credibility bias space using the credibility and bias assignments of each user. We then analyze these distributions to estimate how positively each cluster reacts to high-bias or low-credibility news sources.

Data

In this section we summarize the contents and the preparation of the datasets we use throughout the paper.

Reddit Dataset

This study uses real-world data collected from Reddit,[3] which is a social news website based in the U.S. Reddit discussions are organized in broad, community-generated groups, called *subreddits*, and consist of an original *post*, followed by *comments* and *replies* to the comments. The data consists of the posts and the comments from three consecutive years, starting from January 2016 and ending in December 2018. We chose four subreddits to collect the data from: *r/Conservative, r/Libertarian, r/Republican, r/democrats.* We chose these subreddits as they span a wide range of political biases.

We prune the data by removing all deleted posts, and comments that contain less than three words. We also remove all users with less than 10 comments in any given year, as these

[3]Accessible from https://www.reddit.com. All data collected from https://pushshift.io.

users have too few comments to reliably assign them latent-space representations. We also remove all of the posts that do not contain any comments after pruning as they do not contain any information on users. Table 2 shows the distribution of posts and comments after pruning. The entire corpus of the comments in this pruned data is authored by $3,155$ users, providing an ample source of comments per user to achieve accurate clustering.

(Dis)agreement Dataset

To train and validate the LoRA model, we use an agreement/disagreement dataset (Pougué-Biyong et al. 2021), which contains $42,894$ comment-reply pairs with manually annotated stances between each pair. The dataset covers Reddit posts from five political subreddits: *r/democrats*, *r/Republican*, *r/Brexit*, *r/BlackLivesMatter* and *r/climate*. The stance between each comment-reply pair in the dataset reflects a consensus of at least two out of three annotators, selected randomly from a pool of 519 English-proficient annotators. The dataset reports both the overall stance of each comment-reply pair and the overall confidence of the stance based on the amount of consensus among the annotators.

We generate training and validation sets by first removing all comment-reply pairs that are unlabeled or have conflicting labels between multiple experts. This pruning effectively filters out most of the outliers in the data, which is required as LoRA models are often susceptible to outlier-caused performance losses, a known side effect of their high data efficiency. Next, we sample $10,000$ of these comments-reply pairs and split it into two partitions of $9,000$ and $1,000$ corresponding to training and validation sets respectively.

Negation Dataset

The negation dataset (Günther et al. 2023) consists of $10,000$ sentence triplets. Each triplet consists of the following strings:

- Anchor: A base string,
- Entailment: A string that follows logically from the anchor,
- Negation: A string that contradicts the anchor.

The dataset is based on the SNLI dataset (Bowman et al. 2015), which is composed of sentence pairs, consisting of an anchor and a hypothesis, with human-generated labels describing whether the hypothesis logically follows the anchor, or contradicts the anchor. The entailment and negation strings come directly from the positively and negatively labeled hypotheses that share a common anchor. We embed all of these sentences using the 'all-distilroberta-v1' model. We then split the available sentences into $9,000$ training samples and 1000 validation samples.

Naturally, the entailment and negation strings are negations of each other. When training the comment embedding layer, we thus inflate this dataset by constructing all possible tuples in the form of [entailment, negation], or [negation, entailment], yielding $20,000$ tuples of sentences that contradict each other, yielding a total of $18,000$ training samples and $2,000$ validation samples

Model	Train Acc.	Val. Acc.	F1
Base Model	-	0%	0
Random Guess	-	33%	33%
Rank 4 LoRA	65%	61%	58%
Rank 8 LoRA	68%	65%	63%
Rank 16 LoRA	$\mathbf{73}\%$	$\mathbf{68}\%$	$\mathbf{66}\%$
Rank 32 LoRA	74%	67%	65%
(Luo et al. 2023)	–	68%	67%

Table 3: Comparison of model accuracies of the base LLaMa-2-7b model and four fine-tuned versions. We use the model indicated in bold for the rest of the experiments.

Results

We first present the training and validation results of the LoRA fine-tuning model for stance detection and the affine transformation model for negation embeddings. We then present our main results.[4]

LoRA Fine-Tuned Stance Detection

LoRA fine-tuning of the base LLaMa-2-7b model contributes to a dramatic increase in the accuracy of the stance detection. The majority of this performance increase is due to the fine-tuned model being much less averse to returning text completions that are outside the three allowed categories, with a slight and gradual increase in performance in later epochs due to the fine-tuned model becoming more capable of identifying language nuances in Reddit comment discussions. Table 3 summarizes the classification accuracy of the fine-tuned and the base model, along with some other LoRA training parameters we have tested. We calculate these results empirically over the validation split of the **(dis)agreement dataset**, which contains 1000 comment-reply pairs with manually evaluated ground-truth stances for comparison. These results show performances that are comparable with the reported state-of-the-art results for stance detection tasks using **(dis)agreement dataset** (Luo et al. 2023) which reports a mean F1 score of 66.91%.

In addition to evaluating the results on the validation split of **(dis)agreement dataset**, we also randomly sampled 100 posts from the Reddit Dataset and manually labeled them to create a second manual validation set. Testing LLaMa-2-7b model with Rank 16 LoRA on this validation set yields a 69% accuracy with an F1 score of 64%, which are within one standard deviation of the results we provide in Table 3. Thus our manual evaluation is in accordance with the large-scale validation tests we conduct on the **(dis)agreement dataset**

The entire LoRA fine-tuning took 2 hours and 45 seconds running on a single NVIDIA RTX A5000 GPU.

Affine Negation Model

We train the affine model for sentence embedding negation using **negation dataset** as explained in the methods section. It took 31 epochs for the affine model to reach minimal mean

[4]The codes used to generate the results can be found in: https://github.com/ege-bayiz/reddit-community-susceptibility.

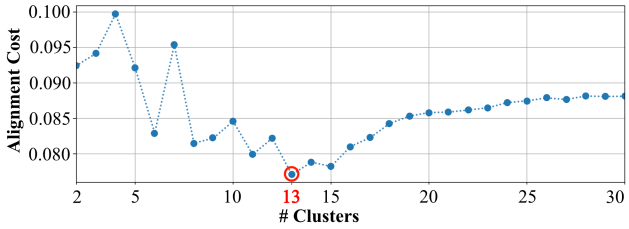

Figure 2: Alignment cost across different numbers of clusters for users' latent space embeddings.

squared loss across the validation set. The training results yielded a mean squared loss of 4.11×10^{-4} and the validation results yielded a loss of 4.88×10^{-4}. The training took approximately 5 minutes and 30 seconds running on a single NVIDIA RTX A5000 GPU.

To verify the effectiveness of the affine negation model, we also compare the cosine similarity between the ground truth negation embeddings in the validation set and the predicted negation embeddings returned by the negation model. We discover that there is an average cosine similarity of 0.79 between the predicted embedding and the ground truth. We contrast this error with the naive approach of inverting the sign of the embedding for negation, which yields a cosine similarity of -0.12, a significant decrease in performance compared to using the affine negation model considering that a cosine similarity of 0 indicates that the ground truth embeddings and predicted embeddings are orthogonal.

Main Results

We investigate the variation in the susceptibility to news sources of differing credibility among Reddit user communities. Results reveal marked differences in how these clusters respond to biases and credibility in news sources (**RQ1**). Specifically, in some clusters, susceptibility to low-credibility news is as much as three times higher than in others. This finding indicates that cluster association is indicative of users' susceptibility to low-credibility news (**RQ2**). The susceptibility of users against biased media shows similar, trends with some communities having significantly more highly biased users than others. These insights underscore the importance of cluster-specific strategies in combating low-credibility and highly biased news propagation.

User Clustering We cluster users based on their latent space embeddings using self-tuning spectral clustering, a technique that captures the underlying patterns in user interactions with posts (Zelnik-manor and Perona 2004). To identify the optimal number of clusters, we calculate the alignment score for each potential cluster number, ranging from 2 to 30. Figure 2 shows alignment scores as a function of the number of clusters, showing that 13 clusters attain the minimal alignment cost, indicating the optimal number of clusters. We find that using 4 clusters, which corresponds to the number of subreddits in our study, attains the highest alignment cost. This disparity suggests that subreddit categories alone do not provide a meaningful way to cluster users based on their interaction patterns with posts. Thus, our analysis justifies the decision not to rely on subreddit categorization for defining user communities

We employ *uniform manifold approximation and projection* (UMAP) (McInnes et al. 2018) to visualize the distribution of the 13 user clusters in a reduced two-dimensional space. UMAP helps in simplifying complex, high-dimensional user data into a format that's easier to visualize. Figure 3A shows the distribution of the users in a two-dimensional UMAP representation. This color-coded UMAP representation demonstrates distinct user communities, visibly separated in the two-dimensional space. This separation serves as visual evidence of the spectral clustering method's effectiveness in categorizing users into discrete and meaningful groups. The UMAP plot also reveals patterns in user behavior, such as the concentration of certain clusters, which warrants further investigation.

The descriptive analysis of clusters reveals significant variations in size and spread. Of the total population of $3,155$ users, 61 are in the smallest cluster and 674 belong to the largest cluster. The populations of other clusters are roughly distributed according to a power law. We measure the spread each cluster occupies in the embedding space by calculating the principal component standard deviation (PC-std), the greatest standard deviation of a cluster across all possible directions. Calculating PC-std equates to finding the square root of the largest eigenvalue of the covariance matrix of each cluster. The PC-std values ranged from 0.036 in the most compact cluster to 0.100 in the most dispersed one, indicating a nearly threefold variation in cluster spans. We sort the clusters based on increasing order of PC-std, meaning cluster 1 is the cluster with the tightest distribution and cluster 13 has the widest distribution. Note that Figure 3A does not represent the spread of these clusters accurately due to the non-linear projection of UMAP, for example, cluster 9, which is the cluster with the fourth highest variance in the embedding space, appears tightly distributed in Figure 3A.

Correlation to Credibility and Bias Figure 3B shows the projection of the user embedding onto the source credibility and political bias space where each color represents a different user cluster. Notice that despite there being no explicit dependence between cluster assignments and user credibility and bias scores, there is a visible separation between the distributions of each cluster. We also visualize the individual cluster distributions in Figure 3C and summarize their distributional characteristics in Table 4.

The results indicate a notable correlation between political bias and credibility with a Pearson's correlation constant of -0.76 among users on social news websites. Specifically, users with left-leaning tendencies tend to have higher credibility scores on average. This correlation contradicts the news source distributions in the Ad Fontes Media dataset. We discuss the implications of this representation further in the discussion section.

A significant observation is that overall, the clusters that have tighter distributions in the latent space also have a tight distribution in the credibility bias space. Indeed, we find a Pearson's correlation of 0.41 between the PC-stds of the la-

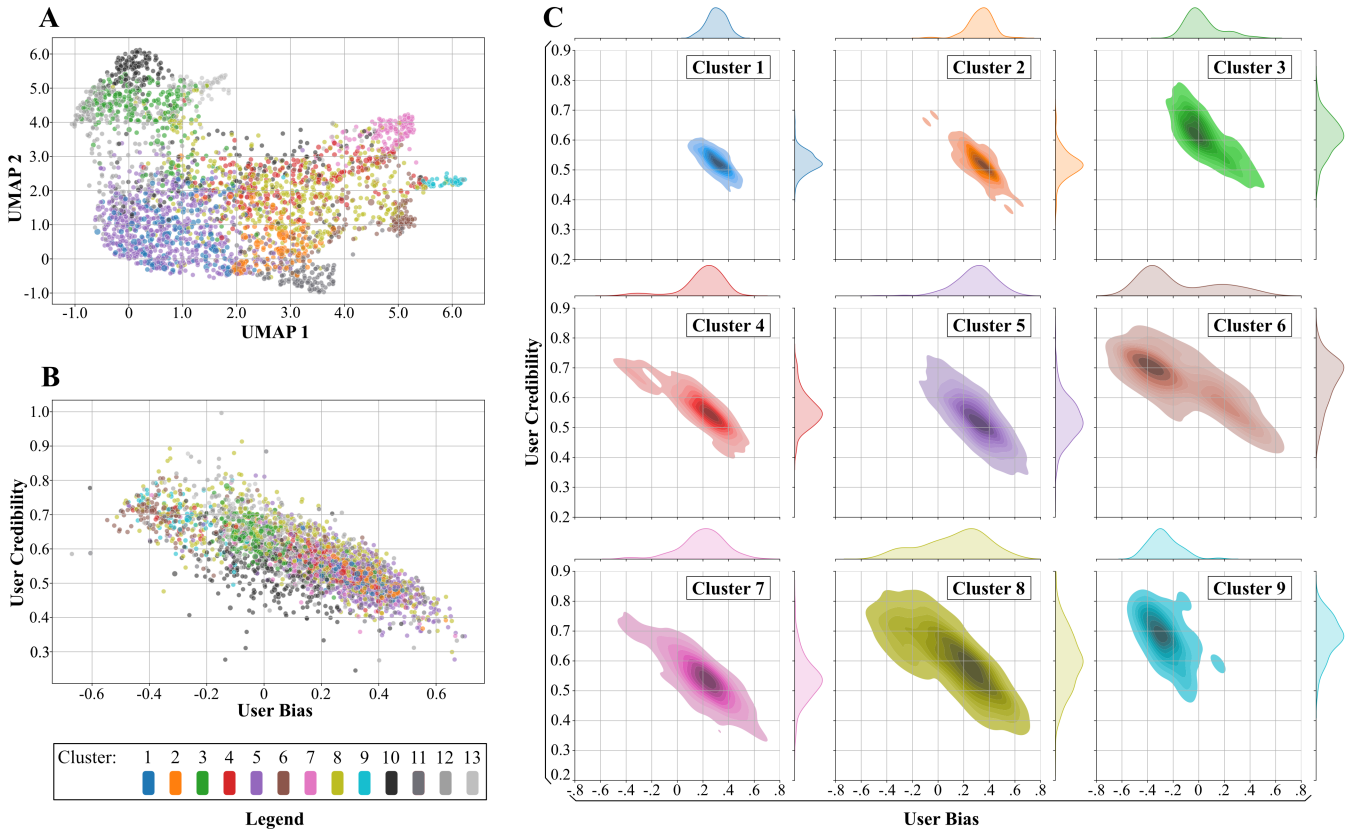

Figure 3: User distributions of all 13 clusters across years 2016-2018. **A:** Latent space embedding visualization of users using UMAP reduction. **B:** Credibility-bias mappings of all users. Larger numbers denote higher credibility and right-wing political bias in their respective axes. **C:** Credibility-bias distributions of 9 clusters with least maximal covariance eigenvalue, together with marginal distributions.

tent space embeddings and the credibility bias embedding, meaning that users that have similar latent space embeddings also tend to have similar credibility and bias scores. The significance of this finding lies in the independent nature of the credibility-bias assignment from the latent space embeddings. This correlation shows that some latent space features are indicative of susceptibility to highly biased or uncredible content.

To understand how these clusters align with subreddits, we examine their *dominant subreddit*—the subreddit where users in a cluster most frequently post comments. We find that none of the 13 clusters have r/Republican as their dominant subreddit, mainly due to the total number of comments from r/Republicans being small compared to other subreddits. There is a major variability between clusters in the proportion of comments included in the dominant subreddit. 99.2% of comments made by users from cluster 9 is in r/democrats, the dominant subreddit of cluster 9. In contrast, this ratio is only 51.0% for cluster 10. The ratio is between these two extremes for other clusters. Table 4 includes the dominant subreddits for all clusters.

We determine the threshold for low-credibility users as those with a credibility score less than 0.5; likewise, we determine the highly biased users as those with bias greater than 0.5 or less than -0.5. These thresholds correspond to low credibility and hyper-partisanship thresholds in the Ad Fontes Media dataset. Overall, the results show that cluster associations strongly impact users' susceptibility to uncredible and highly biased news. Cluster 5 and cluster 9 achieve the highest and the lowest proportion of low-credibility users at 34.7% and 1.6% respectively. This difference is significant as it means a member of cluster 5 is over 20 times more likely to be a low-credibility user than a user in cluster 9. This ratio comparison is not possible for comparing the proportion of highly biased users as three of the 13 clusters have 0 that are highly biased. Table 4 presents the proportion of highly biased and low-credibility users across all clusters.

Our work introduces a credibility/bias score that proves instrumental in identifying echo chambers. We can visualize echo chambers within user clusters by measuring users' tendencies toward biased sources. For instance, clusters 1, 2, and 9 exhibit features characteristic of echo chambers, such as a high degree of tight spread in bias and predominant subreddits. Conversely, clusters like 6 and 8 demonstrate a broader range of spread in terms of bias, indicating a more diverse mix of opinions.

Our findings align with those of Morini et al. (2021), who explored echo chambers' existence and temporal dynamics

Cluster	# Users	Dom. Subreddit	Mean Bias	Std. Bias	Mean Cred.	Std. Cred.	$\lvert\text{Bias}\rvert > 0.5$	Cred. < 0.5
1	164	r/Conservative	0.305	0.076	0.526	0.033	00.0%	18.3%
2	112	r/Conservative	0.331	0.099	0.519	0.045	02.7%	29.5%
3	232	r/Libertarian	0.032	0.138	0.614	0.063	00.4%	04.3%
4	194	r/Conservative	0.197	0.173	0.558	0.057	00.5%	12.9%
5	674	r/Conservative	0.292	0.164	0.528	0.071	08.0%	34.7%
6	168	r/democrats	-0.152	0.297	0.655	0.078	03.6%	04.8%
7	155	r/Conservative	0.198	0.176	0.542	0.071	02.6%	25.8%
8	520	r/Conservative	0.127	0.257	0.601	0.098	05.6%	16.0%
9	61	r/democrats	-0.271	0.116	0.680	0.065	00.0%	01.6%
10	291	r/Libertarian	0.021	0.155	0.531	0.079	00.3%	32.6%
11	169	r/Conservative	0.286	0.220	0.532	0.078	08.3%	34.3%
12	308	r/Libertarian	0.055	0.205	0.622	0.100	03.2%	10.7%
13	107	r/Libertarian	-0.001	0.142	0.641	0.077	00.0%	03.7%

Table 4: Comparison of credibility and bias distributions of clusters.

on specific topics. This is similar to our observations in clusters 1, 2, 5, and 9. Meanwhile, De Francisci Morales et al. (2021) challenge the ubiquity of echo chambers in certain political discussions, a notion supported by our analysis of clusters 3, 6, and 8. The strength of our approach lies in its foundation on embeddings derived from the posts. This enables our model to distinguish not only users' biases and reactions but also their responses to different topics. This provides a significant leap towards automating such analyses, allowing for a more nuanced and detailed study of user behavior in online communities by not relying on users' comments but rather on their reactions to topics.

Discussion

In this section, we interpret and discuss the results of our analysis. We provide the major limitations and discuss avenues for future work.

Implications of Dataset Bias

The **Reddit dataset** features a significant portion of posts from r/Conservatives. This subreddit predominantly features right-leaning content, with over 91% of its shared news originating from right-leaning sources. We explore the implications of this bias in the following discussion.

As mentioned in the results section, our results display a correlation between political bias and credibility in the **Reddit dataset**, where right-leaning users also scored lower in their credibility assignment. It is important to note that this correlation does not necessarily imply any cause-and-effect relationship between political bias and credibility. It is indicative of neither a correlation in the wider population nor in the news sources. Contrarily, the credibility and bias distribution of the news sources in the Ad Fontes Media dataset show that both the extreme right and extreme left news sources are associated with low credibility in a similar fashion. We hypothesize that this discrepancy between user credibility and source credibility may stem from the predominant sharing of extreme-right sources over extreme-left ones in the subreddits under our study. This imbalance causes right-leaning posts to have a lower credibility than left-leaning posts on average in our dataset, which likewise affects user credibility.

Our analysis of the distribution of clusters also shows that most of the clusters have a mean bias score leaning towards right-wing politics. In addition, observing Figure 3C, there are some user clusters, such as cluster 6, that have a distribution that contains both highly left-leaning and highly right-leaning users. This is likely due to incorrectly clustered right-leaning users presenting as noise in cluster 6, which otherwise mainly contains left-leaning users.

Despite these issues, the user clustering in the latent space admits meaningful separation of right-leaning and left-leaning users. This is mainly thanks to the local scaling step we use in the spectral clustering method, which separates tightly packed clusters from the more spread-out background noise caused by incorrectly embedded users.

Limitations

Our study faces several limitations concerning the validity of our results. First, the task of determining the credibility of news sources is inherently complex and somewhat subjective. Relying on a single source for credibility and bias analysis constrains the validity of our conclusions. Second, our method for defining user credibility and biases involves an implicit assumption: users who regularly react negatively to high-credibility content, as opposed to low-credibility content, are considered to have lower credibility. This assumption is not completely unjustified, as users who consistently respond negatively to high-credibility content, in contrast to low-credibility content, demonstrate a discernible preference for the latter. However, it overlooks the additional reasons for a highly credible user to reach negatively to a highly credible news source, such as conflicting political views. These considerations motivate future work aimed at

detecting user credibilities more accurately.

Due to differences in user interface design, adapting our clustering approach to different social news platforms requires fine-tuning the stance detection and comment embedding modules. Hence, we have restricted our analysis to the Reddit platform, and the specific correlations and patterns we observe from the Reddit data may not extend to all social news platforms. Nevertheless, the method we present in Figure 1 does not rely on any data structure that is specific to Reddit, such as subreddits or user upvotes. The methodology we present in this paper generalizes to any social news platform where users can comment on news content. This generalizability is one of the reasons we decided to rely on clustering-based, ideological communities rather than using the subreddits to infer community structure.

The method we present in this paper relies heavily on large language models for the comment embedding and stance detection tasks. Thus, it shares some of the limitations that are present in modern large language models. First, the style and content in online discourse evolve rapidly, and as such, large language models trained on outdated data might produce erroneous results. For example, new phrases, mannerisms, and words in comments might lead to comments that are not representable in the embedding space of SBERT if they are not found in its training dataset. It is, therefore, critical for these models to be up to date with the data they are used to analyze. In our experiments, we ensure this by restricting the user data to be from before the model was last updated. A second, related concern is that since the large language models we use in this paper are mainly tuned for the English language, their performance may suffer when applied to discourse in other languages. Extending the methods we present to other languages may require using sentence embedding and stance detection models fine-tuned to the language of discourse.

Future Work

Future work can expand our analysis in two key directions. Firstly, to improve the validity of our results, we suggest expanding the dataset and refining the methods to assess user credibility and political biases. Advances in stance detection and sentence embedding methods can lead to more accurate user embeddings. These advancements could produce clusters with tighter distributions, allowing a more granular analysis of their credibility and bias distributions. The second major direction for future work is to evolve our proposed user embedding pipeline to include more nuanced effects, such as integrating the content of user comments alongside their stances towards posts. Additionally, employing graph-based methods to capture the interactions among comments could further refine the embeddings and consequently reveal richer conclusions on the credibility and bias susceptibilities of user clusters.

Conclusion

This paper introduces a novel pipeline to derive latent space embeddings of users from the sentence analysis of posts and comments on Reddit. We show that this embedding pipeline induces a clustering of users into communities that have distinctive susceptibilities to incredible and highly biased news sources. Our experiments demonstrate that clusters that are tightly distributed in the embedding space tend to have a tight distribution in the credibility bias space, indicating that the user embeddings we derive are indicative of the credibility and bias scores of the users. Additionally, our research demonstrates that these user-generated communities do not inherently produce natural clusters in the latent space embeddings. This observation suggests that participation in subreddits does not necessarily mirror users' opinions or their responses to specific subjects.

Acknowledgments

This work was supported in part by grants ONR N00014-21-1-2502 and ARO W911NF-23-1-0317.

References

Arakelyan, E.; Arora, A.; and Augenstein, I. 2023. Topic-Guided Sampling For Data-Efficient Multi-Domain Stance Detection. In *Annual Meeting of the Association for Computational Linguistics*, 13448–13464. ACL.

Augenstein, I.; Rocktäschel, T.; Vlachos, A.; and Bontcheva, K. 2016. Stance Detection with Bidirectional Conditional Encoding. In *Conference on Empirical Methods in Natural Language Processing*, 876–885. ACL.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing*. ACM.

Buntain, C.; and Golbeck, J. 2014. Identifying social roles in reddit using network structure. In *Proceedings of the 23rd International Conference on World Wide Web*, 615–620. ACM.

Cinelli, M.; De Francisci Morales, G.; Galeazzi, A.; Quattrociocchi, W.; and Starnini, M. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9): e2023301118.

Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Conference on Empirical Methods in Natural Language Processing*, 670–680. ACL.

Damle, A.; Minden, V.; and Ying, L. 2018. Simple, direct and efficient multi-way spectral clustering. *Information and Inference: A Journal of the IMA*, 8: 181–203.

Datta, S.; and Adar, E. 2019. Extracting inter-community conflicts in reddit. In *Proceedings of the international AAAI conference on Web and Social Media*, volume 13, 146–157.

De Francisci Morales, G.; Monti, C.; and Starnini, M. 2021. No echo in the chambers of political interactions on Reddit. *Scientific reports*, 11(1): 2818.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Eke, C. I.; Norman, A. A.; Shuib, L.; and Nweke, H. F. 2019. A Survey of User Profiling: State-of-the-Art, Challenges, and Solutions. *IEEE Access*, 7: 144907–144924.

Gül, I.; Lebret, R.; and Aberer, K. 2024. Stance Detection on Social Media with Fine-Tuned Large Language Models. arXiv:2404.12171.

Günther, M.; Milliken, L.; Geuter, J.; Mastrapas, G.; Wang, B.; and Xiao, H. 2023. Jina Embeddings: A Novel Set of High-Performance Sentence Embedding Models. arXiv:2307.11224.

Hardalov, M.; Arora, A.; Nakov, P.; and Augenstein, I. 2022. A Survey on Stance Detection for Mis- and Disinformation Identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 1259–1277. ACL.

Henderson, M.; Budzianowski, P.; Casanueva, I.; Coope, S.; Gerz, D.; Kumar, G.; Mrkšić, N.; Spithourakis, G.; Su, P.-H.; Vulić, I.; and Wen, T.-H. 2019. A Repository of Conversational Datasets. *arXiv preprint arxiv:1904.06472*.

Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; and Qi, P. 2023. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. arXiv:2309.12247.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Ji, Y.; and Eisenstein, J. 2014. Representation Learning for Text-level Discourse Parsing. In Toutanova, K.; and Wu, H., eds., *Annual Meeting of the Association for Computational Linguistics*, volume 1, 13–24. ACL.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-Thought Vectors. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Küçük, D.; and Can, F. 2020. Stance Detection: A Survey. *ACM Computing Surveys*, 53(1).

Liedke, J.; and Wang, L. 2023. Social Media and News Fact Sheet. *Pew Research Center*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Luo, Y.; Liu, Z.; Li, S. Z.; and Zhang, Y. 2023. Improving (Dis)agreement Detection with Inductive Social Relation Information From Comment-Reply Interactions. arXiv:2302.03950.

McInnes, L.; Healy, J.; Saul, N.; and Grossberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software*, 3(29): 861.

Monti, F.; Frasca, F.; Eynard, D.; Mannion, D.; and Bronstein, M. M. 2019. Fake News Detection on Social Media using Geometric Deep Learning. arXiv:1902.06673.

Morini, V.; Pollacci, L.; and Rossetti, G. 2021. Toward a standard approach for echo chamber detection: Reddit case study. *Applied Sciences*, 11(12): 5390.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The Development and Psychometric Properties of LIWC2015.

Pougué-Biyong, J.; Semenova, V.; Matton, A.; Han, R.; Kim, A.; Lambiotte, R.; and Farmer, D. 2021. DEBAGREEMENT: A comment-reply dataset for (dis)agreement detection in online debates. In *Neural Information Processing Systems*.

Qazvinian, V.; Rosengren, E.; Radev, D. R.; and Mei, Q. 2011. Rumor has it: Identifying Misinformation in Microblogs. In *Conference on Empirical Methods in Natural Language Processing*, 1589–1599.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*. ACL.

Sakketou, F.; Plepi, J.; Cervero, R.; Geiss, H. J.; Rosso, P.; and Flek, L. 2022. FACTOID: A New Dataset for Identifying Misinformation Spreaders and Political Bias. In *Language Resources and Evaluation Conference*, 3231–3241. European Language Resources Association.

Sawicki, J.; Ganzha, M.; Paprzycki, M.; and Watanobe, Y. 2023. Reddit CrosspostNet—studying Reddit communities with large-scale Crosspost graph networks. *Algorithms*, 16(9): 424.

Shu, K.; Zhou, X.; Wang, S.; Zafarani, R.; and Liu, H. 2020. The Role of User Profiles for Fake News Detection. In *ASONAM*, 436–439. New York, NY, USA: ACM.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15: 1929–1958.

Touvron, H.; Martin, L.; Stone, K. R.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; ...; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arxiv:2307.09288*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yu, S.; and Shi, J. 2003. Multiclass spectral clustering. In *IEEE International Conference on Computer Vision*, 313–319.

Zelnik-manor, L.; and Perona, P. 2004. Self-Tuning Spectral Clustering. In Saul, L.; Weiss, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Zhou, X.; Jain, A.; Phoha, V. V.; and Zafarani, R. 2020. Fake News Early Detection: A Theory-Driven Model. *Digital Threats*, 1(2).

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, see Results and Discussion for advancements, and Ethical Statement sections for privacy considerations.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, they provide a concise summary and motivation.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Methods.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see Discussion.
   (e) Did you describe the limitations of your work? Yes see Discussion
   (f) Did you discuss any potential negative societal impacts of your work? Yes, our result shows a correlation between political leaning and credibility. We explain this and connect it with the potential dataset bias in the Discussion section.
   (g) Did you discuss any potential misuse of your work? Yes, we cover them partially in the Discussion section. We minimize it by ensuring the privacy of the users.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we remain politically unbiased in our Results, and Discussion. We also ensure Privacy.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA.
   (b) Have you provided justifications for all theoretical results? NA.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA.
   (e) Did you address potential biases or limitations in your theoretical framework? NA.
   (f) Have you related your theoretical results to the existing literature in social science? NA.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA.
   (b) Did you include complete proofs of all theoretical results? NA.
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, all of the codes that were used to generate our results, along with the necessary instructions to run them, are publicly available in https://github.com/egebayiz/reddit-community-susceptibility. We also point to this URL in the paper in our Results section, under footnote 4.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, See Data for data splits and Methodology for training details.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No. Our main results involve using the models we train on separate datasets and performing unsupervised learning afterward. The results are not permissive to report error bars. The performance results of the models are included in the Results section separately.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see Results
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see Methods and Results
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No, Only one of the trained models (stance detection) is permissive to such analysis, and we did not find any literature discussing fault (in)tolerance. We do compare with reported accuracies of the baselines.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, immediately after they are introduced.
   (b) Did you mention the license of the assets? Yes, we mention that all sources are open source and publicly available.
   (c) Did you include any new assets in the supplemental material or as a URL? No, we do not introduce new assets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No. All data we use are publicly available from existing datasets. We did not collect any new data.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, in our analysis, we discard and anonymize all personally identifiable information,

such a usernames and post/comment contents, see Ethical Statement.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Ethical Statement

All of the experiments and methods in this paper use publicly available data. We do not disclose any personal information in a manner that jeopardizes anonymity. We did not collect personal information on users except the information available in public datasets, and we anonymized all sample messages and discourse that we show explicitly in this paper. This paper was not subject to an academic IRB process.

Appendix A: Querying for Stance Detection

The stance detection method we use relies on providing a generative large language model with a prompt containing the original post, a parent text, and a comment to the parent text. We then query the model—LLaMa-2-7b—to generate an answer of at most 7 tokens long. Below is the exact query we use in our experiments.

> *[INST]<SYS>You are a helpful, respectful, and honest assistant that detects the stance of a comment with respect to its parent. Stance detection is the process of determining whether the author of a comment is in support of or against a given parent. You are provided with: post: the text you that is the root of discussion. parent: the text which the comment is a reply towards. comment: text that you identify the stance from.*
>
> *You will return the stance of the comment against the parent. Only return the stance against the parent and not the original post. Always answer from the possible options given below:*
>
> *favor: The comment has a positive or supportive attitude towards the post, either explicitly or implicitly.*
>
> *against: The comment opposes or criticizes the post, either explicitly or implicitly.*
>
> *none: The comment is neutral or does not have a stance towards the post.*
>
> *unsure: It is not possible to make a decision based on the information at hand.</SYS>*
>
> *post:* {*post*}
>
> *parent:* {*parent*}
>
> *comment:* {*comment*}
>
> *stance: [/INST]*

Here, the {*post*}, {*comment*}, and {*parent*} represent the original post title, the comment to the post, and the parent to the comment, respectively.