Two-Stage Stance Labeling: User-Hashtag Heuristics with Graph Neural Networks

Joshua Melton[0000−0001−9813−1049], Shannon Reid[0000−0002−4318−4271], Gabriel Terejanu[0000−0002−8934−9836], and Siddharth Krishnan[0000−0002−9570−0186]

University of North Carolina at Charlotte, Charlotte NC 28223, USA
{jmelto30,sreid33,gterejan,skrishnan}@charlotte.edu

Abstract. The high volume and rapid evolution of content on social media present major challenges for studying the stance of social media users. In this work, we develop a two stage stance labeling method that utilizes the user-hashtag bipartite graph and the user-user interaction graph. In the first stage, a simple and efficient heuristic for stance labeling uses the user-hashtag bipartite graph to iteratively update the stance association of user and hashtag nodes via a label propagation mechanism. This set of soft labels is then integrated with the user-user interaction graph to train a graph neural network (GNN) model using semi-supervised learning. We evaluate this method on two large-scale datasets containing tweets related to climate change from June 2021 to June 2022 and gun control from January 2022 to January 2023. Our experiments demonstrate that enriching text-based embeddings of users with network information from the user interaction graph using our semi-supervised GNN method outperforms both classifiers trained on user textual embeddings and zero-shot classification using LLMs such as GPT4. We discuss the need for integrating nuanced understanding from social science with the scalability of computational methods to better understand how polarization on social media occurs for divisive issues such as climate change and gun control.

Keywords: graph neural networks · polarization · social network analysis · stance labeling.

1 Introduction

In today's digital society, online social media have become the main platform for dissemination of partisan and political information as well as the main forum for public discussion of political and social issues. The immense scale and interconnected nature of social media has meant that the scope of public discussion online permits varied and diverse viewpoints about a broad range of topics. Despite the potential for the co-existence of diverse perspectives online, research has shown that online communities are increasingly *polarized*, especially with respect to contentious issues [7, 21, 23, 28]. The increasingly polarized nature of online communities poses a real challenge to many democracies in today's world.

Studies have shown that over the last 30 years, both Democrats and Republicans have become more negative in their views towards the opposition party [7], and these negative views have affected outcomes in areas such as scholarship fund allocation, mate selection, employment decisions, and acceptance of public health measures [9, 20]. Over the last several years, social media platforms like Twitter, Reddit, YouTube, and others have amassed millions of users who engage both with mainstream and fringe media outlets as well as engaging directly with other users on numerous political and social issues. As such, large social media platforms like Twitter present an excellent opportunity to study web-scale user behavior and polarization among differing viewpoints on contentious issues such as climate change and gun control.

The study of polarization online has broadly been divided into two main categories. The majority of computational social science research has focused on analyzing *interactional polarization*, which occurs when users interact to a large extent exclusively with like-minded individuals in highly balkanized communities and are only minimally exposed to users with opposing viewpoint. Such *echo chambers* isolate users in an in-group community that reinforces users' existing beliefs and skews user perspectives towards a particular issue [12]. Another important form of polarization that is less well-studied is known as *affective polarization*. Affective polarization refers to highly negative sentiments expressed by users towards out-groups of users with opposing viewpoints [23, 28]. Prior work on affective polarization has been limited in part due to reliance on manually annotated datasets and challenges in identifying user stances and group memberships in a computationally efficient manner. With recent advances in language modeling, stance detection, sentiment analysis, and network science-based approaches to community detection, researchers have begun to study affective polarization using large, unannotated corpora and interaction networks. Despite this recent progress, much work on affective polarization remains *ad hoc* and incommensurate between studies.

Stance labeling remains a particularly challenging task due to the informal and noisy nature of social media language, the presence of sarcasm and irony, and the potential biases introduced by human annotators and automated models. To further facilitate research into polarization online, we develop a stance labeling method that incorporates both textual content and user social interactions in a scalable two-stage pipeline. In the first stage, we apply a reciprocal label propagation algorithm to the user-hashtag bipartite graph that generates a soft-labeled set of users. In the second stage, we construct a signed, weighted, and attributed user-user interaction network and use transformer language models to represent the content of user posts and to determine the sentiment of interactions between users. Using the soft labels generated from the label propagation stage, we employ semi-supervised training to train a graph neural network classifier to classify user stances.

Our stance labeling approach utilizes an understanding about linguistic and network homophily. Linguistic homophily states that people who share similar ideologies also share similar textual content, not only through similar keyword

and hashtag usages but also through sentiment and linguistic features [15, 27], and the inverse phenomenon—that users with dissimilar views use different language—has also been investigated by researchers [21]. Likewise, the concept of network homophily, which describes the phenomenon that similar nodes tend to be connected to one another has long been studied in graph theory [14, 17] and informs the success of graph neural networks [8, 13, 26]. By combining the content-based and network-based approaches into a single, scalable pipeline for stance labeling, we facilitate further large-scale analysis of both interactional and affective polarization online.

In sum, the contributions of this work are:

- We develop an approach for estimating user stance labels that incorporates both textual features and social network interactions.
- A two-stage pipeline allows our approach to scale to large datasets with soft labeling and semi-supervised training.
- We discuss challenges and potential biases in stance determination of users when relying on short text content as commonly found on social media.

2 Related Work

As social media has become the primary forum for discussion on many topics, including political issues, there has been growing interest in user stance detection [2, 6, 10]. Many existing methods can be divided into two broad categories: textual-based methods that use the content of messages, and network-based methods that use social or conversational structure. Textual-based methods use the content of tweets, user profiles, and hashtags to determine a user's stance on a particular topic. Several studies have used supervised classification of user stance labels where classifiers are trained using a set of features encoding tweet text, hashtags, and user profile information. These techniques have been applied to determine user political ideology [1, 3]. These approaches all generally rely on manually annotated sets of users to train the classifier models, and such methods are often limited by the size and diversity of the training content as manual annotation is both time-consuming and expensive. Furthermore, these methods can be susceptible to annotator bias and poor generalizability due to the limited distribution of training examples.

Topic modeling and user clustering are two unsupervised techniques that have been employed for stance labeling. These unsupervised methods generally embed text or user profile information in a latent feature space and then apply a variety of topic modeling or cluster algorithms to identify texts of similar topics or clusters of similar users. While such techniques avoid the issues with manual annotation for supervised training, they suffer from the *curse of dimensionality*—the search space for a solution grows exponentially with the increase in dimensionality—as there are many more possible patterns than in lower-dimensional subspaces. Selecting useful and informative features, and the computation and memory complexity of clustering methods on high-dimensional data create problems for unsupervised clustering-based methods. For high-dimensional

data, many clustering techniques fail to produce meaningful clusters, but on the other hand, such techniques can be very efficient on low-dimensional features spaces [6]. Several studies have therefore employed dimensionality reduction techniques prior to clustering to first project content embeddings to a low-dimensional space, and such techniques can be robust to class imbalance between different stance groups.

Network-based approaches use the structure of social interaction networks to determine user ideology or stance. Such methods employ techniques such as label propagation, min-cut, and node embedding techniques to leverage structural information from retweet [10] and user follower [4] networks. Xiao et al. [25] generate a multi-relational network using retweets, mentions, likes, and follows for binary classification of user ideologies.

Recent works have combined both textual content and network features in socially-infused text mining. Li and Goldwasser [16] combine user interactions and user sharing of news media to detect bias in news articles, and Johnson, Jin, and Goldwasser [11] combine multiple representations from lexical features and social interactions to categorize tweets by politicians on a variety of political topics. Pan et al. [19] use the network structure and node content to learn node representations for classifying the category of scientific publications. Such works combining textual and network information have demonstrated improvements in user stance detection compared with methods that rely solely on textual content or network structure alone.

Our approach is likewise informed by socially-infused text mining that combines information from text and network relationships for user stance labeling, but the approach we present in this paper is unique from the methods described above in that we i) combine language features from state-of-the-art transformer language models with the structure and sentiment of social interactions using graph neural networks, and ii) our two-stage process allows our method to scale to large datasets with minimal supervision.

3 Methods

This section describes our proposed method to estimate the stance of Twitter users on political issues. We first describe the collection of data for the gun control and climate changes datasets. Then we characterize the hashtag-based heuristic method to generate soft labels for two stance groups of users, which are used as seed users for training stance labeling methods. We then introduce several baseline methods followed by our GNN-based approach for stance labeling.

3.1 Data Collection

We collect tweets related to two prominent and contentious political issues—climate change and gun control. The keywords listed in Table 1 are used to scrape the initial set of relevant tweets related to each topic. To fully capture

conversation cascades and user interactions, we then recursively collect all referenced tweets from each tweet in the initially matched set. This ensures that all available tweets from relevant conversations are included in our analysis, even if particular tweets do not explicitly use any of the keywords. The climate change dataset consists of tweets published between June 1st, 2021 and May 31st, 2022. In all, the climate change dataset consists of a total of 46M tweets authored by 4.8M unique users and contains 726,378 conversation threads of at least three tweets. The gun control dataset consists of tweets published between January 1st, 2022 and December 31st, 2022. In all, the gun control dataset consists of a total of 14.4M tweets from 2.66M unique users containing 335,000 conversation threads of at least three tweets.

Table 1. Set of keywords used to collect tweets related to Climate Change and Gun Control from Twitter.

Climate Change		Gun Control	
climate change	global warming	gun control	gun rights
climate hoax	global warming hoax	second amendment	2nd amendment
global cooling	#ActOnClimate	#guncontrol	#guncontrolnow
#ClimateChange	#climatechangehoax	#gunreform	#gunviolence
#globalwarminghoax	#globalcooling	#endgunviolence	#2a
#globalwarmingisahoax	#climatehoax	#nra	#gunrights
		#secondamendement	#shallnotbeinfringed
		#righttobeararms	

These issues are good models for understanding the differences in behavior on social media for topics that are highly event driven, like gun control, and those that are less focused on particular events, like climate change. Figure 1 illustrates that conversation threads related to climate change are started at a fairly consistent rate over time with some spikes in conversation activity connected with specific events. But compared with the conversation activity on Twitter surrounding gun control, illustrated in Figure 2, the frequency of discussions surrounding gun control spikes sharply in response to mass shooting events. The summer of 2022 saw several deadly and highly publicized mass shootings in Buffalo, Uvalde, and Chicago that generated large amounts of discourse surrounding gun control and gun violence on social media, illustrated by the sharp spikes from the end of May through late July. Shortly after mass shooting events though, the level of discourse abates, returning to near baseline levels between early August until November, when a shooting event at UVA caused a small spike in conversation activity. Event driven issues such as gun control exhibit different posting behavior and user interactions that we hypothesize lead to differences in polarization and require different techniques for intervention and mitigation of polarization on social media. Developing approaches for stance labeling that

are both topic independent and robust to different types of discourse online is an important step for facilitating further analysis of polarization online.

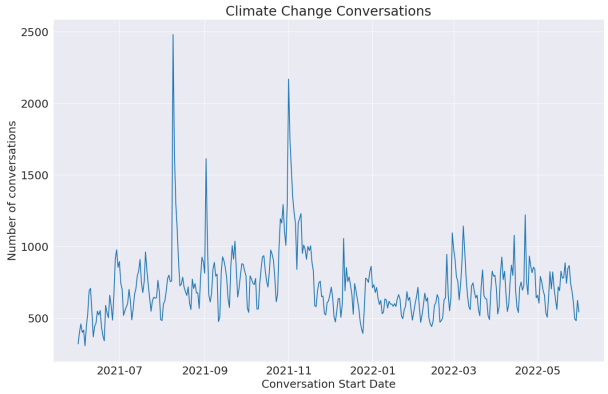

Fig. 1. Timeline of the number of Twitter conversations about climate change started each day.

3.2 User-Hashtag Stance Labeling

From the set of collected tweets for a given topic, we first construct a bipartite graph connecting users with the hashtags that each user posted. The user-hashtag bipartite graph $\mathcal{G} = (V, E)$ consists of two disjoint node sets containing the users and the hashtags posted in tweets by the user set. An edge $(u, h) \in E$ indicates that a user u posted a tweet containing hashtag h. The weight $w(u, h)$ represents the number of times the author posted a particular hashtag across all of their authored tweets. Using the bipartite user-hashtag graph, we then develop a reciprocal label propagation-based method for automatically assigning stance labels to users and hashtags.

Table 2. Sets of seed hashtags used for heuristic stance labeling of users in the climate change and gun control datasets.

Climate Change		Gun Control	
Believe	Disbelieve	Pro	Anti
actonclimate	climatechangehoax	guncontrolnow	shallnotbeinfringed
climatecrisis	globalwarminghoax	endgunviolence	righttobeararms
climateaction	globalcooling	gunreform	gunrights
climateemergency	globalwarmingisahoax		
climateactionnow	climatehoax		

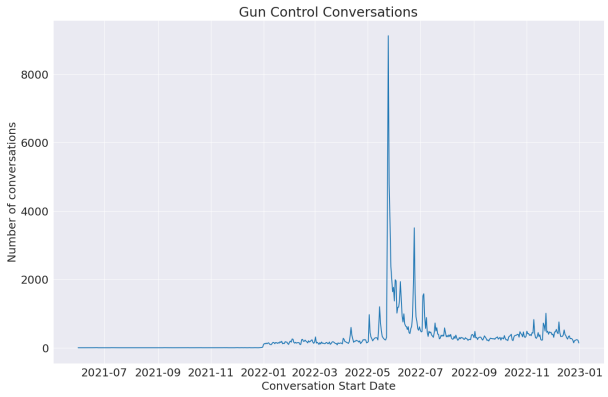

Fig. 2. Timeline of the number of Twitter conversations about gun control started each day.

Based on a small (3-5) set of seed hashtags associated with each stance, we first propagate the stance associated with each seed hashtag across the user-hashtag bipartite graph to the set of users who posted a given labeled hashtag. The importance of a hashtag's stance to a particular destination user is weighted by the number of times the user posted each hashtag, as described in Algorithm 2. Users are then assigned a label based on the number of hashtag usages for each stance group. User labels are then propagated back to the hashtags used by each user, as described in Algorithm 3. Each hashtag is then scored based on the difference between the normalized count of usages from users of each stance group. Hashtag scores are normalized via min-max normalization, and the mean and standard deviation of hashtag scores is computed. Hashtags that score one or more standard deviations above or below the mean are assigned the corresponding stance label. We then repeat alternating label propagation until the model converges, i.e. the set of labeled users is the same from the prior iteration to the next.

While the user-hashtag bipartite label propagation method can quickly label a number of users with high precision. It suffers from the limitation that a large proportion of users do not use hashtags, and while it is well suited for determining hashtags that are highly associated with one particular stance group as compared with the other, hashtags that are used across both stances albeit in different contexts will remain unlabeled.

3.3 GNN Stance Labeling

The user-hashtag stance labeling method can efficiently determine the stance group of users and discover important hashtags associated with each stance

Algorithm 1 Bipartite stance label propagation algorithm

Require: Bipartite user-hashtag graph G, seed hashtag set H^0, s stance label

$H^0 \leftarrow \{(h, s) \in \text{Set of labeled seed hashtags}\}$
$U^0 \leftarrow \emptyset$
while $U^{(k-1)} \neq U^k$ **do**
    $U^k \leftarrow propagateTagsToUsers\left(G,\, H^{(k-1)}\right)$
    $H^k \leftarrow propagateUsersToTags\left(G,\, U^k\right)$
end while
return U^K, H^K

Algorithm 2 Propagate hashtag labels to users

Require: Bipartite user-hashtag graph G, labeled hashtag set H, s stance label, $\mathcal{N}$ neighbor set

user_counts = {}
for $h, s \in H$ **do**
    for $u, w \in \mathcal{N}(h)$ **do**
        user_counts[u][s] += w
    end for
end for
for $u \in$ user_counts **do**
    $U \leftarrow (u,\, \max_s(\text{user\_counts[u]}))$
end for
return U

Algorithm 3 Propagate user labels to hashtags

Require: Bipartite user-hashtag graph G, labeled user set U, s stance label, $\mathcal{N}$ neighbor set

tag_counts = {}
for $u, s \in U$ **do**
    for $h \in \mathcal{N}(u)$ **do**
        tag_counts[h][s] += 1
    end for
end for
for $h \in$ tag_counts **do**
    $h_{score} = \frac{\text{tag\_counts}[h][s_2]}{|U_{s2}|} - \frac{\text{tag\_counts}[h][s_1]}{|U_{s1}|}$
end for
$\mu = \text{mean}(h_{score} \,|\, h \in \text{tag\_counts})$
$\sigma = \text{stdev}(h_{score} \,|\, h \in \text{tag\_counts})$
for $h \in$ tag_counts **do**
    if $h_{score} \geq \mu + \sigma$ **then**
        $H \leftarrow (h, s_1)$
    else if $h_{score} \leq \mu - \sigma$ **then**
        $H \leftarrow (h, s_2)$
    end if
end for
return H

group, but it is limited by the fact that not all users post hashtags and that hashtags consist of only a small part of the overall textual content contained in tweets. Furthermore, the method does not take into account any social interactions between users. The hashtag-based heuristic has high precision but relatively low recall due to the limited breadth of users covered by the stance labeling method. To develop a scalable model capable of labeling users in large-scale datasets, we employ the user-hashtag bipartite method as a means to generate a soft-labeled set of users from which we train a graph neural network classifier over the attributed user-user interaction graph.

We define a user interaction as a retweet, mention, reply, or quote from the author of the tweet towards another user. We then construct a weighted, signed, and attributed user-user interaction network $G = (V, E, X)$, where V is the set of users and a directed edge $(u, v)_i \in E$ connects the author user to the target user of their interaction as defined above. The edge weight $w(u, v)_i$ corresponds to the sentiment of the tweet, scored by a BERTweet-based sentiment analysis model [18]. To characterize the overall interactions between users, we consolidate individual tweets between users by combining each directed edge $(u, v)_i$ into a single edge (u, v) with the composite edge weight $w(u, v)$ computed as the mean sentiment score across all tweets between two users. Each user $u \in V$ is also associated with a feature vector $x_u \in X$ corresponding to the element-wise mean of the BERTweet embeddings of all tweets authored by the user. We train a graph neural network model as a stance label classifier using semi-supervised learning with the set of seed users labeled by our hashtag-based heuristic.

For the GNN, we compare two implementations from the prominent families of graph neural networks—convolutional and attentional. For convolutional GNNs, the neighbor coefficients are fixed weights; whereas, in attentional GNNs these weights are computed implicitly by an attention mechanism. Specifically, we evaluate graphSAGE [8] as an example of a convolutional GNN and graph attention network (GAT) [24] as an example attentional GNN.

GraphSAGE defines a message passing function as:

$$\hat{\boldsymbol{h}}_v^{(k)} = \sigma\left(\mathbf{W^k} \sum_{u \in \tilde{\mathcal{N}}(v)} \frac{1}{|\tilde{\mathcal{N}}|} \boldsymbol{h}_u^{(k-1)}\right) \tag{1}$$

where $\tilde{\mathcal{N}}(v)$ is the local neighborhood of node v with added self-loop, $\mathbf{W}$ is a trainable weight matrix, and σ is a non-linear activation function. We use the mean aggregator to aggregate the neighboring node features.
We also consider a GAT layer operator, defined as:

$$\hat{\boldsymbol{h}}_v^{(k)} = \sigma\left(\mathbf{W^k} \sum_{u \in \tilde{\mathcal{N}}(v)} a_{vu} \boldsymbol{h}_u^{(k-1)}\right) \tag{2}$$

where aggregation coefficient a_{vu} is computed dynamically in a node-dependent manner via edge softmax attention. In this case, the GAT function is a first-order

approximation of a convolution applied over the weighted adjacency matrix, where the edge weights are computed via the attention mechanism.

Table 3. Models utilized for user stance labeling and indication whether the models incorporate textual based information and/or network structural information.

Model Type	Model	Text Content	Network Structure
Random	Weighted random	✗	✗
Transformer	BERT-base-uncased	✓	✗
	BERT-large-uncased	✓	✗
	RoBERTa-base	✓	✗
	RoBERTa-large	✓	✗
	BERTweet	✓	✗
	GPT-4	✓	✗
GNN	GraphSAGE	✓	✓
	GAT	✓	✓

3.4 Baselines

We compare our stance labeling method against six transformer language models that leverage the textual content of tweets to represent users. With each language model, we compute a user embedding by computing the element-wise mean of the embedding of all the users' tweets from BERT, RoBERTa, and BERTweet models. Using the set of labeled users from our heuristic method, we train an MLP classifier to predict user stance label.

We evaluate classification performance of the stance labeling methods using a set of manually annotated users and a set of users annotated via GPT-4 using zero-shot classification. Two social science experts independently annotated a set of 250 users for the climate change dataset and 350 users for the gun control dataset by analyzing the top 20 most popular tweets from each user. We provide the same set of tweets to GPT-4 and prompt the LLM to provide the sentiment of content with respect to gun control or climate change. For the gun control dataset, the function call prompt is: "The stance of the content. Is the message pro gun control, anti gun control, or neutral?" with three classification options of "pro", "anti", or "neutral". For the climate change dataset, "The stance of the content. Does the message capture the user's belief in climate change, disbelief, or is it neutral?" with classification options of "belief", "disbelief", or "neutral". We then aggregate the stance label for each individual tweet to determine a user's overall stance. Users who have a majority of tweets belonging to one stance are labeled accordingly. Users whose tweets are all labeled as neutral or have an equivalent number of tweets labeled as both stances are classified as undetermined.

Table 4. Results for stance labeling classification on the `Gun Control` dataset ($N = 350$). The best scores for each model type are shown in bold and the best overall scores are underlined.

	GPT4 Labels			Manual Labels		
Model	Prec.	Recall	F1	Prec.	Recall	F1
Weighted random	47.36	49.99	39.45	40.76	49.92	38.11
BERT-base-uncased	**84.25**	85.38	**84.66**	83.08	82.75	82.88
BERT-large-uncased	80.47	82.16	79.96	81.44	81.71	81.13
RoBERTa-base	83.33	83.94	83.60	84.79	83.99	84.25
RoBERTa-large	81.39	82.69	79.52	84.41	84.18	83.14
BERTweet	83.59	**85.43**	83.97	85.45	**85.76**	85.13
GPT-4	-	-	-	**89.81**	85.17	**87.05**
GraphSAGE	**87.58**	88.75	**88.02**	91.09	90.46	90.70
GAT	87.17	**88.99**	87.56	**91.28**	**91.46**	**91.36**

Using the set of manual and GPT-4 annotated users for the two datasets, we compare the classification performance of both text-based methods and our GNN-based method informed by socially infused text mining. We additionally compare the performance of GPT-4 for zero-shot stance labeling of users against the manual annotation by domain experts.

4 Results

We report the macro averaged precision, recall, and F1 score for each model, and we use the F1 score as the primary metric due to the class imbalance in the datasets. Metrics reported are averaged across five trials for all models.

4.1 Gun Control Dataset

The classification results for the stance labeling models on the gun control dataset are shown in Table 4. On the gun control dataset, BERT and RoBERTa transformer methods achieve a macro averaged F1 score of between 82%-84% when evaluated on the set of manually annotated users. BERTweet, having been finetuned on tweets, surpasses the performance of the other transformer models achieving an F1 score of 85.13%. GPT-4 zero-shot classification outperforms the trained BERTweet classifier by two percent on F1 score and a four percent increase in precision while maintaining an equivalent recall performance.

For the gun control dataset, the GNN-based models that incorporate network structure information with the content-based representations of users produced by BERTweet, outperform both the trained text-based classifiers and zero-shot classification with GPT-4. Using graphSAGE and GAT to enrich users' content-based representations results in an improvement in stance labeling classification

Table 5. Results for stance labeling classification on the `Climate Change` dataset ($N = 250$). The best scores for each model type are shown in bold and the best overall scores are underlined.

	GPT4 Labels			Manual Labels		
Model	Prec.	Recall	F1	Prec.	Recall	F1
Weighted random	47.36	49.99	39.45	45.26	47.55	42.87
BERT-base-uncased	76.74	79.82	77.36	78.60	82.54	79.36
BERT-large-uncased	75.63	79.14	73.86	76.11	80.28	74.17
RoBERTa-base	76.30	79.92	76.21	76.63	80.92	76.50
RoBERTa-large	75.72	79.36	75.16	76.91	81.33	76.29
BERTweet	**77.26**	**80.78**	**77.64**	78.35	82.66	78.78
GPT-4	-	-	-	**85.36**	**86.08**	**85.71**
GraphSAGE	**84.62**	86.06	**85.25**	**83.80**	85.91	**84.67**
GAT	83.73	**86.43**	84.70	83.47	**87.02**	84.62

compared with the unenriched BERTweet embeddings. With GAT, incorporating users' local neighborhood information with their text-based embeddings improves classification performance by four percent in F1 score. Compared with zero-shot classification using GPT-4, the BERTweet + GAT model performs 3%-4% better in F1 score on classifying user stance labels for the manually labeled user set. Compared with graphSAGE, which takes the mean representation of a node's neighborhood effectively weighting each incident edge on the node equally, GAT utilizes edge softmax attention to dynamically compute edge weights when aggregating node neighborhood information. The dynamic computation of edge importance for users results in a marginal improvement in classification performance.

4.2 Climate Change Dataset

The classification results for the stance labeling models on the climate change dataset are shown in Table 5. Compared with the gun control dataset, where pro gun control users outnumber anti gun control users by roughly a ratio of 2:1 in the heuristic labeled user set, the imbalance in the climate change dataset is significant with climate change believers outnumbering disbelievers by a ratio of nearly 10:1. We also observe that climate change related conversations are less topic-focused than gun control discussions on Twitter. Due in part to the severe class imbalance in the dataset in addition to the less focused of climate change conversation activity on Twitter, stance determination of users in the climate change dataset is more difficult than in the gun control dataset.

Of the BERT and RoBERTa based classifier models, BERT-base-uncased performs the best with an F1 score of 79.36%, narrowly outperforming BERTweet with an F1 score of 78.78%. Enriching the BERTweet user embeddings with structural information from the user interaction graph improves the classifica-

tion performance by six percent to 84.67% F1 score. Despite the performance uplift from the user social network, for the climate change dataset zero-shot classification with GPT-4 outperforms all other models with an F1 score of 85.71%, outperforming other transformer methods by six percent and GNN-based approaches by one percent.

In all, enriching text-based representations of users with structural information from the user-interaction graph results in a significant improvement in classification performance compared with the text-based embeddings alone. Zero-shot classification of user stances using GPT-4 consistently outperforms BERT and RoBERTa-based classifier models, but incorporating social network information surpasses GPT-4 classification performance in the gun control dataset and significantly narrows the gap in the climate change dataset.

5 Discussion

When researching online polarization, serious divides exist in how questions regarding user stance are structured, and this divide is also reflected in how different researchers integrate the literature into future work. Oftentimes, this split can be seen in how social science researchers integrate computer science research into their studies and vice versa. This methodological gap can inhibit the understanding of the phenomena of online polarization and radicalization. Big data research can provide scalability not feasible with traditional qualitative social science research, while social science expertise provides understanding of the subtleties of these interactions on human behavior. As seen in the aftermath of events like radicalized mass shooters, the online behavior of these individuals is not well understood, and we need to understand how and why these processes take place online. The integration of subject-matter expertise with AI is necessary for the application of big data methodology into human-level polarization research and policy work, requiring a balance of the nuance provided by social science subject-matter experts with the speed and volume provided by computation experts. Qualitative research offers tools like narrative analysis to analyze the cultural and historical contingency of the terms, beliefs, and issues narrators address in their writings [22]. The small narratives created from tweets can provide insights into the range of ongoing societal arguments that create collective identity online [5]. The wealth of this data adds critical information to understanding how people discuss their thoughts and opinions about polarizing topics like gun rights or climate change, and the ability to combine these two approaches can greatly benefit the understanding of online polarization.

One area of difficulty in automating stance detection on topics like gun control or climate change is the need to dichotomize beliefs that really exist on a spectrum. The extreme ends of these beliefs are straight-forward to categorize, but the majority of people exist in the space in-between. A better goal is to qualitatively code the tipping point as people's stance moves between stances on issues. Automated stance analysis can mislabel users because of difficulties in discerning written cues like tone, connotation, coded words, or pictures. Some of

these coding errors are due to a purposeful misdirection of a tweet that is clarified by the picture. For example, user 2945287090 said "I'm pro-choice... are you? https://t.co/WtYW6zohSB" but the picture has them wearing a shirt that says pro-choice with an array of guns to "choose" from. Since the picture cannot be assessed by many stance labeling algorithms, these misdirections would be potentially miscoded. There are also users who frame their arguments in neutral language or the language of their opposition. For example, user 66811907 states "Gun control advocates do not have a monopoly on outrage or sorrow. They do not have a monopoly on ideas for addressing gun violence. They do not have a monopoly on wanting to fix the problem." The larger context of this user's tweets balance being pro-Second Amendment and being critical to the responses to mass shooting events. As a whole, this user falls more strongly on the side of being anti-gun control but certainly has criticism of open market gun purchasing. The neutral framing can be difficult to for automated approaches to parse since the full narrative of the users' tweets need to be taken into account.

However, coding errors are not only on the side of computational approaches. One of the key limitations to using only human-based qualitative coding is that the deep understanding of stances is highly time-consuming and delayed—especially in online space where posting moves rapidly. This means that qualitative researchers are forced to limit their focus to a subset of tweets or users, and balancing efficiency and accuracy means that there will be miscoded information that could potentially be corrected by looking at all the available data. The ability to integrate big data and computational methods can be a force multiplier for qualitative researchers as a mixed methods approach greatly improves the scope and generalizability of work on online polarization.

As mentioned previously, these stances exist on a spectrum and the middle ground can often be coded either way by a researcher making a qualitative decision on which end of spectrum the user exists. For example, is a person who owns guns for hunting but pushes for very strong hand gun control measures, pro-gun control or anti-gun control? Does wanting strong background checks or other restrictions inherently make someone pro/anti gun? Even in the context of a large body of tweets, it may not be fully clear how this stance should be coded. One example is user 66811907 who has a large body of tweets that different coders could code differently "This is your friendly reminder that New York already bans "assault weapons", that the shooter procured his NY-compliant "non-assault weapon" legally, and that the cops who stopped him showed up with their "assault weapon" version of the same gun to combat a civilian threat." and "Rifles of any kind account for fewer annual homicides than hands and feet. Mass shooters don't need pistol grips, etc., to cause horrific carnage, but on the flip side those features can be quite useful for lawful owners using them in self-defense. That's why cops carry them, too." This example is just one that highlights that neither AI nor human coding is without issues, error, or disagreement and the reporting of inter-rater reliability could be a useful measure to add for both human annotators and automated approaches.

6 Conclusion

In this work, we take a step towards a broader understanding of user stance labeling through socially infused text mining—incorporating both textual content and network interactions using language models and graph neural networks. Both gun control and climate change are issues that have a range of historical contexts, political connotations, and intersections with race and gender. It is in this space that subject-matter expertise can provide the contextual information to inform the work of computational approaches. Social science insights provide the ground truths necessary to properly contextualize computational approaches to complex human stances to ensure that the models are acknowledging the real-world context, minimizing errors, and maximizing the relevance of the generated outputs to policymakers and to the broader research community.

Acknowledgments. The research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-22-1-0035. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Addawood, A., Badawy, A., Lerman, K., Ferrara, E.: Linguistic cues to deception: Identifying political trolls on social media. In: ICWSM. vol. 13, pp. 15–25 (2019)
2. ALDayel, A., Magdy, W.: Stance detection on social media: State of the art and trends. Information Processing and Management **58**(4), 102597 (2021). https://doi.org/10.1016/j.ipm.2021.102597
3. Badawy, A., Ferrara, E., Lerman, K.: Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In: ASONAM. pp. 258–265 (2018). https://doi.org/10.1109/ASONAM.2018.8508646
4. Barberá, P., Jost, J.T., Nagler, J., Tucker, J.A., Bonneau, R.: Tweeting from left to right: Is online political communication more than an echo chamber? Psychological Science **26**(10), 1531–1542 (2015). https://doi.org/10.1177/0956797615594620
5. Coupland, C., Brown, A.D.: Constructing organizational identities on the web: A case study of royal dutch/shell. Journal of management studies **41**(8), 1325–1347 (2004)
6. Darwish, K., Stefanov, P., Aupetit, M., Nakov, P.: Unsupervised user stance detection on twitter. ICWSM **14**(1), 141–152 (May 2020). https://doi.org/10.1609/icwsm.v14i1.7286
7. Garimella, K., Morales, G.D.F., Gionis, A., Mathioudakis, M.: Quantifying controversy in social media. CoRR **abs/1507.05224** (2015)
8. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: NeurIPS. pp. 1024–1034 (2017)

9. Jiang, J., Ren, X., Ferrara, E.: Social media polarization and echo chambers in the context of covid-19: Case study. JMIRx med **2**(3), e29570 (2021)
10. Jiang, J., Ren, X., Ferrara, E.: Retweet-bert: political leaning detection using language features and information diffusion on social networks. In: ICWSM. vol. 17, pp. 459–469 (2023)
11. Johnson, K., Jin, D., Goldwasser, D.: Modeling of political discourse framing on twitter. In: ICWSM. vol. 11, pp. 556–559 (2017)
12. Karlsen, R., Steen-Johnsen, K., Wollebæk, D., Enjolras, B.: Echo chamber and trench warfare dynamics in online debates. European Journal of Communication **32**(3), 257–273 (2017). https://doi.org/10.1177/0267323117695734
13. Kipf, T.N., Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks. In: ICLR (2017)
14. Kossinets, G., Watts, D.J.: Origins of homophily in an evolving social network. American journal of sociology **115**(2), 405–450 (2009)
15. Kovacs, B., Kleinbaum, A.M.: Language-style similarity and social networks. Psychological Science **31**(2), 202–213 (2020). https://doi.org/10.1177/0956797619894557
16. Li, C., Goldwasser, D.: Encoding social information with graph convolutional networks for political perspective detection in news media. In: ACL. pp. 2594–2604 (2019)
17. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual review of sociology **27**(1), 415–444 (2001)
18. Nguyen, D.Q., Vu, T., Tuan Nguyen, A.: BERTweet: A pre-trained language model for English tweets. In: Liu, Q., Schlangen, D. (eds.) EMNLP. pp. 9–14. ACL (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.2
19. Pan, S., Wu, J., Zhu, X., Zhang, C., Wang, Y.: Tri-party deep network representation. In: IJCAI. pp. 1895–1901. AAAI (2016)
20. Peretti-Watel, P., Seror, V., Cortaredona, S., Launay, O., Raude, J., Verger, P., Fressard, L., Beck, F., Legleye, S., L'Haridon, O., Léger, D., Ward, J.K.: A future vaccination campaign against covid-19 at risk of vaccine hesitancy and politicisation. The Lancet Infectious Diseases **20**(7), 769–770 (2020)
21. R. KhudaBukhsh, A., Sarkar, R., Kamlet, M.S., Mitchell, T.: We don't speak the same language: Interpreting polarization through machine translation. AAAI **35**(17), 14893–14901 (May 2021). https://doi.org/10.1609/aaai.v35i17.17748
22. Rosenwald, G.C., Ochberg, R.L.: Storied lives: The cultural politics of self-understanding. Yale University Press (1992)
23. Tyagi, A., Uyheng, J., Carley, K.M.: Affective polarization in online climate change discourse on twitter. In: ASONAM. pp. 443–447 (2020). https://doi.org/10.1109/ASONAM49781.2020.9381419
24. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph Attention Networks. In: ICLR (2018)
25. Xiao, Z., Song, W., Xu, H., Ren, Z., Sun, Y.: Timme: Twitter ideology-detection via multi-task multi-relational embedding. In: SIGKDD. pp. 2258–2268 (2020)
26. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? In: ICLR (2019)
27. Yang, Y., Eisenstein, J.: Overcoming language variation in sentiment analysis with social attention. Trans. Assoc. for Comput. Linguist. **5**, 295–307 (2017)
28. Yarchi, M., Baden, C., Kligler-Vilenchik, N.: Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. Political Communication **38**(1-2), 98–139 (2021). https://doi.org/10.1080/10584609.2020.1785067