Towards Author-informed NLP: Mind the Social Bias

Inbar Pendzel and Einat Minkov
Faculty of Computer and Information Sciences
University of Haifa

Abstract

Social text understanding is prone to fail when opinions are conveyed implicitly or sarcastically. It is therefore desired to model users' contexts in processing the texts authored by them. In this work, we represent users within a social embedding space that was learned from the Twitter network at large-scale. Similar to word embeddings that encode lexical semantics, the network embeddings encode latent dimensions of social semantics. We perform extensive experiments on author-informed stance prediction, demonstrating improved generalization through inductive social user modeling, both within and across topics. Similar results were obtained for author-informed toxicity and incivility detection. The proposed approach may pave way to social NLP that considers user embeddings as contextual modality. However, our investigation also reveals that user stances are correlated with the personal socio-demographic traits encoded in their embeddings. Hence, author-informed NLP approaches may inadvertently model and reinforce socio-demographic and other social biases.

1 Introduction

Social natural language processing (NLP) tasks aim at understanding a given text as intended by its author. In particular, the task of stance prediction involves inferring one's attitude towards a specified target, or topic, as expressed in a given text (Mohammad et al., 2016). Typically, automated models of stance prediction consider the text as sole evidence, relying on sentiment cues, as well as on word usage and semantic frames that characterize each stance in a topic-sensitive manner, e.g., (Somasundaran and Wiebe, 2009; Reuver et al., 2021). However, the availability of relevant labeled databases is limited, and it has been shown that linguistic features often fail to generalize across targets and data distributions (Wullach et al., 2021; Ng and Carley, 2022; Waldis et al., 2024). Moreover, inferring stances from text alone is prone to fail when opinions are conveyed ambiguously, implicitly or sarcastically, or when the text lacks sufficient context; for example, the text 'I so love Biden' could be genuinely in favor of Biden, or it could be sarcastic. To that end, researchers have pointed the necessity of modeling the social and cultural context of the author, or speaker, for correctly interpreting the text as intended by them (Nguyen et al., 2021; Hovy and Yang, 2021).

In practice, various works have demonstrated the benefit of incorporating user-level information about the text authors. In particular, inferring user representations from their historical posts on social media has been shown to improve stance detection performance (Lynn et al., 2017; Zamani et al., 2018; Samih and Darwish, 2021). In addition, information about the accounts that a user links or interacts with is commonly used to model community structures and homophily among users. Typically, user embeddings are learned from dedicated social graphs constructed for this purpose, e.g., (Pan et al., 2019; Islam and Goldwasser, 2021). This transductive approach is limited in that the learned embeddings are tailored to a specific domain, task, or dataset. Consequently, data extension and retraining are required for representing new users, preventing generalization across datasets and tasks. Instead, it is desired to elicit user embeddings that encode comprehensive and generalizable social semantics.

In this work, we represent users within a social embedding space learned from the social network of Twitter at large-scale.[1] Specifically, we exploit pre-trained network embeddings of popular Twitter accounts–referred to as *entities*–which were efficiently learned from the network informa-

[1] https://x.com; The network data used to learn the social embeddings, as well as the experimental datasets, were drawn from Twitter in 2022, prior to its rebranding as X.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34825–34838
November 4-9, 2025 ©2025 Association for Computational Linguistics

tion of a large randomly selected sample of Twitter users (Lotan and Minkov, 2023). Similar to word embeddings that capture word meaning based on word co-occurrences in text, the entity embeddings represent social semantics, having popular accounts that individual Twitter users tend to co-follow reside close to each other in the embedding space. Importantly, this approach allows to compose social user embeddings in an inductive fashion. Similar to words, popular entities on social media form a general vocabulary of social meaning. Given information about the accounts that an individual user follows, she is projected onto the social embedding space based on the respective pre-trained entity embeddings.

Presumably, the encoded social network information is indicative of the user's beliefs, personal preferences and traits (Culotta et al., 2015; Xiao et al., 2020; Mueller et al., 2021; El-Kishky et al., 2022; Pendzel et al., 2024). To perform author-informed text processing, we begin by converting the text into semantic embeddings using transformer-based text encoders, which we finetune on the task at hand. These are combined with the social user embeddings, which provide complementary author context, and both sources of evidence are then processed jointly using a dedicated classifier.

We report comprehensive experiments of stance detection, showing substantial gains on multiple benchmark datasets using author-informed classification over a textual baseline. We further experiment with cross-topic stance prediction, showing improved generalization due to social user context modeling. Additional results on the tasks of hate speech and political toxicity detection further demonstrate the generality of this approach. Thus, a main contribution of our work is the validation of the proposed framework for modeling social user context in a general and inductive fashion that is applicable within and across datasets and tasks.

Another main contribution of this work is an investigation of the social features that come into play in performing author-informed stance classification. Utilizing existing models of personal trait prediction, we hypothesize and show that user stances are often correlated with various socio-demographic factors, including *gender*, *age*, *ethnicity*, *education* and *income* levels, as well as *political affiliation*. To our knowledge, this work is the first to highlight the social biases that are manifested in existing stance-labeled benchmark datasets.

In summary, our results suggest that representing users within a social embedding space that has been learned from large-scale social network data provides comprehensive and generalizable author context for decoding their intent. Provided with labeled examples, downstream classifiers can potentially leverage latent dimensions of social semantics encoded in the user embeddings to enhance performance of social NLP tasks. The learned user-level information may generalize across related topics and tasks. As personal stances and beliefs are inherently correlated with social factors, this direction shows promise–but it also carries the risk of introducing social biases in author-informed NLP.

2 Related Work

Methods that process language within the context of social factors, such as demographics or location, have led to improvements on a variety of NLP tasks. Learning age and gender-specific word embeddings improved on the tasks of topic classification and sentiment analysis (Hovy, 2015). Researchers further found that learning community-specific word embeddings from Twitter benefits sentiment analysis, due to a more subtle modeling of word meaning within inter-connected individuals (Yang and Eisenstein, 2017; Hovy and Fornaciari, 2018). To date, the adaptation of SOTA transformer-based language models to user-level factors, such as demographic information or personality traits, remains challenging (Hung et al., 2023; Soni et al., 2025; Caplan et al., 2025).

Assuming that one's viewpoint in a post is related to their identity, researchers have previously utilized information about the text author as contextual evidence for stance detection. It has been shown that modeling information about the user's network, including inter-user interactions and links, improves stance prediction performance (Lynn et al., 2017; Aldayel and Magdy, 2019a; Samih and Darwish, 2021). Others showed that learning user embeddings from social graphs serves to model homophily among users, thus promoting generalization (Darwish et al., 2020; Islam and Goldwasser, 2021; Sutter et al., 2024). A main disadvantage of the latter approach is that it involves transductive learning, generating user representations that are specialized for a given topic and limited to a given population of users.

Unlike previous works, we utilize pre-trained entity embeddings that were learned from large-scale social network data to inductively construct social user embeddings. A recent related work represented the Twitter network as a heterogeneous graph at production scale (El-Kishky et al., 2022), applying the TransE algorithm (Bordes et al., 2013) to embed users and tweets in a common low-dimension space. They proposed to represent 'out-of-vocabulary' entities, such as new users, as a mixture over the embeddings of existing users who share similar network patterns. Their evaluation focused on personalized recommendation and search in Twitter, while showing an improvement also on offensive content detection using an author-informed approach (see Sec. 6). The source data and the embeddings learned in their work are not publicly accessible, however. In our work, we utilize SocialVec (Lotan and Minkov, 2023), publicly available embeddings of more than 200K popular accounts in Twitter, that have been learned from a large sample of the Twitter network and shown to encode social semantics. Our approach is inductive in that the embeddings of individual users are constructed from the pre-trained embeddings of popular accounts which they follow.

Previous works have shown that the accounts that one follows on social networks are indicative of their socio-demographic traits, interests, and beliefs (Culotta et al., 2015; Xiao et al., 2020; Dangur et al., 2020; Mueller et al., 2021; Lotan and Minkov, 2023). Our research follows this line of works, showing that this information also provides meaningful context for inferring personal stances on disputable topics. While it is recognized that stance taking by individuals is shaped by sociocultural contexts (Du Bois, 2007), this work may be first to explicitly analyze, quantify and highlight the socio-demographic biases underlying existing stance-labeled benchmark datasets and tasks.

Considering the limited availability of stance-labeled examples, there is growing interest in zero-shot cross-target stance prediction, where pre-trained stance detection models are applied to new targets (Zhang et al., 2024). Prior research has emphasized the importance of accounting for socio-cultural similarity between target pairs (Reuver et al., 2021; Schlangen, 2021).[2] Additionally, it has been suggested that network information captures relevant social context, and that user stances across different topics may be correlated (Samih and Darwish, 2021). We report zero-shot cross-target stance detection results, showing that author-informed stance detection improves learning generalization across topics, and that the social user embeddings further enhance cross-topic generalization. As in prior work, we manually define the source and target topic pairs in our experiments. Identifying relevant topic pairs based on shared socio-demographic contexts remains an open question for future research.

3 Author-informed stance detection

The task of stance prediction from text is traditionally defined as follows. Let D be a dataset of examples, each consisting of a document (a tweet) d, which is assigned a stance label y towards a specified topic t. The task is to predict the label y given d and t. A tweet may not communicate a definite stance towards the target. The label y therefore typically takes the values of 'in favor', 'against' or 'neutral'. We frame the task of *author-informed* stance prediction in a similar fashion. Assuming that context information about the text author u is available, then it is desired to predict the stance label y towards topic t given both d and u.

Social user embeddings. While a tweet t may be processed into a semantic embedding vector using general pre-trained language models, it is an open question how to represent a user u in a low-dimension space that encodes social semantics. There exist multiple traces on social media that can be utilized to form social user representations, including the content posted, consumed, liked or shared by each user, as well as the accounts that users interact with, distribute their posts ('retweet'), the accounts which the users follow, and those which follow them (El-Kishky et al., 2022). In this work, we construct social network-based user embeddings based on the accounts that each user follows. In addition, we consider the historical texts posted by the user as content-based evidence (Volkova et al., 2015; Pritsker et al., 2017). Beyond considerations of efficiency and data accessibility, we rely on previous findings indicating the relative importance of these information types for our purposes, e.g., (Aldayel and Magdy, 2019b).

We argue that similar to text, social network information should be represented within an embedding space inferred from large-scale social network data, which encodes latent dimensions of social meaning. To that end, we employ pre-trained 100-dimension embeddings of popular user accounts on Twitter that were learned in the following fash-

[2] The terms *target* and *topic* are used inter-changeably.

ion (Lotan and Minkov, 2023).[3] Based on a large random sample of 1.5K Twitter users and the accounts they follow, the most popular accounts–those with the highest number of followers within the sample–were identified to create a 'vocabulary' of 200K *entities* of general interest. The social embeddings of these popular entities were then learned from the sampled network data, using an adaptation of the Word2Vec method (Mikolov et al., 2013). Specifically, the well-known skip-gram variant of Word2Vec is used to learn word embeddings by predicting neighboring words around a given focus word, assuming that local word sequences exhibit topical and syntactic coherence. Analogously, in learning social entity embeddings, it is assumed that the popular accounts that are co-followed by individual users form coherent social contexts, as they all reflect the preferences and interests of those users. Unlike textual data, the popular accounts that one follows form a set rather than a sequence. Accordingly, for each entity account that a user follows, embeddings were learned with the goal of predicting any of the other popular entities followed by the same user. As a result, entities that are frequently co-followed by users are positioned close to one another in the resulting embedding space. The learned embeddings have been shown to capture political biases, as well as topical semantics at varying levels of granularity. For instance, sports-related accounts may have similar embeddings, reflecting the tendency of avid sports fans to co-follow multiple accounts within that domain (Lotan and Minkov, 2023).

Importantly, given information about popular entity accounts that an individual user u follows, that user may be projected onto the social embedding space by averaging the respective entity embeddings. Presumably, the resulting representation is indicative of the user's interests and preferences. Furthermore, it has been shown that variety of socio-demographic traits, including *age*, *gender*, *ethnicity*, *education* level, and *political affiliation* can be inferred from the resulting user embeddings at high precision, using a classifier trained on labeled examples (Lotan and Minkov, 2023). In this work, we hypothesize and demonstrate that this user representation scheme effectively encodes relevant social context for stance prediction.

While alternative user embedding schemes are generally plausible, we believe that the following key requirements must be satisfied: (a) The embedding space should represent general dimensions of social meaning, inferred from large-scale data; and (b) individual user representations should be composed in an inductive manner from existing social embeddings, in order to enable transfer learning across datasets and tasks.

Contextual inference. Recent efforts to adapt pre-trained large language models to account for social factors indicate this to be an open challenge (Soni et al., 2024). One may condition text processing tasks on prior texts by the same author (Soni et al., 2022). However, network-based user representations form a social modality that is incompatible with text-based neural architectures.

Here, provided with small labeled datasets, we integrate the user representations with tweet information using linear probing, e.g., (El-Kishky et al., 2022). Initially, and as a baseline approach, we finetune a pre-trained transformer-based language encoder on each dataset and task. Author-informed stance detection is then performed by concatenating the author representation vector with the tweet embedding, generated using the finetuned encoder.

In our experiments, the user representation vector includes either the social network-based user embedding, the content-based summary embedding of the user's historical tweets, or both. Having experimented with several variants of BERT (Devlin et al., 2019) and the larger RoBERTa architecture (Liu et al., 2019), we observed comparable performance and similar trends using these models. We therefore report our results using the 12-layer BERT model.[4]

Our experiments are not aimed at achieving state-of-the-art performance, which could potentially achieved using very large pre-trained and instruction-tuned language models (Zhao et al., 2024). In fact, such models may have already been exposed to the benchmark datasets, potentially resulting in evaluation contamination. Instead, our goal is to demonstrate the potential benefits and biases associated with author-informed NLP using a controlled and well-understood experimental framework. We hope our findings will serve future work towards author-informed NLP.

4 Experimental results

In our experiments, we first wish to gauge, *to what extent, and using what types of evidence, is user-*

[3] https://github.com/nirlotan/SocialVec

[4] https://huggingface.co/bert-base-uncased

Dataset	**Target**	**Size**	**Favor [%]**	**Level**
COVID-19	vaccination	1706	68.8	tweet
SemEval	Feminism	249	47.4	tweet
	H. Clinton	264	29.5	tweet
P-stance	Biden	1657	48.0	tweet
	Sanders	1529	57.9	tweet
Yoga		1907	51.7	user
Vegetarianism		1173	48.1	user

Table 1: Dataset statistics. The tweet-level datasets are labeled as in-favor or against the target; The user-level datasets include supportive (favor) vs. random users.

	C19	Hillary	Fem.	Biden	Sanders
Tweet (t)	0.76	0.63	0.55	0.72	0.63
		User-based			
History (h)	0.67	0.68	0.64	0.74	0.66
Social (s)	0.63	0.53	0.68	**0.80**	0.71
s+h	0.70	0.70	0.64	0.74	0.70
		Author-informed			
t+s	0.78	0.72	**0.70**	0.74	0.73
t+s+h	**0.79**	**0.73**	0.66	0.78	**0.75**
Δ	4.0%	15,9%	20.7%	8.3%	19.1%

Table 2: Within-dataset Macro-F1 test results given the tweet (t), the author's tweet history (h) or social embedding (s). Δ is the relative gain using author-informed (t+s+h) vs. content-based (t) classification.

	Veg. (V)	**Yoga (Y)**	Y→V	V→Y
History (h)	0.72	0.89	0.64	0.76
Social (s)	**0.82**	0.88	**0.66**	0.76
s+h	0.80	**0.91**	0.64	**0.78**

Table 3: Vegetarianism and yoga user-level macro-F1 test results, within and across datasets.

level information predictive of their stance on a given topic? To address this question, we compare stance detection performance using textual versus author-informed evidence, using multiple user representation schemes and a variety of experimental datasets. Then, we ask, *whether, and to what extent, does user-level modeling generalize across topics?* Similar to previous works, we perform cross-topic experiments for selected topic pairs. Our results show that social author modeling can substantially improve within-dataset, as well as zero-shot cross-topic stance detection.

4.1 Datasets

We experiment with several public datasets of stance-labeled tweets. When possible, we retrieved the most recent (up to 200) tweets posted by the tweet authors, as well as the list of accounts that they followed as of June 2022. Due to the removal of some tweets and user accounts over time, we were able to recover only a subset of the original datasets. Considering the limited size of the datasets, and applying a uniform labeling scheme, we retained tweets labeled as either in favor of or against the target. Table 1 details the size and class distribution of the resulting stance-labeled datasets. *SemEval* (Mohammad et al., 2016) is a benchmark dataset collected in 2016. We consider the targets of Hillary Clinton and Feminism, having retrieved roughly 250 tweets with author information per target. Similarly, the *COVID* dataset includes tweets collected during 2020-21 that express either support for or opposition to COVID vaccinations (Poddar et al., 2022). Finally, the *P-Stance* dataset includes $\sim$1.6K tweets about Biden and Sanders as candidates for Presidency in the 2020 U.S. elections (Li et al., 2021).

We also consider a couple of datasets that directly associate users with life choices. The *Yoga* dataset includes Twitter users identified as fans of yoga based on hashtag usage (Islam and Goldwasser, 2021). We complemented this dataset with an equal number of randomly selected users as counter examples. In addition, we constructed a dataset that associates users with the life choice of *vegetarianism*. For this purpose, we sampled users from Twitter, whose self-description contained the words 'vegetarian' or 'vegan'. Also in this case, we consider random users (who do not mention the specified terms in their description) as counter examples. The number of users and class distribution of these user-level datasets are included in Table 1.

4.2 Within-topic stance detection results

We split each dataset into class-stratified train and test sets at proportions of 80:20, where we make sure that the users in the train and test sets are distinct. Table 2 details our baseline test results reported in terms of macro-F1 following finetuning using the labeled tweets (t) as sole evidence. In general, we observed that tweet classification by means of finetuning the BERT model was preferable to classification using a discrete and sparse n-gram representation of the tweets (as detailed in the appendix). The *feminism* dataset forms an exception, possibly due to the small size of the dataset. The table reports our best results for each of the datasets.

Table 2 includes also the results of predicting one's stance using user-level information as sole evidence, considering either the encoding of the user based on their historical tweets (h), their social embedding (s), or the concatenation of these

	F→H	H→F	S→B	B→S
Tweet (t)	0.45	0.41	0.67	0.63
Social (s)	**0.70**	0.62	0.64	0.50
s+t	0.57	**0.64**	**0.72**	**0.64**
Δ	26.7%	56.1%	7.5%	1.6%

Table 4: Cross-dataset stance detection macro-F1 results, given the tweet embedding (t) (The Feminist dataset is an exception, where n-grams is preferred for tweet representation); the user's social embedding (s); and, the author-informed setup. Δ is the relative gain using author-informed vs. text-only classification.

vectors (s+h). We observe that for most datasets (except COVID-19), user-level evidence yields superior stance classification results. In particular, the social user embeddings (s) or the combined user representation vector (s+h) achieve the best user-level results. Additional experimental results, using a discrete and sparse representation of the popular accounts that users follow as features, is included in the appendix.

Finally, the bottom part of the table shows the results of author-informed stance detection, using the social author embedding as context (t+s), and considering also their other posted texts (t+s+h). It is shown that author-informed stance prediction, using both feature schemes, yields the best overall results across all datasets (as marked in boldface). The improvement rate compared with text-based processing (Δ) ranges between 4-21%.

Our results on the task of identifying users who practice vegetarianism and yoga are given in Table 3. In this case, high classification performance is achieved, exceeding 0.8 and 0.9 in macro-F1, respectively. The best results are achieved using either the social user embedding as sole evidence, or its combination with the user's historical tweet information.

4.3 Stance prediction across topics

Due to the scarcity of labeled examples, it is desired to leverage existing datasets in learning to identify stances towards other topics of interest. This setup of zero-shot transfer learning is typically applied for related topics pairs, assuming that some lexical features overlap between topics that share a common latent theme (Wei and Mao, 2019; Allaway et al., 2021). We therefore experiment with target pairs that are topically related, for which the examples were collected around the same time as part of the same benchmark dataset: (i) Sanders and Biden (*P-Stance*); and, (ii) Feminism and Hillary Clinton (*SemEval*). Presumably, the latter pair is related in that Clinton is a female liberal politician, who is known for her feminist views. We perform zero-shot cross-dataset stance prediction: For each dataset pair, a model is trained using all of the labeled examples from one dataset, being applied and tested on the other dataset, and vice versa.

As shown in Table 4, learning author-informed stance prediction models improves generalization across targets in all cases. The improvement rates over the text-only baseline are particularly pronounced for the SemEval datasets (26.7-56.1%), possibly due to their small size, which makes generalization more challenging. Substantial improvements are also obtained for the P-Stance dataset, ranging between 1.6-7.5%.

We report cross-target results also for the user-labeled datasets on the topics of yoga and vegetarianism. These topics share the latent themes of commitment to health, wellness, or ethics (Islam and Goldwasser, 2021). As shown in Table 3, while there is a drop in performance in shifting targets, the models remain informative, yielding F1 performance of 0.64-0.78–well beyond an uninformative guess. Additional results show that the social user embeddings exhibit superior cross-target generalization in all cases compared with discrete network feature representations; See the appendix.

5 Analysis: What's in a stance?

Next, we explore the salient social features that contribute to stance prediction. We show that, beyond topic-specific features, there are general user-level characteristics that correlate with particular stances. Inferring and outlying users' socio-demographic profiles further illustrates pronounced social biases associated with each topic and stance.

Feature analysis. Let us examine the lexical and social patterns that characterize each stance per target in our experimental datasets. For convenience and clarity, we consider discrete features of word (1-3) ngrams, derived from the labeled tweets, and the popular Twitter accounts which users follow.

Table 5 details the most distinctive features that characterize each stance based on pointwise mutual information (PMI) analysis (Rudinger et al., 2017). As expected, some salient lexical features simply express positive (*in favor*) or negative (*against*) sentiment, e.g., 'good news', 'amazing' and 'happy', as opposed to 'lie', 'corrupt', 'problem', and 'wrong'. Other lexical features are grounded in semantic frames that are topic-specific. For example, tweets that support COVID vaccines tend to use scientific terms like 'result', 'development' and 'phase', whereas opposing tweets use the words 'poison' and 'experimental'. As another example, Tweets that support Sanders include terms like 'revolution', 'change' and 'democracy', while tweets that oppose his candidacy mention 'socialism' and 'Cuba'. As noted, such terms may not generalize well, and may even provide misleading evidence across topics, e.g., (Reuver et al., 2021).

Somewhat analogously, social features may be topic-specific. For example, in the case of COVID vaccines, authors of supportive tweets tended to follow the account of Laurene Tribe, who served as the advisor to the President on COVID-19, whereas users who opposed the vaccines followed the account of Karol Sikora, an academic who argued against the vaccines. Otherwise, prominent social features seem to be describe general correlated factors. For example, COVID vaccine supporters tended to follow the account of Greta Thunberg, while its opponents followed accounts of conservative political orientation. Likewise, supporters of Sanders tended to follow the account of Sanders himself, and other politicians of the Progressive camp. In addition, they followed the accounts of Greta Thunberg and Snowden. Those who opposed Sanders followed figures of the mainstream Democratic party like Chelsea Clinton, or Republican politicians, e.g., Ted Cruz. Presumably, this social user-level information should assist in inferring a stance as intended by the tweet author.

Considering the datasets of Vegetarianism and Yoga, we observe that Yoga practitioners often follow the account of the Dalai Lama, a spiritual leader of Tibetan Buddhism, but also follow the accounts of Google and TED talks. In contrast, random users distinctively follow accounts related to sports and rap music. Users who identify with a Vegetarian lifestyle tend to follow celebrities who are liberal and also known for practicing a plant-based diet, e.g., Al Gore and Camala Harris. Also in this case, they tend not to follow accounts of sports and rap music as opposed to random users. Therefore, user-level modeling is expected to capture some common and thus transferable social characteristics among these user populations.

Socio-demographic analysis. A variety of personal traits can be inferred with high precision based on the popular accounts that one chooses to follow on social media (Culotta et al., 2015; Mueller et al., 2021). We exploit available classifiers that were trained in a supervised fashion to predict the users' socio-demographic traits from their social network embedding vectors (Lotan and Minkov, 2023). In line with the reference human-labeled dataset (Volkova et al., 2015), the trait values are binary, pertaining to *gender*, *age*, *ethnicity*, *parenthood*, *education* and *income* levels, and *political affiliation*.

Figure 1 presents an aggregate view of the resulting user statistics per target and stance. Each trait is represented using a dedicated axis in a radar plot, where the binary trait values reside on the circle's center and diameter, and the proportion of trait values within the relevant population of users is denoted using a point along that axis. Hence, the closed shape connecting the observed proportions over all traits represents a collective social profile. The profiles of users who support vs. oppose each target are presented on top of each other, allowing to grasp the social differences between these populations. Overall, we observe notable differences for most datasets. Let us consider the dataset of vegetarianism. It is shown that among those who endorse vegetarianism, there are distinctively more women compared with random users (72 vs. 43%), a larger ratio of white ethnicity (84 vs. 75%), more liberals (96 vs. 71%), as well as higher education and income levels. Somewhat similar patterns are observed for Yoga, where those who practice Yoga include a lesser proportion of young people compared to vegetarians. Further, as expected, users who expressed support of Biden or Clinton include a higher ratio of Democrats compared with users who criticized them (91% vs. 39% for Clinton, and 81% vs. 52% for Biden, respectively). Likewise, the ratio of women among those who support them is higher. According to our analysis, those who support feminism include a higher ratio of women, voters of the Democratic party, and a higher proportion of young users compared to those who expressed negative stances towards feminism. We note that for the COVID19 dataset, there are relatively mild differences between advocates and opponents of the vaccines, with somewhat higher proportions of women, highly educated and liberal users among the supporters. Importantly, the observed distributions may reflect a sampling bias rather than real-world statistics. Nevertheless, to the extent that there exist social differences between those who hold opposite stances on disputable topics, so-

	PRO	**AGAINST**
	COVID-19	
N-grams	today, good news, result, development, phase	hell, skeptical, dna, recovery, posion, lol, experimental, flu vaccine, fear, lie, fact, cold, die
Social	White House Press Secretary, Dr. Tedros (served as Director of the World Health Org.), BuzzFeed (Internet media company, with focus on tracking viral content), Laurence Tribe (served as chief advisor to the president on COVID-19), NIH, dog_feelings, TED Talks, UNICEF, Greta Thunberg, WHO, Science Magazine	Candace Owens (conservative political commentator), Dan Bongin (conservative political commentator), White House press secretary Kayleigh McEnany, BreitbartNews, Rudy Giuliani, Eric Trump, Sean Hanity, Senator Ted Cruz, Prof. Karol Sikora (argued against COVID-19 vaccines)
	SemEval: Hillary Clinton	
N-grams	readyforhillary, human rights, amazing, mom, awesome, gay, gun control	gop, lie, black, democrat, obama, old, whilte, berniesanders, liberal, corrupt
Social	Lady Gaga, Human rights compaign, Oprah, Chelsea Clinton, Sarah Silverman, Bill Clinton, Michelle Obama, Hillary Clinton, Ellen DeGeneres, Elizabeth Warren	Fox news, VP45 (Trump), Elon Musk, MurrayNewlands (an entrepreneur), the Economist, ricky gervais, Kanye West
	SemEval: Feminism	
N-grams	feminist, woman, yesallwomen, equality, stop	sexist, feel, feminism, look, time
Social	Lasy Gaga, Taylor Swift, NYT, Rihanna, EllenDeGeneres, UN, WHO, Oprah	NONE
	P-STANCE: Bernie Sanders	
N-grams	fighting, revolution, end, medicareforall, corporate, damn, democracy, change, middleground	democraticparty, russian, cuba, communist, idiot, racist, lol, worse, destroy, killed, promise, socialism
Social	Ro Khanna, JusticeDems, Nina Turner, ProudSocialist, Rashida Tlaib, Bernie Sanders, Snowden, Greta Thunberg	Ben Shapiro, Sean Hannity, Chelsea Clinton, Ted Cruz, GOP (Rep. National Committee)
	P-STANCE: Joe Biden	
N-grams	happy, thank, community, team, united, strong, leader, senator, great, democracy	child, kid, creepy, problem, lying, corrupt, paid, hair, old, progressive, wrong, democratic party, ukranian
Social	Jill Biden, Jon Cooper, VP44 (VP Biden), John Kerry, Philip Rucker (Editor of the Washington Post), Project Lincoln, Shannon R Watts	Ben Shapiro, Sean Hannity, DonaldJ Trump Jr, Judicial Watch, kanye west, Fox News, Nina Turner (liberal)
	Yoga	
Social	Deepak Chopra (alternative medicine), Dalai Lama, UNICEF, Tony Robbins, TED talks, WHO, Google, UN, Bill Gates	Drake, Kayne West, NFL, SportsCenter, Snoop Dogg, Seth Rogen, Alyssa Milano
	Vegetarianism	
Social	Michael Woods, Al Gore, Camala Harris, Alexandria Ocasio-Cortez, Senator Cory Brooker, WholeFoods, Jerry Seinfeld	Diddy, Drake, SportsCenter, NFL, James Lebron, Rihanna, Kim Kardashian, Fox News, Kanye West, justin bieber, Nike

Table 5: The top word (1-3) n-grams and the most distinctive popular accounts followed per target and stance. All of the features that are included in the table were assigned high positive PMI scores (greater than 0.5)

cial author modeling is expected to improve stance detection performance. Likewise, author-informed prediction is expected to support generalization across socially correlated topics.

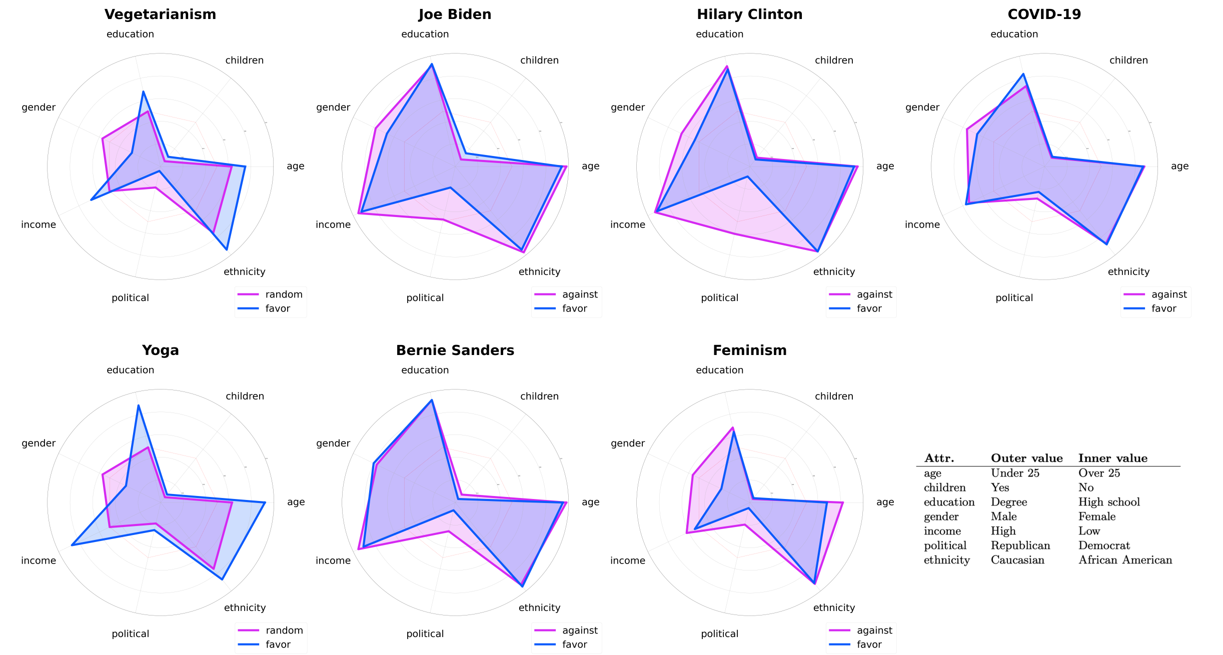
Evidence	Precision	Recall	F1
Tweet (t)	0.774	0.583	0.665
Author-informed (t+s)	**0.778**	**0.647**	**0.707**

Table 6: Political incivility detection results

6 Author-informed toxicity detection

While we focused on analyzing socio-demographic user traits from the social user embeddings, the social embedding space encodes latent fine semantics. It has previously shown that modeling user network information benefits models of hate detection, as hateful users tend to form social communities. In particular, a relative improvement of up to 9% in terms of PR-AUC has been obtained on a proprietary dataset of offensive content detection in Twitter using network user embeddings learned from Twitter at large-scale (El-Kishky et al., 2022).

We consider a pubic dataset of political tweets labeled as uncivil, where we extend a finetuned model of RoBERTa with the social embeddings of the text authors (Pendzel et al., 2024). The dataset includes 9.5K tweets, out of which 42.7% are labeled as uncivil. In applying author-informed classification, we split this dataset into 80:20 class-stratified train and test sets, assuring there was no overlap between the tweet authors among those sets. The results are detailed in Table 6. As shown,



Attr.	Outer value	Inner value
age	Under 25	Over 25
children	Yes	No
education	Degree	High school
gender	Male	Female
income	High	Low
political	Republican	Democrat
ethnicity	Caucasian	African American

Figure 1: Aggregate socio-demographic profiles of the users who support and oppose each target in our datasets. Each trait is represented using an axis in a radar plot, where the binary trait values reside on the circle's center and diameter. The proportion of trait values within the relevant population of users is denoted as a point along each axis.

incorporating the social embeddings of the tweet authors improved F1 performance by absolute 5.7 points and a relative improvement of 6.3%. Based on previous analyses (Pendzel et al., 2024), we conjecture that relevant social dimensions correspond to high political engagement, which tends to correlate with increased political incivility. In general, these results show that inductive social user embeddings learned from large-scale network data are applicable and beneficial for a variety of social text processing tasks.

7 Conclusion

Familiarity with the parties in a conversation provides important pragmatic context in human communication. Similarly, modeling information about text authors can help infer the meaning of their utterances as intended by them. So far, social user embeddings were typically learned in a transductive fashion, from small graphs constructed for specific datasets, domain and task. We argue that similar to word embeddings, which encode general lexical semantics, social user semantics should be learned from large-scale social network data. We proposed to project individual users onto a social embedding space using the pre-trained embeddings of prominent entities which they follow. Our experiments demonstrate that the resulting social user embeddings are general, comprehensive, and effective across a range of social NLP tasks. In supervised stance detection with limited labeled data, as well as in cross-topic transfer learning scenarios, we found that author-informed text processing leads to substantial and consistent improvements in model generalization. Additionally, we observed gains in toxicity detection, further illustrating the broad applicability of the proposed approach to diverse social NLP tasks.

It is an open question how to effectively inject social author context into SOTA large language models. Potentially, inductive social user embeddings could be processed jointly with textual evidence using multi-modal architectures (Jin et al., 2025). This direction may pave way to many exciting applications of social NLP, e.g., sarcasm detection, or text generation using personas that are represented as vectors in the social embedding space. Importantly, the social user embeddings capture socio-demographic attributes as well as fine-grained latent social dimensions. Our detailed analyses revealed socio-demographic biases within popular stance detection benchmarks. Social user modeling must therefore carefully consider the implications of encoding these latent social biases.

8 Limitations

While this research was conducted using datasets and user embeddings learned from Twitter, the proposed approach can be potentially applied using other social network platforms. A key requirement is that the social network data should be diverse, associating users with entities or aspects that are indicative of personal interests, tastes and beliefs. Future research may also expand the evaluation to non-English data, and other cultures. Another main limitation that is inherent to Twitter data concerns replicability, as accounts may be deleted or suspended and posts may be removed from the social network platform over time. This limitation applies to all Twitter datasets, which require tweet recovery via rehydration. We will make our datasets, including the user representations, available to researchers upon request.

As the social user embeddings are based on network information, the representation of users with richer network information is expected to be more reliable. We experimented with varying thresholds over the number of accounts that users followed. We found that information about as few as 10 (or better yet, 15) popular accounts that the user followed sufficed for achieving solid performance. It may be possible to extend the information about user-entity associations based on retweets and user mentions. Collecting this information is more demanding, however. In the context of dialogue systems, we envision future use cases where end users agree to provide access to their followees on the social network, or agree to indicate which popular accounts they favor, to achieve improved understanding of their intentions and preferences within a human-agent dialogue.

We acknowledge the potential significance of other contextual factors for social NLP, such as the context or history of a conversation, e.g., (Barel et al., 2025). Nevertheless, modeling author context is orthogonal and complementary to other types of social contexts.

9 Ethical statement

This research is intended to highlight both the potential and the risks involved in the modeling of social user information for NLP tasks. On one hand, social user modeling may lead to improved decoding of the user's utterances, thus improving personalization, as well as promoting social computation research. On the other hand, as we have shown, automatic models that incorporate social user representations are prone to modeling social biases, which may lead to stereotypical NLP. Future research should explore methods for mitigating these risks. Considering that explicit and latent personal traits may be predicted from user information and utilized for downstream application, as well as the biases that may be inflicted by social user modeling, parties who implement such an approach must inform users of these risks, request and obtain their consent.

This research has been IRB approved (031/23). The data that we experimented with has been obtained from Twitter for research purposes. Notably, the user social embeddings are practically anonymized in that users are projected onto a low-dimension embedding space, based on aggregate information about a subset of popular accounts that they follow. As user-level information is sampled, aggregated and obfuscated, and considering the dynamic nature of social networks and the immense size of the social media user base, we believe that user identity cannot be recovered. We will provide access to the user representations in our datasets to researchers who are similarly IRB authorized and adhere to the same the data usage agreement with Twitter, now X.

Acknowledgments

We thank Nir Lotan and Sagi Pendzel for their help and contribution to this research, and the reviewers for their useful comments. This research was supported by the Israeli Science Foundation, Grant no. 2671/22.

References

Abeer Aldayel and Walid Magdy. 2019a. Your Stance is Exposed! Analysing Possible Factors for Stance Detection on Social Media. *Proceedings of the ACM on Human-Computer Interaction*, 3.

Abeer Aldayel and Walid Magdy. 2019b. Your stance is exposed! analysing possible factors for stance detection on social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–20.

Emily Allaway, Malavika Srikanth, and Kathleen McKeown. 2021. Adversarial Learning for Zero-shot Stance Detection on Social Media. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Guy Barel, Oren Tsur, and Dan Vilenchik. 2025. Acquired TASTE: multimodal stance detection with textual and structural embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the International Conference on Neural Information Processing Systems*.

Eylon Caplan, Tania Chakraborty, and Dan Goldwasser. 2025. Splits! a flexible dataset for evaluating a model's demographic social inference. *arXiv preprint arXiv:2504.04640*.

Aron Culotta, Nirmal Kumar, and Jennifer Cutler. 2015. Predicting the Demographics of Twitter Users from Website Traffic Data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Ido Dangur, Ron Bekkerman, and Einat Minkov. 2020. Identification of topical subpopulations on social media. *Information Sciences*, 528:92–112.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. 14.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

John W Du Bois. 2007. The stance triangle. *Stancetaking in discourse: Subjectivity, evaluation, interaction*, 164(3):139–182.

Ahmed El-Kishky, Thomas Markovich, Serim Park, Chetan Verma, Baekjin Kim, Ramy Eskander, Yury Malkov, Frank Portman, Sofía Samaniego, Ying Xiao, and Aria Haghighi. 2022. Twhin: Embedding the twitter heterogeneous information network for personalized recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22.

Dirk Hovy. 2015. Demographic Factors Improve Classification Performance. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL*, pages 752–762.

Dirk Hovy and Tommaso Fornaciari. 2018. Increasing in-class similarity by retrofitting embeddings with demographic information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 671–677.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Chia-Chien Hung, Anne Lauscher, Dirk Hovy, Simone Paolo Ponzetto, and Goran Glavas. 2023. Can Demographic Factors Improve Text Classification? Revisiting Demographic Adaptation in the Age of Transformers. In *Findings of the Association for Computational Linguistics: EACL*.

Tunazzina Islam and Dan Goldwasser. 2021. Analysis of Twitter Users' Lifestyle Choices using Joint Embedding Model. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 242–253.

Yijie Jin, Junjie Peng, Xuanchao Lin, Haochen Yuan, Lan Wang, and Cangzhi Zheng. 2025. Multimodal transformers are hierarchical modal-wise heterogeneous graphs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nir Lotan and Einat Minkov. 2023. Social world knowledge: Modeling and applications. *Plos one*, 18(7):e0283700.

Veronica Lynn, Salvatore Giorgi, Niranjan Balasubramanian, and H. Andrew Schwartz. 2019. Tweet Classification without the Tweet: An Empirical Examination of User versus Document Attributes. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human Centered NLP with User-Factor Adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*.

Aaron Mueller, Zach Wood-Doughty, Silvio Amir, Mark Dredze, and Alicia Lynn Nobles. 2021. Demographic representation and collective storytelling in the me too twitter hashtag activism movement. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–28.

Lynnette Hui Xian Ng and Kathleen M. Carley. 2022. Is my stance the same as your stance? A cross validation study of stance detection datasets. *Information Processing and Management*, 59(6).

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in nlp: a sociolinguistic perspective. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jiaqi Pan, Rishabh Bhardwaj, Wei Lu, Hai Leong Chieu, Xinghao Pan, and Ni Yi Puay. 2019. Twitter homophily: Network based prediction of user's occupation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*.

Sagi Pendzel, Nir Lotan, Alon Zoizner, and Einat Minkov. 2024. A closer look at multidimensional online political incivility. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Soham Poddar, Mainack Mondal, Janardan Misra, Niloy Ganguly, and Saptarshi Ghosh. 2022. Winds of Change: Impact of COVID-19 on Vaccine-related Opinions of Twitter Users. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 782–793.

Evgenia Wasserman Pritsker, Tsvi Kuflik, and Einat Minkov. 2017. Assessing the contribution of twitter's textual information to graph-based recommendation. In *Proceedings of the International Conference on Intelligent User Interfaces*.

Myrthe Reuver, Suzan Verberne, Roser Morante, and Antske Fokkens. 2021. Is stance detection topic-independent and cross-topic generalizable? - a reproduction study. In *Proceedings of the 8th Workshop on Argument Mining*, pages 46–56, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79.

Younes Samih and Kareem Darwish. 2021. A Few Topical Tweets are Enough for Effective User Stance Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2637–2646, Online. Association for Computational Linguistics.

David Schlangen. 2021. Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing Stances in Online Debates. In *Proceedings of the Joint Conference of the Annual Meeting of the ACL and the International Joint Conference on Natural Language Processing of the AFNLP*.

Nikita Soni, Pranav Chitale, Khushboo Singh, Niranjan Balasubramanian, and H. Schwartz. 2025. Evaluation of LLMs-based hidden states as author representations for psychological human-centered NLP tasks. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H. Andrew Schwartz. 2022. Human language modeling. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Nikita Soni, H. Andrew Schwartz, João Sedoc, and Niranjan Balasubramanian. 2024. Large human language models: A need and the challenges. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Maia Sutter, Antoine Gourru, Amine Trabelsi, and Christine Largeron. 2024. Unsupervised stance detection for social media discussions: A generic baseline. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Andreas Waldis, Yufang Hou, and Iryna Gurevych. 2024. Dive into the chasm: Probing the gap between in- and cross-topic generalization. In *Findings of the Association for Computational Linguistics: EACL 2024*.

Penghui Wei and Wenji Mao. 2019. Modeling transferable topics for cross-target stance detection. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Zhiping Xiao, Weiping Song, Haoyan Xu, Zhicheng Ren, and Yizhou Sun. 2020. Timme: Twitter ideology-detection via multi-task multi-relational embedding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Yi Yang and Jacob Eisenstein. 2017. Overcoming Language Variation in Sentiment Analysis with Social

Attention. *Transactions of the Association for Computational Linguistics*, 5:295–307.

Mohammadzaman Zamani, H. Andrew Schwartz, Veronica Lynn, Salvatore Giorgi, and Niranjan Balasubramanian. 2018. Residualized Factor Adaptation for Community Social Media Prediction Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Zhao Zhang, Yiming Li, Jin Zhang, and Hui Xu. 2024. LLM-driven knowledge injection advances zero-shot and cross-target stance detection. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*.

Chenye Zhao, Yingjie Li, Cornelia Caragea, and Yue Zhang. 2024. ZeroStance: Leveraging ChatGPT for open-domain stance detection via dataset generation. In *Findings of the Association for Computational Linguistics: ACL 2024*.

A Experimental setup

In fine-tuning the models per dataset, we used a batch size of 32 examples and early stopping on a held-out validation set containing 10% of the training data, using the AdamW optimizer with a learning rate of 3×10-5. Following tuning, we learned the final models using all of the training examples.

B Dense vs. sparse network-based user representations

In this paper, we advocate the representation of user in a low-dimension embedding space, learned from a large sample of the Twitter network. The users are projected onto the social embedding space using the pre-trained embeddings of popular entity accounts followed by them (Lotan and Minkov, 2023). Alternatively, information about the accounts that one follows can be represented using discrete and sparse feature vectors. We assessed stance detection performance using the low-dimension representations against this baseline approach, implemented as follows.

Following previous work Lynn et al. (2019), we identified the most popular 5K accounts followed by the users within the training portion of each dataset. Individual users were then represented using a one-hot vector of length 5K. Each feature in this vector denote a specific Twitter account, indicating if the user follows that account or not.

The experimental results using this discrete representation scheme ('Top 5K') are reported in Tables 7 and 8 for each of our experimental datasets. For clarity, we focus on network information as single evidence. For comparison, the tables include our results using the social user embeddings (s). We also report the results of stance detection using the concatenation of both user representations (s+5K). Overall, the performance of the dense and sparse representation schemes are comparable, where the concatenated vector yields the best results in most cases.

The advantage of the social embeddings becomes apparent in the cross-topic stance detection setup. The results of the cross-dataset experiments using each of the user network representations are given in Tables 8 and 9. There, we observe superior performance using the social user embeddings in all cases, where the user embedding information generalize gracefully to related topics, while discrete user modeling fails to generalize.

	C19	Hillary	Fem.	Biden	Sanders
Top 5k	0.62	**0.68**	0.68	0.81	**0.76**
Social (s)	0.63	0.53	0.68	0.80	0.71
s+5k	**0.64**	0.64	**0.72**	**0.82**	**0.76**

Table 7: Within-dataset stance detection macro-F1 test results using alternative representations of user network information: discrete ('Top 5K'), social embeddings (s), and their concatenation.

	Veg. (V)	**Yoga (Y)**	Y→V	V→Y
Top 5k	0.76	0.86	0.54	0.50
Social (s)	**0.82**	**0.88**	**0.66**	0.76
s+5k	**0.82**	**0.88**	0.65	**0.78**

Table 8: Yoga and Vegetarianism: Within- and cross-dataset stance detection macro-F1 test results: Dense vs. sparse user network representations

	Fe→Hi	Hi→Fe	Sa→Bi	Bi→Sa
Top 5k	0.64	0.55	0.50	0.34
Social (s)	**0.66**	**0.62**	**0.66**	**0.50**
s+5k	**0.66**	0.55	0.51	0.36

Table 9: Cross-dataset stance classification results: Dense vs. sparse user network representations

C Sensitivity to train set size

Considering author context information is beneficial whenever the text lacks concrete evidence, or in case of ambiguity. In general, however, content-based stance classification is expected to improve as the number of relevant labeled tweets for the specified topic and target increases. We conducted additional experiments to explore the impact of

	50	100	200	500	1000	All
Tweet (t)	0.54	0.54	0.54	0.75	0.77	0.78
User-based:						
History (u)	**0.63**	**0.69**	0.71	0.71	0.75	0.73
Social (s)	0.55	0.54	0.57	0.71	0.70	0.70
s+h	**0.63**	**0.69**	0.71	0.72	0.74	0.74
Author-informed:						
t+s	0.61	0.63	**0.73**	**0.76**	**0.80**	0.81
t+s+h	0.59	0.67	0.72	**0.76**	0.79	**0.82**

Table 10: Weighted F1 results on COVID-19 test set using increasing train set size, starting from 50 examples and up to the full train set

	50	100	200	500	1000	All
Tweet (t)	0.49	0.46	0.66	0.71	0.72	0.73
User-based:						
History (h)	0.65	0.65	0.70	0.75	0.78	0.74
Social (s)	0.52	0.63	0.70	0.78	0.78	**0.79**
s+h	0.71	0.70	**0.74**	**0.79**	**0.79**	0.75
Author informed:						
t+s	0.66	0.68	0.72	0.78	0.75	0.75
t+s+h	**0.72**	**0.74**	0.72	0.76	**0.79**	0.78

Table 11: Weighted F1 results on Joe Biden test set using increasing train set size, starting from 50 examples and up to the full train set

training set size on test performance using the various tweet and author representations. We considered the COVID-19 and Biden datasets for this purpose, as both of these datasets contain about 1.7K labeled tweets. In the experiment, we progressively increased the size of the training set, learning a model using class-stratified subsets of 50, 100, 200, 500, and 1K labeled examples, as well as the full train set. To ensure consistency in learning evaluation, each subset subsumes the labeled examples includes in the smaller subsets. The test set remains unchanged however to enable a direct comparison of performance across different training set sizes and evidence representation schemes.

Tables 10 and 11 show the test set results in terms of weighted F1 for the COVID-19 and Biden datasets, respectively. Macro F1 results showed very similar trends, and are omitted for brevity. The top-performing methods for each experiment are marked in boldface. In general, we observe improvements in performance as the train set increases. Tweet embeddings (t) yield lower performance compared with the discrete and sparse n-grams representation (omitted from the table for clarity). Modeling user-based information presents superior performance given as few as 50 training examples. Thereafter, the combination of user-based and tweet information achieves best results (COVID-19), or comparable results to user-based prediction (Biden). In summary, user-based information is preferable when the number of labeled examples for a given target is small. And, it consistently enhances performance as more examples become available.

D Example tweets

Presumably, modeling user context information can resolve text ambiguity due to possible sarcasm or insufficient context. For example, a text such as 'I so love Biden' may be interpreted as either sincere and positive or sarcastic and negative, depending on the user's political affiliation. Following are randomly selected examples of tweets from the experimental datasets, which were misclassified using the tweet embeddings as sole evidence (t), while being correctly classified using the author-informed model ($t + s$). The first example (COVID) may require relevant world knowledge about MRHA approval, as well as non-trivial reasoning, to infer a positive stance towards the vaccine. Considering additional information about the user can provide additional clues about their stance. The last example tweet (Feminism) is implicit and sarcastic, yet the user features may suggest that the underlying stance is negative.

- @user In fact, the third vaccine the moderna one is also MRHA approved. (COVID-19; Favor)

- Maybe you should let JoeBiden know it wasn't the food service staff either. (Joe Biden; Against)

- I keep seeing Brihana Joy Gray pop into my feed. Can we just all agree to ignore her and let her fade into obscurity? #justdontlook #byefelicia #Bernie (Bernie Sanders; Favor)

- Hillary flew on private jet to speak to speak about Income & Equality. Dinner plates were $2700.00 ! #EqualityForAll #PJNET #CCOT (Hillary Clinton; Against)

- Husband: "They should know that she's always born with it. It's never Maybelline." #beauty (Feminism; Against)