PLOS Complex Systems

RESEARCH ARTICLE

Weakly supervised contrastive representation learning to encode narrative viewpoint of COVID-19 tweets

Kin Wai Ng[1,2], **Nathan Wendt**[1], **Jasmine Eshun**[1], **Emily Saldanha**[1]*

1 Pacific Northwest National Laboratory, Richland, Washington, United States of America, **2** Department of Computer Science and Engineering, University of South Florida, Tampa, Florida, United States of America

* emily.saldanha@pnnl.gov

OPEN ACCESS

Citation: Ng KW, Wendt N, Eshun J, Saldanha E (2026) Weakly supervised contrastive representation learning to encode narrative viewpoint of COVID-19 tweets. PLOS Complex Syst 3(2): e0000089. https://doi.org/10.1371/journal.pcsy.0000089

Editor: Hocine Cherifi, Université de Bourgogne: Universite de Bourgogne, FRANCE

Received: April 07, 2025

Accepted: January 8, 2026

Published: February 13, 2026

Data availability statement: Data supporting this study are derived from the publicly available Avax dataset (https://github.com/gmuric/Avax-tweets-dataset). Code for data processing and modeling is available here: https://github.com/pnnl/NARREMB.

Abstract

Viewpoint detection is a crucial step for characterizing online narratives and understanding their diffusion and evolution across online social spaces. Existing methods for semantic understanding of online posts have shown strong progress towards grouping documents with similar topics but struggle to differentiate between different viewpoints towards those topics. The purpose of this work is to infuse semantic embedding spaces with improved viewpoint information under the constraint of low data availability. To address this task, we develop a novel weakly supervised contrastive learning approach that leverages social proximity of users as a self-supervised signal of shared viewpoint likelihood. We demonstrate the utility of this method on a use case of X (formerly Twitter) discussion related to COVID-19. We show that fine-tuned embeddings which were trained to predict social proximity signals present in retweet networks demonstrate the capability to infuse learned embeddings with viewpoint information. Finally, we demonstrate that these viewpoint-infused embeddings show improved effectiveness at identifying clusters of tweets with shared viewpoints and topics when used in a topic modeling pipeline. Such viewpoint-infused embeddings have strong potential to support multiple semantic reasoning tasks including topic modeling, stance detection, and narrative detection.

Author summary

Information that spreads on social media can impact public opinion and real-world outcomes such as health behaviors. To understand these effects, researchers need tools to analyze large volumes of text and detect coherent narratives from disparate sets of posts. One prominent method for understanding such patterns is to represent posts in a mathematical embedding space, where posts with similar meanings appear close together. While existing methods succeed at grouping

Funding: The research described in this paper was conducted under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy under contract number DE-AC05-76RL01830. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

social media text by topic, they often fail to distinguish posts that have opposing viewpoints on the same topic. In this paper, we develop an approach to improve embedding spaces for social media text data by using information from the social interactions of users on the social media platform. By leveraging the fact that users who interact with each other are more likely to agree, this information helps distinguish conflicting views. Using a case study of social media discussion related to COVID-19 vaccination, we demonstrate that our novel approach leads to an improved ability to separate opposing viewpoints, providing a strong basis for narrative discovery within large social media datasets.

Introduction

The growing prevalence and ubiquity of social media usage has gone hand in hand with rising concerns about the spread of misinformation on these social platforms. The challenges of misinformation can be particularly profound in areas such as public health, where online information can drive real-world behaviors. The ability to observe and characterize the narratives that are spreading on social media can increase the awareness needed to mitigate and prepare for the potential adverse consequences of misinformation spread.

Viewpoint detection is a crucial step for characterizing these online narratives and understanding their diffusion and evolution in the online space. The detection of shared narrative viewpoint in social media text has many inherent challenges. Such posts are typically short, noisy, and written using unique styles and lexicons [5]. Further, online posts usually provide little context or only a small portion of the viewpoint they promote.

Existing methods have shown remarkable progress in identifying collections of documents sharing similar topics via learned semantic embeddings [2,9]. However, such approaches often struggle to differentiate between different points of view related to the same topic, which is necessary for analytical studies aimed at understanding public opinion, especially on controversial issues with prevalent misinformation such as the COVID-19 pandemic. Another challenge is related to the lack of availability of labeled data for training, which can be attributed to the high costs of the data annotation process and the large scale of social media data.

The purpose of our work is to address these challenges and develop an embedding approach which can encode both viewpoint and topic information based on only the text of a social media post. To that end, we develop weakly supervised contrastive methods that leverage self-supervised signals present in social media data to learn a latent space representation of tweet content that is organized by shared narrative viewpoint. The use of self-supervised signals also has the benefit of avoiding reliance on manually annotated data, which would not allow easy generalization to new topic areas or emerging events within the domain of interest. Motivated by the observed homophily of online social networks [20], we leverage social proximity as the weak proxy signal for our contrastive methods and train the model to infer the social network proximity of the authors of a given pair of tweets based only on

the text of those tweets. Our hypothesis is that a model trained on this proxy task will learn to group together texts with similar points of view. In contrast to methods which are designed for specific tasks such as stance detection or narrative discovery, we aim to develop a novel general purpose embedding method which can infuse user viewpoint information into embeddings of social media text. By ensuring that they encode both topic and viewpoint information, these improved embeddings can then support improved performance on many downstream social media analytics tasks. To the best of our knowledge, our work is the first to use social network information as a weak signal to improve embeddings generated from text alone.

In this paper, we demonstrate our methodology on X discussions related to COVID-19. Through multiple experiments, we show that fine-tuned embeddings leveraging social proximity signals present in retweet networks demonstrate the capability to infuse learned embeddings with viewpoint information. We find that domain-specific embeddings are crucial for improving the encoding of viewpoint information compared to out-of-domain pre-trained embeddings. We show that, while viewpoint encoding tends to improve when trained with this approach, it comes with a trade-off in the topic encoding performance. To address this, we propose an approach that incorporates topic-aware representations, which maintain the original topic information and augment the existing representations with learnable viewpoint-encoding dimensions. Our results demonstrate that improvement on the narrative viewpoint task correlates with performance on the proxy task of weak social signals. This observation highlights the usefulness of weakly supervised approaches for this task and calls for future research into the incorporation of additional signals that can further improve the embeddings. Lastly, we apply a topic modeling technique using our fine-tuned embeddings to demonstrate their effectiveness in clustering tweets with shared viewpoints compared to baseline pre-trained embeddings. The code to replicate our methods can be found here: https://github.com/pnnl/NARREMB.

This work extends the analysis performed in Ng et al. [39]. We build on our prior efforts in several ways. First, we provide more comprehensive data and methodology descriptions, perform more extensive experiments to quantify performance across different hyperparameters, and provide quantitative and qualitative performance evaluation at the individual topic level in addition to the aggregate level. Finally, we demonstrate the application of the approach for cluster-based topic modeling analysis, showing that the improvement in viewpoint encoding can support improved detection of cohesive single-viewpoint, single-topic clusters. This key analysis step demonstrates the utility of our improved text embeddings for downstream tasks such as narrative discovery or stance detection.

Related work

The target task of narrative viewpoint detection is related but not identical to several other tasks designed to group and categorize text such as topic detection or stance detection. Viewpoint refers to the understanding of subjective perspectives or interpretations of a topic, reflecting how users perceive the topic. We employ the term viewpoint to convey a more general concept than specific tasks such as topic, stance, or sentiment detection. Topic detection involves the detection of groups of documents related to the same subject area. Methods for topic detection typically do not differentiate between different points of view related to the same topic. For example, discussion in support of mask wearing and discussion opposed to mask wearing will likely be grouped into the same topic unless the vocabulary used by the two groups becomes highly disjoint. Meanwhile, the task of stance detection typically focuses on the detection of different opinion categories toward a given specific entity or viewpoint. Our task differs from stance detection, as our objective is not to categorize tweets into predefined stance labels. Instead, we aim to implicitly encode viewpoint information into learned tweet embeddings such that tweets which share both a topic and a viewpoint are clustered together. Embeddings with improved encoding of viewpoint can be used for many downstream tasks compared to those that encode only topics. Here we highlight prior work in related areas and modeling techniques.

Related tasks

Topic modeling has been widely applied as an unsupervised method to discover the underlying organization of a corpus into topical groups [2,6]. For example, Gao et al. leverage topic modeling on social media and news data for event summarization [16], while Nerghes and Lee compare topics discovered from news and from social media to perform narrative comparison analysis [38]. Such topical groupings are typically not sensitive to the fine-grained viewpoints of the author toward the subject.

Stance detection focuses on the extraction of an author's stance towards a given target from a fixed set of options (e.g. Favor or Against) [27]. For example, Dey et al. leverage a two-stage LSTM-based model to first classify the overall subjectivity of tweets, followed by a stance detection step [12]. While these methods typically require pre-defined target and stance categories, recent work has tackled the unsupervised stance detection task. Darwish et al. leverage unsupervised clustering of user embeddings to identify core users to represent each stance [11].

Narrative detection approaches seek to represent online messages in a narrative framework. Jachim et al. develop a method called TrollHunter which leverages correspondence analysis to identify nouns and verbs which are commonly used together in a corpus of tweets [21]. Kwak et al. develop FrameAxis which detects semantic dimensions in the word embedding space of the documents which correspond to different framings of a given issue [28]. Tangherlini et al. construct a graph-based representation of the actants and their interactions as described in a set of documents and performs community detection on this graph to extract sub-graphs corresponding to narrative frameworks [47].

Modeling methods

Weak supervision is a model training framework which leverages large datasets of automatically generated noisy labels in lieu of high quality annotations which are time and effort intensive to generate [42]. Weakly supervised approaches have been previously applied within NLP problems ranging from text and document classification [33,34] to information extraction such as named entity recognition [29,30]. In this work, we leverage the social proximity of users observed through interaction networks as our weak signal to infer the similarity of tweets and their viewpoints towards a particular topic. Other recent work has leveraged a combination of textual signals and social network information to infer user stance or viewpoint information [23,46,52,53]. However, while these methods require both tweet content and social interaction information to infer the stance of a new tweet, our weak supervision approach is able to infer tweet viewpoint using tweet content alone.

Contrastive representation learning is often applied to different views of the same data instance. For example, in computer vision, the goal of the contrastive learning framework is to map similar images (e.g., original and augmented) close together while pushing dissimilar ones further away in the embedding space. The adoption of this framework has led to state-of-the-art performance on unsupervised image classification tasks [8,25]. The same idea has also been applied to natural language text, especially in tasks involving text retrieval [48], text generation [1], and sentence embedding representations [15], often resulting in significant performance improvements over multiple baselines. Recent work has leveraged a contrastive training approach for stance detection by incorporating counterfactual data augmentation of matched and unmatched stances [26].

COVID-19 narrative analysis

Many studies have applied topic modeling to understand the topics on online discussion related to COVID-19 [4,7,13,22,31,49,51]. Jing and Ahn leverage FrameAxis to characterize differences in the framing of COVID-19 narratives between different political parties [24]. Shahsavari et al. leverage the narrative framework discovery pipeline developed in Tangherlini et al. [47] to study COVID discussion [45]. In several works Ng et al. leverage sentiment analysis combined with topic modeling to identify prominent topics and themes associated with negative sentiment towards both influenza and COVID-19 vaccination [40,41]. These prior frameworks often leverage contextual embeddings to group texts with

similar semantics, creating candidate clusters for narrative extraction. Our proposed approach aims to enhance the embeddings of tweets by incorporating viewpoint information. The resulting embeddings can help identify more meaningful and fine-grained tweet clusters, and thus improve the characterization of narratives.

Data

We leverage *Avax* [37], a dataset on COVID-19 X (formerly Twitter) discussions, especially about vaccine adoption and hesitancy. It is publicly available (https://github.com/gmuric/Avax-tweets-dataset) and contains 1.2M records posted by 568K users from Oct. 18th, 2020 to April 21st, 2021 that were collected using a set of anti-vaccine keywords. Due to the focus of the dataset used in this case study, we expect to discover narrative patterns related to anti-vaccine oriented narratives rather than COVID-related narratives more broadly. Additionally, due to the timing of the data collection, the observed narratives will be impacted by the events which occurred during this time frame, including vaccine trial results and rollout, new variant emergence, and changing COVID-related policies and regulations. The use of online user-generated content for research purposes carries risks related to user privacy. All information used in this study was publicly available at the time of collection. The *Avax* data was collected and used in accordance with its license and data usage agreement. We do not release or reveal any social media text content or user information as part of this work.

Tweets processing and topic annotation

We preprocessed the tweets as follows: (1) we removed duplicate entries, such as multiple collections of the same tweet ID, (2) we ignored retweets and only considered actions with textual content written by the user (i.e., tweets, replies, and quotes); for simplicity, we refer to all these actions as tweets, (3) we removed mentions, URLs, and special characters from the tweets to focus on the natural language content, we keep only tweets that are assigned a topic by Top2Vec [2], due to its use in our data sampling pipeline, (5) we kept only tweets in English, and (6) we performed a fuzzy text deduplication technique to remove repeated tweets with almost the exact same content by stripping the text of all non-alphanumeric characters and dropping duplicates. We also removed tweets containing less than 20 characters as they constitute less than 1% of the total tweets in the dataset and often lack textual signals necessary for identifying viewpoints. For example, many of these tweets contained minimal natural language text, often including only a few words, emojis, or punctuation marks, which is not useful for our analysis. After preprocessing the tweets, we end up with 362,146 unique tweets from 191,823 users. The preprocessing steps, along with corresponding data amounts, are illustrated in Fig 1.

We employ Top2Vec [2] to assign a topic to each tweet in the dataset using pre-trained embeddings from the universal-sentence-encoder model. For clustering parameters, we use *min_cluster_size* = 300 and *cluster_selection_method* = "*leaf*". This results in 67 topics with sizes ranging from 861 to 26,585 tweets. We further explore these broad Top2Vec topics and identify ten higher-level categories: *Conspiracy*, *Policy*, *Science*, *General Anti-Vaccination*, *Personal*, *Politics*, *Rights*, *General*, *Memes*, and *Negative Vaccine Effects*.

Since the *Avax* dataset does not include ground truth for viewpoint detection, we manually annotate binary user viewpoints for a selected subset of topics to create a reference dataset for model evaluation. The resulting set of manually annotated tweets facilitates the evaluation of various embedding approaches, specifically in their ability to capture both the semantic content and viewpoint of tweets. We focus on topics that are likely associated with controversial issues, identified through manual inspection of Top2Vec-generated keywords.

We found six controversial topics that were prevalent in our collection of tweets: *Face masks*, *Re-opening schools*, *Vaccine passports*, *Depopulation agenda*, *Miscarriage agenda*, and *Vaccines*. The prevalence of these topics was influenced both by the keywords used for collection of the *Avax* data and by the time period of the data collection when these were salient discussion areas. For each of these six topics, one of the authors randomly selected a tweet and was instructed to annotate it with a binary viewpoint towards that topic (either support/oppose or believe/refute). If the tweet did not express

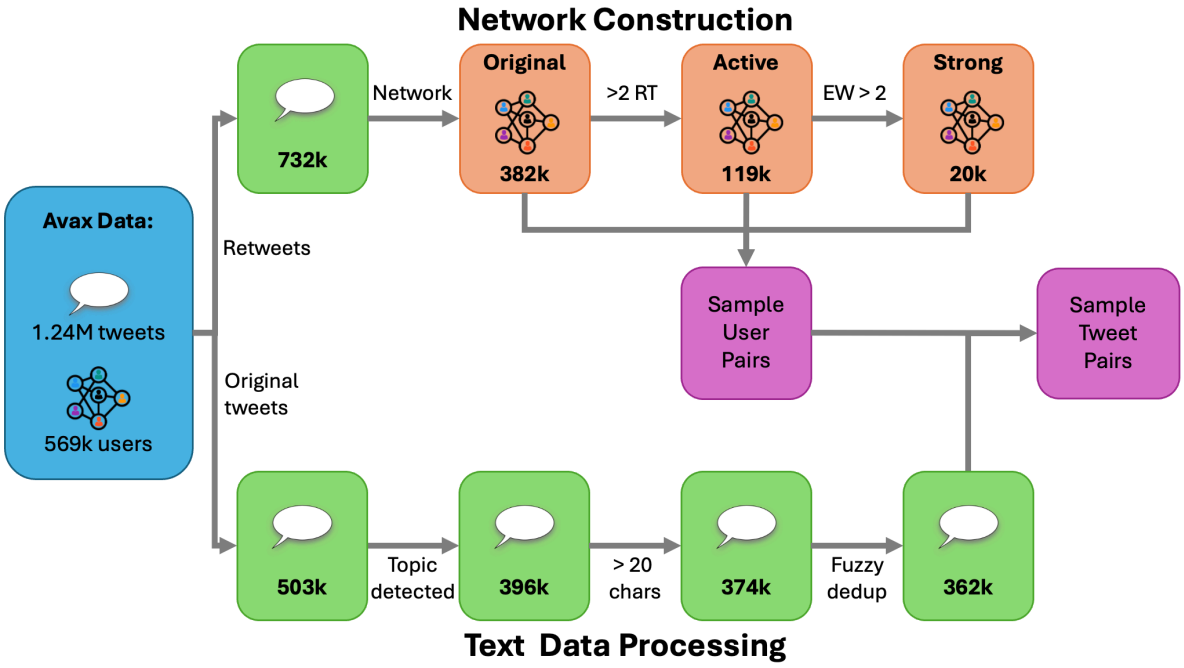


Fig 1. Diagram illustrating the data processing approach to extract corresponding user and tweet pairs for training the contrastive model. The number of tweets (speech bubble icons) and users (network icons) remaining after each processing step are specified.

https://doi.org/10.1371/journal.pcsy.0000089.g001

a clear viewpoint, it was not annotated. The annotation process concluded once we had approximately 30 tweets per viewpoint for each of the six topics. Another author of the paper reviewed the annotations to ensure that each tweet was properly labeled.

Table 1 presents the selected topics, their corresponding viewpoints, and the relevant statistics from the annotated set. We use these 365 annotated tweets as a test set to validate how well the trained models encode shared user viewpoints.

Constructing social networks from Avax tweets

We retrieve tweet information using the *twarc* Python library to rehydrate the tweet IDs. Each record in the dataset includes the following fields: the unique identifier of the retweet, the user identifier of the retweet's author, the timestamp of the retweet, and the original tweet ID along with its author's ID. The X (Twitter) API does not provide complete data on retweet cascades, making it challenging to reconstruct full retweet chains, including intermediate connections between retweets. In this work, we assume that each retweet directly references the original tweet.

Table 1. Identified set of controversial topics and viewpoints for evaluation. The number of tweets for each topic is in parentheses. VP = Viewpoint, S = Support, O = Oppose, B = Believe, R = Refute.

Topic	Category	Discussion Description	VP 1	VP 2
Face masks	Policy	The use of face masks	S (36)	O (30)
Re-open schools	Policy	Reopen schools for in-person learning	S (27)	O (30)
Vaccine passports	Policy	Adoption of vaccine passports	S (31)	O (27)
Depopulation agenda	Conspiracy	Vaccines as part of a depopulation plan	B (29)	R (33)
Miscarriage agenda	Conspiracy	Miscarriages following vaccinations	B (30)	R (30)
Vaccines	General	General opinions of the vaccines	S (33)	O (29)

https://doi.org/10.1371/journal.pcsy.0000089.t001

We construct the social network by leveraging user retweet interactions. We focus on the retweet network because these interactions are most likely to indicate endorsement of a shared viewpoint between users, as opposed to other actions such as replies or mentions, where disagreement may be expressed [14,35,44]. We expect that users who exhibit close proximity in the network would likely share similar viewpoints, while users who are distant in the retweet network are more likely to hold differing views. By leveraging social network information as weak signals, we can improve the representation of tweets, as the implicit connections between users can offer valuable insights into user perspectives and viewpoints.

To effectively capture the diverse nature of social connections and understand the impact of noise and outliers, we construct three distinct retweet networks. Each network is designed to measure the strength of social connections in different ways, allowing for a more nuanced analysis of user interactions and their influence on viewpoint detection. We represent the retweet network as a directed graph, $G = (V, E)$, where nodes $v \in V$ represent user accounts, and an edge $e \in E$ refers to the number of times a social media user retweeted another user. The different types of retweet networks constructed are as follows: (1) *Original* network, which includes all retweet interactions without filtering; (2) *Active* network, which includes users who have made at least two retweets over the entire observation period; and (3) *Strong* network, which includes only edges with an edge weight of at least 2 (i.e., each user must have retweeted another user at least twice). The Active network is a subset of the Original network, and the Strong network is a subset of the Active network. This network creation workflow is illustrated in Fig 1.

Table 2 shows the summary statistics for the giant connected component of each network. Our analysis focuses on the interactions within the giant connected component in each retweet network, as these comprise the majority of users (94%, 98.6%, and 81.3% in the original, active, and strong networks, respectively). We chose to disregard disconnected components because of their limited occurrence and the possibility that they could introduce noisy examples not directly relevant to discussions about COVID-19. We observed that all networks exhibit a high level of modularity, which indicates a strong division of the network into communities. Additionally, the filtered versions of the networks are significantly smaller, specifically by 69% for the active network and 95% for the strong network when compared to the original. Similarly, the number of engaged users (i.e., those users who retweet and also post original tweets) decreases consistently. While we expect both the active and strong networks to be less noisy, there is a trade-off in the number of examples we can generate from the smaller pool of candidate users.

Methodology

To motivate the need for our proposed approach, we first demonstrate that baseline sentence embedding methods struggle to encode viewpoints. In Fig 2a we show a 2D UMAP projection [32] of our set of tweets with annotated binary viewpoints, as shown in Table 1. We extract embeddings for each tweet using a standard pre-trained sentence embedding model from the SentenceTransformers library (https://huggingface.co/sentence-transformers/all-distilroberta-v1) [43]. We use dark versus light colors to represent different viewpoints on the same topic. We observe that the tweets are strongly clustered by topic but achieve no separation by viewpoint. This observation highlights the importance of infusing the

Table 2. Social network properties for the giant connected component in each of the three retweet networks in this study. The Engaged Users property denotes the number of users who participated in both retweeting and generating original posts in *Avax*.

Network Properties	Original	Active	Strong
Users	382,044	118,858	19,638
Edges	646,596	378,820	27,307
Modularity	0.77	0.67	0.83
Mean Degree	3.38	6.37	2.78
Mean Edge Weight	1.09	1.15	2.90
Engaged Users	57,193	41,999	7,488

https://doi.org/10.1371/journal.pcsy.0000089.t002

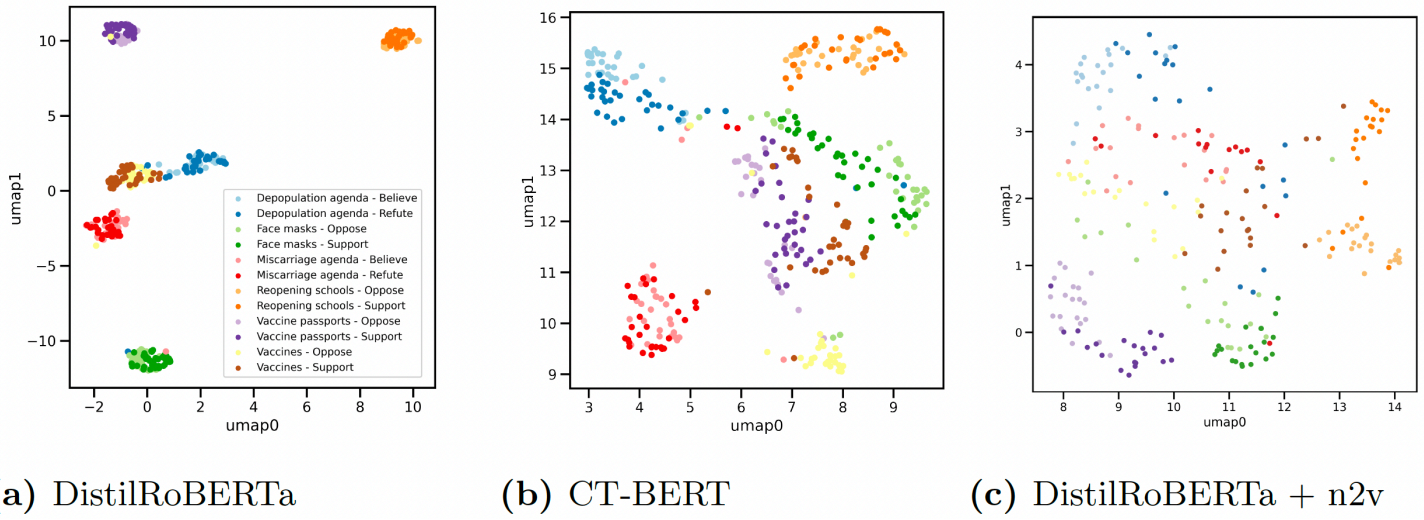


Fig 2. UMAP latent space visualization of two baseline embedding methods (a and b) and a concatenated embedding combining DistilRoBERTa with node2vec embeddings (c) for a selected set of topics with manually annotated binary viewpoints. (a) DistilRoBERTa. (b) CT-BERT. (c) DistilRoBERTa + n2v.

https://doi.org/10.1371/journal.pcsy.0000089.g002

embedding space with viewpoint information such that tweets are organized by both topic and viewpoint. To achieve this, we propose a weakly supervised contrastive approach that leverages social signals to enhance viewpoint representation.

Self-supervised social signals

To demonstrate the utility of social signals for viewpoint encoding, we first generate Node2Vec [19] embeddings for each user in the *Avax Original* retweet network. We concatenate the baseline DistilRoBERTa embeddings for tweets with the corresponding Node2Vec embeddings of their authors, incorporating network position information to augment the initial content embeddings. The 2D UMAP projections of this concatenated embedding space are shown in Fig 2c. We observe that this approach is able to partially separate the annotated test set by both topic and user viewpoint, in contrast with the sentence embeddings alone, showing the utility of social information. However, this approach relies on the availability of a complete view of the retweet network. In the *Avax* dataset, only 53% of tweets were authored by users who have interacted with other users via retweet interactions. Hence, we cannot utilize the Node2Vec representations to infer the stance or viewpoint of the users who authored the remaining 47% of the tweets. This problem is further exacerbated in less comprehensive data collections that have more limited interaction information.

Our weak supervision approach is designed to transfer the weak signals from tweets where social network information is available to those where they are lacking. For example, we can identify the textual signals that make a pair of tweets appear as if their authors were closely related or distant in the social network. This allows us to cluster all tweets using their text alone with a focus on the textual features that are indicative of the likely social relationships of the user. To this end, we leverage the social proximity between two users, which is represented by their shortest path distance in the retweet network.

To demonstrate the effectiveness of this proxy signal, we compute the correlation between two metrics: the shortest path length between users in the network and the cosine similarity of sentence embeddings for their respective tweets. We randomly sample 1000 user pairs from the *Original* retweet network and compute their pairwise shortest path lengths and pairwise tweet cosine similarity. We found a Pearson correlation of approximately –0.25, a Spearman correlation of

–0.23, and a Kendall's Tau correlation of –0.17 between proximity and tweet similarity across networks. The negative correlation suggests that the smaller the distance in the network, the larger the cosine similarity of the tweets. This indicates that there is a presence of signals that can help to discriminate social proximity based only on the content of tweets. The correlation scores for all three network types and three correlation measures can be found in Table 3, showing similar correlations for all three of the networks.

Contrastive model

We develop a contrastive model to infuse the text embeddings with social network information. The goal of our contrastive fine-tuning task is to encode social proximity likelihood by making the embeddings for pairs of tweets similar if their authors are close (positive samples) and dissimilar if their authors are distant (negative samples). To achieve this, we leverage a contrastive loss function based on cosine similarity to measure proximity in the latent space.

$$\text{loss} = \begin{cases} 1 - \cos(x_1, x_2) & \text{if } y = 1 \\ \max(0, \cos(x_1, x_2) - m) & \text{if } y = -1 \end{cases}$$

where x_1 and x_2 are vectors, y is the label (1 for positive samples and –1 for negative samples), *cos* is the cosine similarity function, and m is a margin value. Through experimentation on the full training data, we tested margin values between 0 and 0.5, selecting m equal to 0. This means that the cosine embedding loss penalizes dissimilar pairs that have positive cosine similarity.

We tested two different pre-trained embedding models implemented in HuggingFace: DistilRoBERTa, which uses the all-distilroberta-v1 model (https://huggingface.co/sentence-transformers/all-distilroberta-v1) [43], and CT-BERT, which uses the covid-twitter-bert-v2 model (https://huggingface.co/digitalepidemiologylab/covid-twitter-bert-v2) [36]. We expect that DistilRoBERTa will better preserve the topical similarity of tweets, as the model was specifically trained on semantic similarity-related tasks, but it may potentially struggle in separating user viewpoints. On the other hand, we expect that the embeddings produced by CT-BERT will better match the data domain. We emphasize that there is no risk of train-test data contamination for CT-BERT as its data collection period is different from that of the *Avax* dataset. Specifically, the CT-BERT model was trained on tweets from January 12th to July 5th, 2020, while the *Avax* data starts on October 18th, 2020. Given that the time span of the CT-BERT training set was in the early phase of the pandemic, its in-domain discussion was likely less vaccine-focused than the *Avax* dataset.

Topic-aware contrastive model

While the pre-trained embedding representations successfully encode the topic similarity of tweets, the embeddings resulting from fine-tuning on the social proximity task might not be able to preserve this information. In fact, due to the nature of the task, the fine-tuned embeddings will most likely encode information from underlying social processes present in the network (e.g., polarization) and might disturb the overall latent topic space.

Table 3. Correlation scores between shortest path length (social proximity) and tweet similarity (content locality) across the retweet networks extracted in *Avax*. All correlation scores yield p-values below 0.05, indicating statistical significance. The tweet embeddings were extracted using DistilRoBERTa.

Network Type	Pearson	Spearman	Kendall's Tau
Original	–0.26	–0.23	–0.16
Active	–0.25	–0.24	–0.17
Strong	–0.26	–0.26	–0.18

https://doi.org/10.1371/journal.pcsy.0000089.t003

To address this issue, we experiment with an approach that consists of concatenating the initial pre-trained embedding representations (topic vector) with an extra n-dimensional vector. This new n-dimensional vector is learned by fine-tuning the embeddings of the pre-trained model and then passing them through a linear layer which reduces the vector to a size of n. The concatenation of the frozen pre-trained embedding and the learned extra dimensions is then used in the contrastive loss function. This approach aims to preserve the topic similarity of tweets in the latent space by using the frozen embedding while also encoding additional information related to shared viewpoints through the additional dimensions.

Evaluation metrics

We evaluate the performance of the fine-tuned contrastive models on three different metrics: The first metric involves measuring the correlation between the content similarity of tweets and the social proximity of users. The other two metrics, ViewpointNN and TopicNN, are used to assess the models' performance on the task of clustering tweets with similar viewpoints and topics.

Correlation We compute the Pearson, Spearman, and Kendall Tau correlation scores between the content similarity of tweet pairs, given by the cosine similarity of their text embeddings, and the social proximity of the authors of the tweets, given by their shortest path length in the retweet networks. The correlation metric does not measure the model's ability to encode viewpoints, but indicates the performance on the proxy task corresponding to weak social signals. A higher correlation score indicates that the latent space has encoded more social proximity information into its semantic representation.

ViewpointNN To probe the ability of the embeddings to encode viewpoint towards a particular topic, we used the idea of k nearest neighbors (k-NN). Specifically, for each tweet in the manually annotated evaluation set, we find its k nearest neighbors in the embedding space of all tweets from the same topic (i.e., tweets belonging to a different topic are not considered). We then predict the viewpoint class (e.g. Support or Oppose) by majority voting. The ViewpointNN metric is the proportion of correctly classified tweets among all the tweets in the annotated set. This metric will be higher if the learned latent space does a better job of separating the labeled data by viewpoint.

TopicNN To evaluate the ability of the embeddings to capture a shared topic, we also leverage k-NN using the manually annotated data. In this case, for each tweet, we find the k nearest neighbors in the embedding space of all other tweets in the annotated set (i.e., we consider all topics), and predict the topic class by majority voting. The TopicNN metric is the proportion of correctly classified tweets among all tweets in the annotated set. This metric will be higher if the learned latent space does a better job of separating the labeled data by topic.

Experimental setup

In this section, we detail the process of generating data splits for training, validation, and testing of the models. We present the hyper-parameters chosen for training each model. Finally, we describe the computing resources used for running experiments.

Data splitting for model training

To create the train, validation, and test data splits for the contrastive models, we sample 70% of users for training and 15% each for validation and testing using a stratified approach to maintain the distribution of topics across splits. For each split, we sample user pairs and compute their shortest path length in the corresponding retweet network. We transform the shortest path length values into a two-class representation: (1) user pairs in close proximity and (2) user pairs in distant proximity. We rely on a two-stage sampling approach to generate candidate pairs. First, we use a snowball sampling strategy to sample close connections by exploring the 1-hop and 2-hop neighborhoods of a random set of users. Second, we incrementally add more examples by randomly selecting two users, computing their shortest path length, and retaining them if their distance is larger than 2. Our sampling approach also has the benefit of mitigating the bias in the dataset towards prolific users. Our dataset, as in most social media sets, exhibits a significantly skewed distribution of tweets per

user, with the most active 10% of users making 42.8% of the tweets. However, because we sample pairs of users for training without regard to their number of tweets, we will not be overly biased towards the more active users, mitigating concerns that the results will be only reflective of a vocal minority.

Furthermore, we augment the number of close pairs in each split by adding pairs of tweets authored by the same user, which are likely to reflect similar viewpoints. We select pairs of tweets from each pair of users considering only tweets from the same topic to encourage the model to learn differing viewpoints towards the same topic. We use a shortest path length cutoff value of 4, which means that path lengths in the range of 1-3 are considered close while values of 4 or greater are distant. We found that models trained using a cutoff of 4 yielded better performance on average (as shown in Fig 3). To address the potential data imbalance arising from the chosen cutoff, we also down-sample the majority class to achieve more balanced class distributions in the resulting data splits.

Table 4 shows the summary statistics for each data split across the various types of retweet networks considered.

Model hyperparameters

For each contrastive model, we use the Adam optimizer with a learning rate of 5e−7 and weight decay of 1e−4. Each model was trained for 200 epochs with a mini-batch size of 32. We used early stopping to prevent over-fitting by tracking the loss on the validation set. A dropout rate of 0.2 was used for each network. We set the margin of the cosine loss embedding function to zero. For the topic-aware contrastive models, we set the extra embedding dimensions of the shared viewpoint vector to 100.

Hardware and software resources

The experiments were conducted on a computer with a Dual Intel Broadwell E5-2620 v4 @ 2.10GHz CPU. Each node has 16 cores with 64GB of memory. The operating system is CentOS Linux version 7. For deep learning training, we use an NVIDIA A100 GPU with 40GB of memory. All code was implemented in Python 3.8, making use of various software libraries, including Pandas and NumPy for data exploration and preprocessing; PyTorch for model building and training; scikit-learn for model evaluation; and BERTopic, Top2Vec, UMAP, and HDBSCAN for clustering and topic extraction. Visualizations were generated using Seaborn and matplotlib.

Results

In this section, we report the performance of our fine-tuned embeddings in the task of encoding viewpoint and topic information. We compare our model against two baseline sentence embeddings to demonstrate its effectiveness and limitations. Finally, we conduct an analysis of tweet representations within the latent space and use the learned embeddings in topic modeling techniques to investigate their efficacy in extracting detailed topics.

Table 4. The number of sample pairs, tweets and users for the training, validation, and test data splits per dataset.

Network	Split	Pairs	Tweets	Users
Original	Train	70,032	63,624	24,842
	Val	15,026	14,902	5,674
	Test	15,026	15,003	5,760
Active	Train	70,031	61,545	20,820
	Val	15,014	14,226	4,714
	Test	15,015	14,565	4,817
Strong	Train	35,031	30,118	5,835
	Val	7,479	7,190	1,313
	Test	7,427	6,916	1,306

https://doi.org/10.1371/journal.pcsy.0000089.t004

Model performance

Table 5 shows the full set of metric results for all three social networks (*Original*, *Active*, and *Strong*), each starting embedding (DistilRoBERTa and CT-BERT), and each weak supervision method (fine-tuned and topic-aware fine-tuned). The ViewpointNN and TopicNN metrics of select models are also shown in Fig 3 to better highlight the performance trade-off between these two metrics.

First, we evaluate our models on the proxy task of encoding the social proximity of users within the tweet embeddings. The contrastive training leads to consistent improvements relative to the starting embeddings in all correlation scores, which measure the similarity between text embeddings and the shortest path distance of their respective authors.

Table 5. Model performance over the three types of retweet networks. We report performance on ViewpointNN and TopicNN metrics with $k = 5$ as well as their average (NN Mean), and the Pearson correlation metric. The Baseline performance for ViewpointNN and TopicNN metrics is the same across different retweet networks as the baselines do not leverage the social network information. The correlation metric is computed on the corresponding test split for each network type. FT refers to fine-tuned contrastive models, and TA-FT refers to topic-aware contrastive models. The best results for each metric is shown in bold.

Base Model	Training	Network	Correlation	ViewpointNN	TopicNN	NN Mean
DistilRoBERTa	Baseline	None	–0.26	0.65	**0.98**	0.83
	FT	Original	–0.32	0.73	0.79	0.76
		Active	–0.24	0.69	0.68	0.69
		Strong	–0.30	0.69	0.59	0.64
	TA-FT	Original	–0.37	0.72	0.90	0.81
		Active	–0.26	0.65	**0.98**	0.83
		Strong	–0.37	0.69	0.86	0.78
CT-BERT	Baseline	None	–0.33	0.82	0.88	0.85
	FT	Original	–0.44	**0.92**	0.53	0.73
		Active	–0.39	**0.92**	0.60	0.76
		Strong	–0.35	0.92	0.60	0.76
	TA-FT	Original	**–0.46**	**0.92**	0.73	0.83
		Active	–0.44	0.89	0.78	0.84
		Strong	–0.43	0.92	0.81	**0.87**

https://doi.org/10.1371/journal.pcsy.0000089.t005

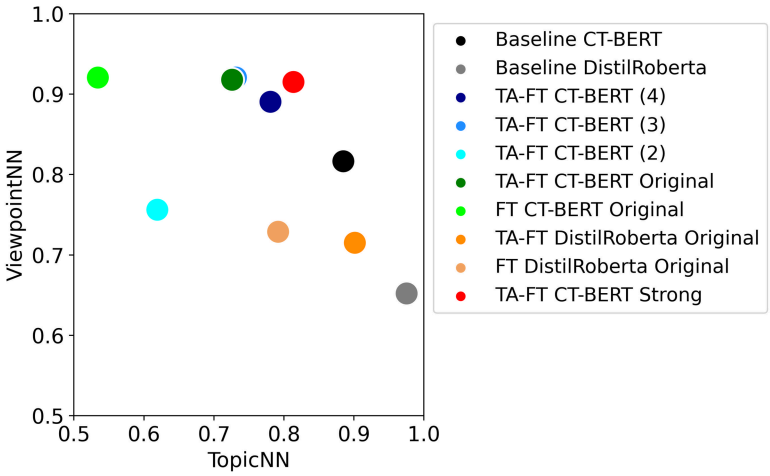


Fig 3. ViewpointNN vs. TopicNN performance with K = 5 of contrastive models and the two baselines. We include performance for the following: three models trained on the active retweet network with different cut-off values of shortest path length (4, 3 and 2), four models trained with different pre-trained embeddings on the original network, and the best-performing model trained on the strong network. FT refers to fully fine-tuned models, and TA-FT refers to topic-aware models.

https://doi.org/10.1371/journal.pcsy.0000089.g003

Our fine-tuned embeddings (FT) have an average improvement of 0.05 in correlation and the topic-aware embeddings (TA-FT) have an average improvement of 0.10 in correlation relative to the starting embeddings. This suggests that our fine-tuned embeddings capture more informative signals for discerning social proximity based solely on the content of the tweets.

To assess the ability of our trained models in encoding text representations with viewpoint information, we evaluate the fine-tuned embeddings and the two baseline embeddings on the annotated set of tweets described in the data section. We observe an improvement in ViewpointNN performance with both fully fine-tuned models (FT) and topic-aware models (TA-FT) compared to their corresponding baselines. Specifically, when considering models trained on the original retweet network, the FT models have a performance improvement of 12.3% with DistilRoBERTa embeddings and 12.2% with CT-BERT embeddings. Similarly, the TA-FT models also present a performance improvement of 10.7% with DistilRoBERTa embeddings and 12.2% with CT-BERT embeddings.

The observed improvement in viewpoint performance is associated with a degradation in topic performance as measured by TopicNN. Specifically, the FT models show a decrease in performance of 19.4% compared to the DistilRoBERTa baseline and 39.8% compared to the CT-BERT baseline. The best performance on the TopicNN metric is achieved by the DistilRoBERTa baseline, which is expected given that its embeddings are trained on sentence-similarity tasks, allowing them to encode more robust semantic similarities. However, these baseline embeddings have the worst ViewpointNN performance.

The observed trade-off between viewpoint and topic performance is driven by the nature of our proxy task. For example, tweets from distinct topics but likely similar viewpoints may be projected into similar regions in the embedding space, which inevitably leads to a reduction in topic performance. TA-FT models can mitigate this problem as they achieve a considerably smaller degradation in TopicNN compared to FT models. Specifically, the TA-FT models show a decrease in TopicNN performance of only 8.1% compared to the DistilRoBERTa baseline and 17.0% compared to the CT-BERT baseline while achieving an improvement on ViewpointNN performance.

The domain-specific pre-training of CT-BERT improves the encoding of viewpoint embeddings by 26% relative to starting from the DistilRoBERTa embeddings. This highlights the importance of leveraging domain-specific embeddings over out-of-domain embeddings, as the former contains more relevant contextual information and vocabulary tailored to the specific domain of interest. In comparing the performance across the three different networks, we found that the model trained with the retweet network of strong relationships demonstrated the best performance with an improvement of 12.2% in ViewpointNN and a reduction of 8.0% in TopicNN compared to the CT-BERT baseline. This can be attributed to its consideration of less noisy interactions, thus avoiding spurious interactions by chance.

We also investigate the impact of different values of k, which controls the number of neighbors to be considered for the ViewpointNN and TopicNN metrics. Fig 4 shows the performance for TA-FT CT-BERT models in ViewpointNN and TopicNN at different values of k. We present results for five models: two baselines, each using a different pre-trained embedding model, and three models fine-tuned using our topic-aware approach on data derived from three distinct types of retweet networks. The latter three models use CT-BERT embeddings as their initialization. We observed that increasing k has little effect on ViewpointNN, which indicates that our embeddings are robust in capturing different viewpoints even when varying the number of nearest neighbors. However, for TopicNN, larger k values result in tweets of different topics being included in the neighboring examples, which leads to a decrease in topic performance.

So far, we have demonstrated the improved performance on viewpoint encodings using our method on an aggregate basis across our annotated data. Next, we explore the performance within individual topics to understand whether the improvements are attributable to specific subsets or appear to generalize across topics. Fig 5 shows the ViewpointNN performance of our best-performing model (TA-FT CT-BERT Strong) and the two baselines across each topic in the annotated set of tweets. We find that our model outperforms the two baselines across all topics, showing that the approach learns improved representations across multiple topic areas within this case study.

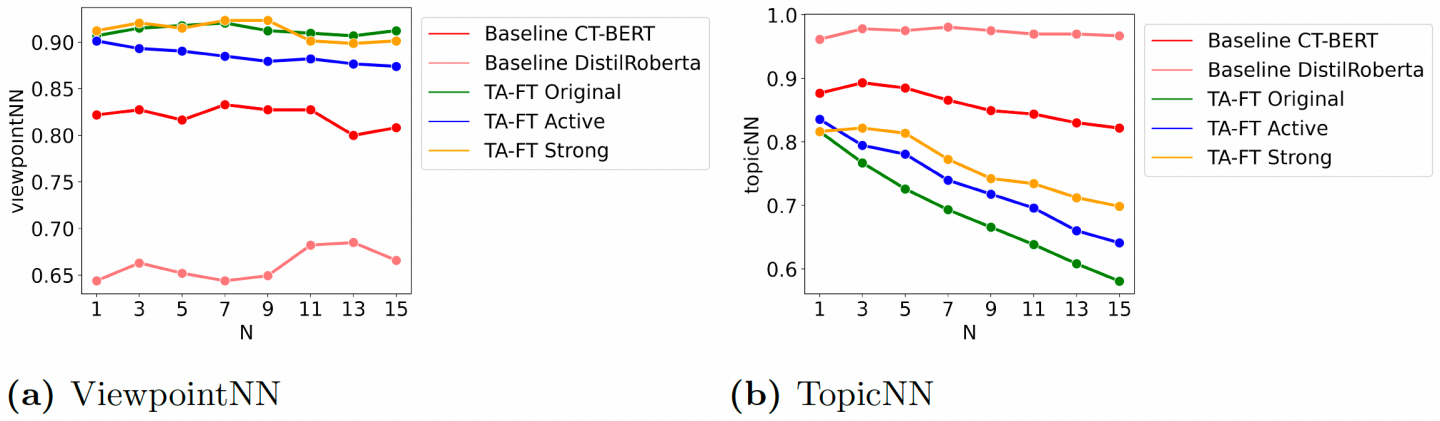


Fig 4. Model performance in ViewpointNN and TopicNN at different values of *k* for topic-aware contrastive models trained with CT-BERT embeddings. TA-FT refers to models leveraging both fine-tuned embeddings and static topic-aware embeddings. Models shown here are trained using the original, active, and strong retweet networks. (a) ViewpointNN. (b) TopicNN.

https://doi.org/10.1371/journal.pcsy.0000089.g004

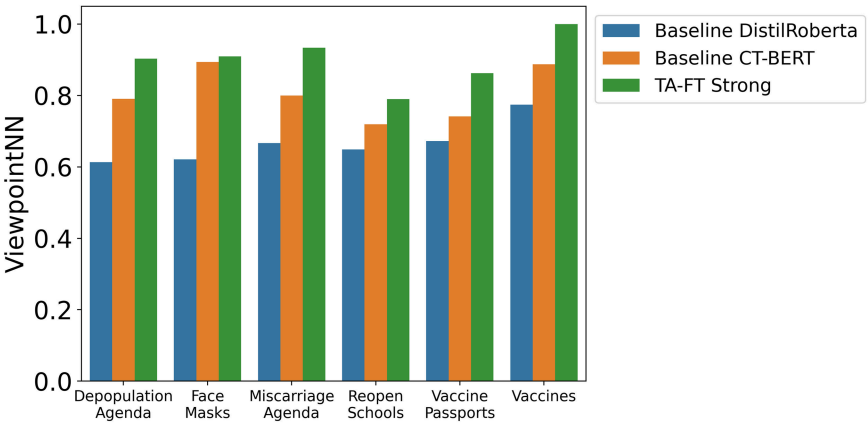


Fig 5. ViewpointNN performance with *k*=15 of the best contrastive model (TA-FT Strong) and baselines across the different topics in the annotated set of tweets.

https://doi.org/10.1371/journal.pcsy.0000089.g005

Latent space analysis

We visualize the latent space of the annotated set of tweets by using the UMAP algorithm to project the embeddings for each tweet to a 2D space. Fig 6a shows the latent space embedding of the model that achieves the best trade-off between ViewpointNN and TopicNN performance, *TA-FT CT-BERT Strong*, which can be compared with the two baseline embeddings shown in Fig 2. DistilRoBERTa baseline embeddings are successful at encoding tweets with similar topics in similar regions of the latent space but fail to distinguish between the binary viewpoints on each topic. CT-BERT baseline embeddings can also identify semantically similar groups of tweets, though the resulting clusters are not as cleanly separated. Interestingly, the CT-BERT baseline is able to capture some partial viewpoint information within some of the topics. The embeddings generated by our topic-aware fine-tuned CT-BERT model demonstrate that tweets sharing similar viewpoints tend to group together, while those with opposite viewpoints are pushed to distant regions of the latent space, showing a binary pattern that likely reflects the polarization of the social network.

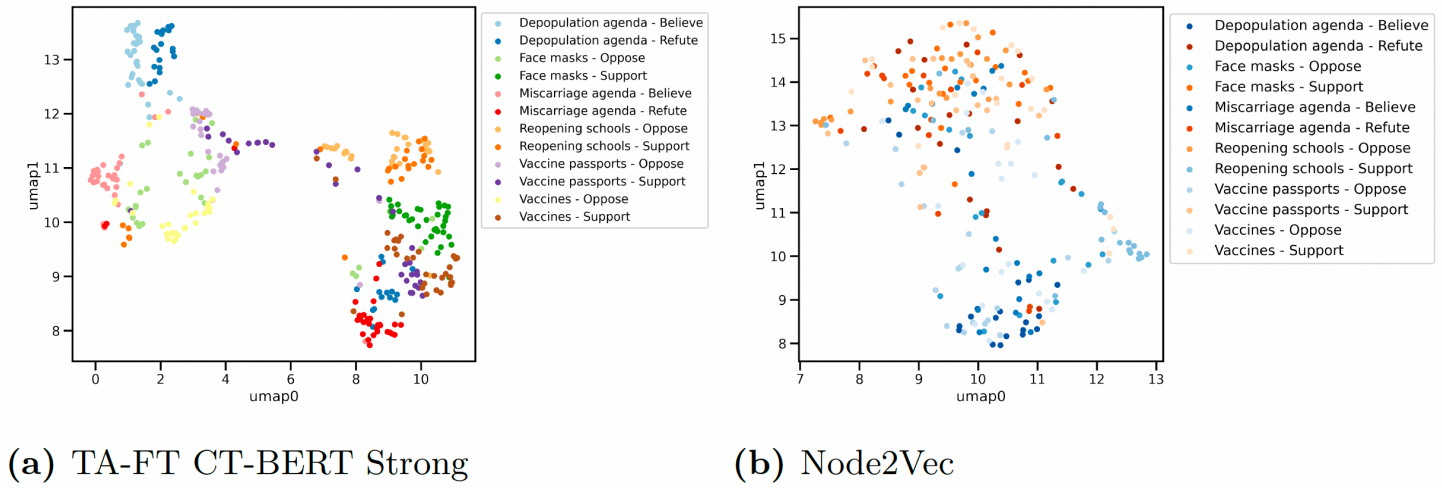


Fig 6. (Left) UMAP visualizations of the embeddings from the best performing topic-aware fine-tuned model trained with the strong retweet network. (Right) UMAP visualization of annotated set of tweets with their corresponding node2vec embeddings. (a) TA-FT CT-BERT Strong. (b) Node2Vec.

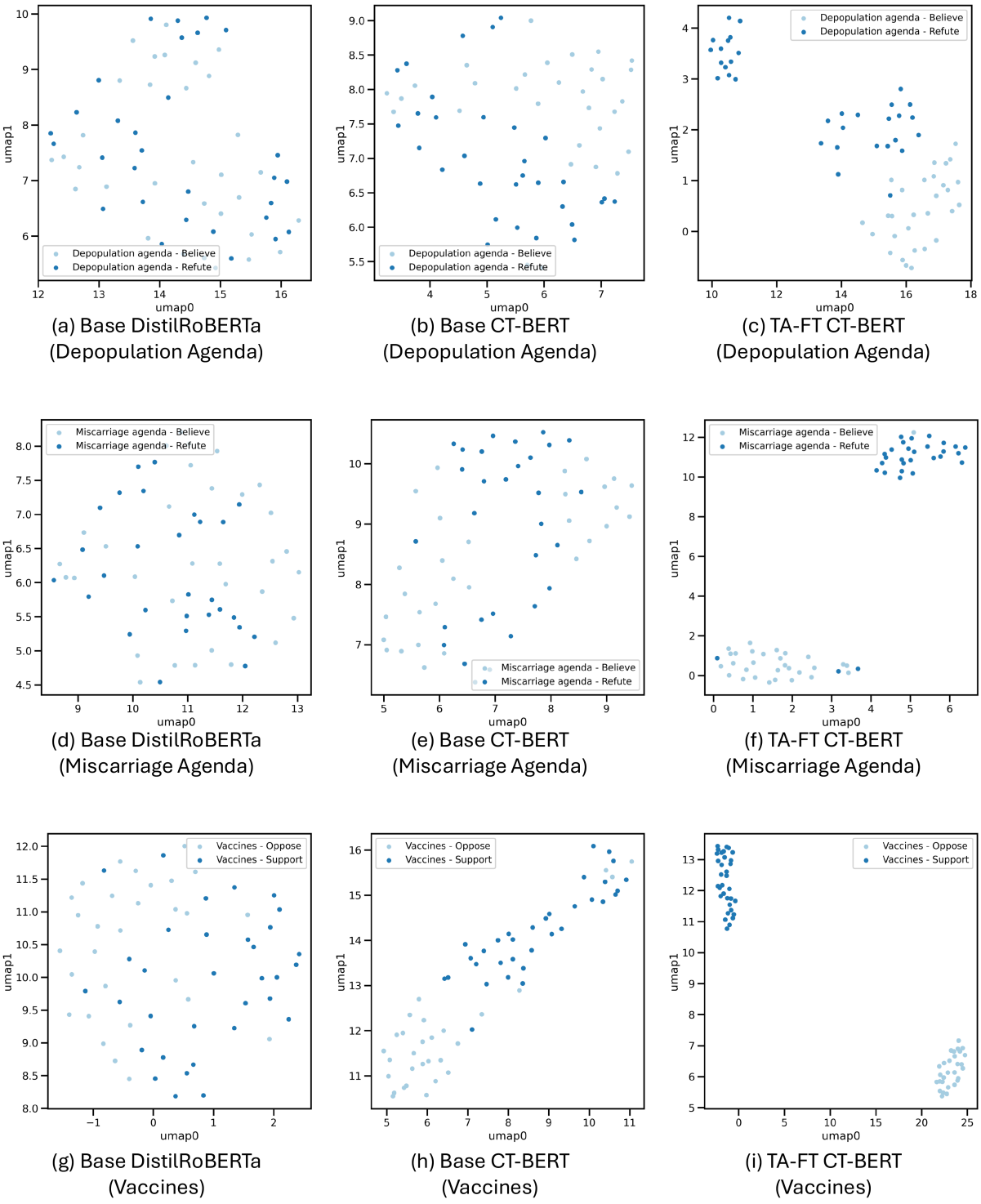
https://doi.org/10.1371/journal.pcsy.0000089.g006

To further demonstrate the ability of the fine-tuned embeddings in capturing narrative viewpoints, we visualize the latent space of three out of the six topics as identified in Table 1: depopulation agenda (Conspiracy), miscarriage agenda (Conspiracy), and vaccines (General). Fig 7 shows the UMAP visualizations per topic for two baselines and the best-performing fine-tuned embeddings. We observed that DistilRoBERTa embeddings struggle the most in distinguishing opposing viewpoints towards a particular topic as there is no clear separation between tweets. The CT-BERT baseline seems to partially capture viewpoint information as shown by visualizations on the depopulation and vaccines topics, but still struggles in other topics such as the miscarriage agenda topic. Overall, the CT-BERT TA-FT embeddings provide the best performance in capturing viewpoint information across multiple topics by visually demonstrating the ability to separate most tweets by their stance in each topic. These improved separations occur even though the model has not been trained directly on the viewpoint tasks but instead has inferred the new clustering based on indirect training on social proximity information.

Topic modeling analysis

Next, we qualitatively explore the clustering behavior of the embedding spaces and aim to determine whether the viewpoint-infused embeddings support improved performance on downstream tasks such as topic modeling. We employed BERTopic [18], a topic modeling technique, on the text representations generated by our best-performing model and the two baselines. BERTopic uses UMAP for dimensionality reduction and HDBSCAN for document clustering. It also offers the flexibility to incorporate custom embeddings. The algorithm assigns documents to a specific topic, and if a document is not representative enough, it is classified as noise. In particular, BERTopic identified 7 topics using DistilRoBERTa baseline embeddings with 20 tweets as noise, 9 topics using CT-BERT baseline embeddings with 40 tweets as noise, and 11 topics using the TA-FT CT-BERT Strong embeddings with 96 tweets as noise.

The extraction of meaningful narratives relies on the ability of each cluster to encompass tweets that share both an underlying topic area and corresponding viewpoints. Fig 8 shows the distribution of annotated topic categories across each topic identified by the BERTopic algorithm. The annotated topic categories are color-coded based on two groups of viewpoints that are more likely to be associated based on the polarization observed in the social network. This association can be seen in Fig 6b, where we projected the annotated tweets with their corresponding Node2Vec embeddings into



(a) Base DistilRoBERTa (Depopulation Agenda)
(b) Base CT-BERT (Depopulation Agenda)
(c) TA-FT CT-BERT (Depopulation Agenda)
(d) Base DistilRoBERTa (Miscarriage Agenda)
(e) Base CT-BERT (Miscarriage Agenda)
(f) TA-FT CT-BERT (Miscarriage Agenda)
(g) Base DistilRoBERTa (Vaccines)
(h) Base CT-BERT (Vaccines)
(i) TA-FT CT-BERT (Vaccines)

Fig 7. UMAP visualizations of baseline embeddings and the embeddings from the best performing topic-aware fine-tuned model across different annotated topics. The TA-FT CT-BERT model shown uses the strong network for training. (a) Base DistilRoBERTa (Depopulation Agenda). (b) Base CT-BERT (Depopulation Agenda). (c) TA-FT CT-BERT (Depopulation Agenda). (d) Base DistilRoBERTa (Miscarriage Agenda). (e) Base CT-BERT (Miscarriage Agenda). (f) TA-FT CT-BERT (Miscarriage Agenda). (g) Base DistilRoBERTa (Vaccines). (h) Base CT-BERT (Vaccines). (i) TA-FT CT-BERT (Vaccines).

https://doi.org/10.1371/journal.pcsy.0000089.g007

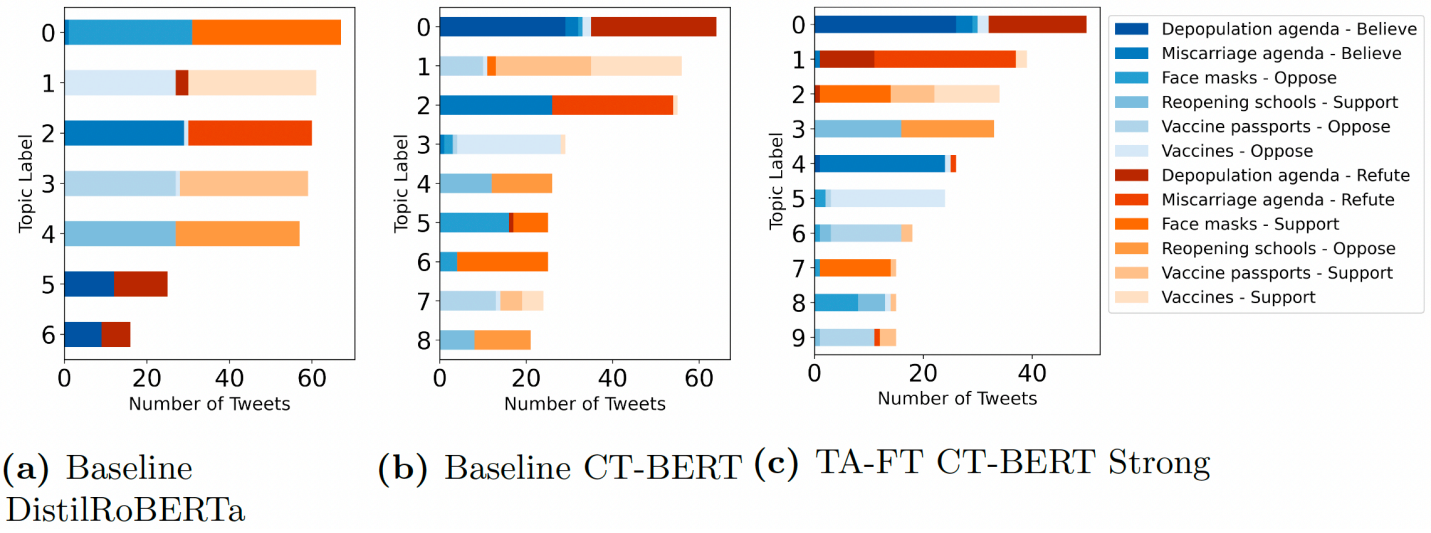


Fig 8. Stacked bar plot showing the distribution of topic categories across BERTopic-identified topics. The y-axis displays the BERTopic topics, while the x-axis represents the number of tweets belonging to the 12 annotated categories, which are listed in the legend. Viewpoints are shaded blue or red based based on two groups of viewpoints more likely to be associated based on the social network polarization. (a) Baseline DistilRoBERTa. (b) Baseline CT-BERT. (c) TA-FT CT-BERT Strong.

https://doi.org/10.1371/journal.pcsy.0000089.g008

a latent space. DistilRoBERTa embeddings generate topics that are roughly evenly split between opposing viewpoints on specific themes, highlighting its strong topic clustering performance and weak viewpoint clustering performance. CT-BERT embeddings exhibit similar results to DistilRoBERTa, with exceptions in topics 3 and 6, where the majority of tweets express vaccine opposition and face masks support, respectively. On the other hand, while our fine-tuned embeddings do not entirely preserve the similarity of tweets sharing the same topic, they effectively group tweets into informative regions of the latent space with shared or associated viewpoints. Specifically, topics 4, 5, 6, 7, and 9 are composed primarily of a single topic and viewpoint, while topics 1, 2, and 8 comprise viewpoints toward multiple topics that tend to be associated. Overall, we found that the majority of the detected topics clearly focus on a similar viewpoint towards one or multiple subject areas.

Discussion

In this work, we have introduced a weakly supervised contrastive approach to encode narrative viewpoint of tweets. Our approach leverages social proximity signals observed through user interaction networks to identify the textual features that make a pair of tweets appear as if their authors would be closely related in the social network. Using an X (Twitter) dataset related to COVID-19 discussions, we show that our models are better than baseline sentence embedding methods in distinguishing tweets with shared and opposing viewpoints towards particular topics, and in effectively grouping them into similar regions of the latent space. We show how the learned representations can be leveraged for the automated identification of narratives through clustering of the latent space. Clustering using our fine-tuned embeddings shows significant improvement over baseline embeddings in terms of discovering viewpoint-cohesive tweet subsets that may represent specific narratives. While our method is trained using social proximity information based on user interactions, we emphasize that the use of the model at inference time requires only the text of the tweets to generate the embeddings. This allows us to generate a latent space representation that encodes both topic and viewpoint information from the textual content of tweets alone, as the model has learned the textual signals that are likely to correlate with social proximity.

While we demonstrate the utility of our method using a clustering workflow, one of the benefits of our approach compared with methods designed specifically for topic modeling or stance detection is that viewpoint-infused embeddings can be leveraged for many potential downstream tasks. One avenue for future work is to evaluate the benefits of the fine-tuned embeddings for tasks such as supervised or semi-supervised tweet classification, narrative and event extraction, and change point analysis. Additionally, it would be interesting to explore whether our embeddings can capture information relevant to the detection of bots engaged in social media discussions. Especially when bots are engaged in coordinated behavior such as botnets, their interactions and posting styles [50] may cause distinct clusters to form our in refined embedding space and provide potential flags for detection.

One limitation of the current evaluation approach for our method is that it relies on binary annotations of viewpoint, while true user viewpoints are likely to be much more complex and nuanced. Further evaluation would be needed to demonstrate whether our method can capture these additional nuances to infuse the embeddings with fine-grained viewpoint information. Additionally, while we have demonstrated the utility of weak supervision leveraging intrinsic social media information, there is significant room to build on the groundwork of our approach. Further exploration of how our embedding approach could be combined with improved topic model approaches such as Contextual Top2Vec [3] would be a fruitful future direction. Additionally, social network proximity provides one source of signal on user viewpoints through the observed assortativity and polarization of such interactions [10,17]. However, the methods described here could be augmented by incorporating additional weak signals into the model supervision. Future work could investigate the incorporation of additional signals, for example, contrastive models that can infer whether two tweets use the same hashtag or share the same URL based only on the natural language text of the tweet. The incorporation of multiple such weak signals would likely provide additional information which could enhance the topic representation of embeddings and further mitigate the trade-off between viewpoint and topic performance, leading to more robust representations. Another potentially interesting direction for further work would be to incorporate temporal information into the method and analysis to better capture how individual, group, and overall viewpoints are evolving over time.

Acknowledgments

The research described in this paper was conducted under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy.

Author contributions

Conceptualization: Emily Saldanha.

Data curation: Jasmine Eshun.

Formal analysis: Kin Wai Ng.

Funding acquisition: Emily Saldanha.

Methodology: Kin Wai Ng, Nathan Wendt, Jasmine Eshun, Emily Saldanha.

Software: Kin Wai Ng, Nathan Wendt.

Supervision: Emily Saldanha.

Validation: Kin Wai Ng.

Visualization: Kin Wai Ng.

Writing – original draft: Kin Wai Ng, Emily Saldanha.

Writing – review & editing: Kin Wai Ng, Emily Saldanha.

References

1. An C, Feng J, Lv K, Kong L, Qiu X, Huang X. CoNT: Contrastive neural text generation. In: Advances in Neural Information Processing Systems. 2022. https://openreview.net/forum?id=mjVZw5ADSbX
2. Angelov D. Top2vec: distributed representations of topics. arXiv preprint 2020. https://arxiv.org/abs/2008.09470
3. Angelov D, Inkpen D. Topic modeling: contextual token embeddings are all you need. In: Findings of the Association for Computational Linguistics: EMNLP 2024. 2024. p. 13528–39. https://doi.org/10.18653/v1/2024.findings-emnlp.790
4. Batrancea I, Balcı MA, Batrancea LM, Akgüller Ö, Tulai H, Rus M-I, et al. Topic analysis of social media posts during the COVID-19 pandemic: evidence from Tweets in Turkish. J Knowl Econ. 2023;15(3):12361–91. https://doi.org/10.1007/s13132-023-01565-6
5. Benamara F, Inkpen D, Taboada M. Introduction to the special issue on language in social media: exploiting discourse and other contextual information. Computational Linguistics. 2018;44(4):663–81. https://doi.org/10.1162/coli_a_00333
6. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research. 2003;3:993–1022.
7. Boon-Itt S, Skunkan Y. Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. JMIR Public Health Surveill. 2020;6(4):e21978. https://doi.org/10.2196/21978 PMID: 33108310
8. Caron M, Misra I, Mairal J, Goyal P, Bojanowski P, Joulin A. Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems. 2020;33:9912–24.
9. Chaudhary Y, Gupta P, Saxena K, Kulkarni V, Runkler T, Schütze H. TopicBERT for energy efficient document classification. In: Findings of the Association for Computational Linguistics: EMNLP 2020. 2020. p. 1682–90. https://doi.org/10.18653/v1/2020.findings-emnlp.152
10. Conover M, Ratkiewicz J, Francisco M, Goncalves B, Menczer F, Flammini A. Political polarization on Twitter. ICWSM. 2021;5(1):89–96. https://doi.org/10.1609/icwsm.v5i1.14126
11. Darwish K, Stefanov P, Aupetit M, Nakov P. Unsupervised user stance detection on Twitter. ICWSM. 2020;14:141–52. https://doi.org/10.1609/icwsm.v14i1.7286
12. Dey K, Shrivastava R, Kaushik S. Topical stance detection for twitter: a two-phase lstm model using attention. In: European Conference on Information Retrieval. 2018. p. 529–36.
13. Doogan C, Buntine W, Linger H, Brunt S. Public perceptions and attitudes toward COVID-19 nonpharmaceutical interventions across six countries: a topic modeling analysis of Twitter data. J Med Internet Res. 2020;22(9):e21419. https://doi.org/10.2196/21419 PMID: 32784190
14. Gaisbauer F, Pournaki A, Banisch S, Olbrich E. Ideological differences in engagement in public debate on twitter. PLoS One. 2021;16(3):e0249241.
15. Gao T, Yao X, Chen D. SimCSE: simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. p. 6894–910. https://doi.org/10.18653/v1/2021.emnlp-main.552
16. Gao, W, Li, P, Darwish, K.: Joint topic modeling for event summarization across news and social media streams. In: Social Media Content Analysis: Natural Language Processing and Beyond. World Scientific; 2018. p. 321–46.
17. Garimella VRK, Weber I. A long-term analysis of polarization on twitter. In: Eleventh international AAAI conference on web and social media. 2017.
18. Grootendorst M. Bertopic: neural topic modeling with a class-based TF-IDF procedure. arXiv preprint 2022. https://arxiv.org/abs/2203.05794
19. Grover A, Leskovec J. node2vec: scalable feature learning for networks. KDD. 2016;2016:855–64. https://doi.org/10.1145/2939672.2939754 PMID: 27853626
20. Halberstam Y, Knight B. Homophily, group size, and the diffusion of political information in social networks: evidence from Twitter. Journal of Public Economics. 2016;143:73–88. https://doi.org/10.1016/j.jpubeco.2016.08.011
21. Jachim P, Sharevski F, Pieroni E. TrollHunter2020: real-time detection of trolling narratives on Twitter during the 2020 U.S. Elections. In: Proceedings of the 2021 ACM Workshop on Security and Privacy Analytics. 2021. p. 55–65. https://doi.org/10.1145/3445970.3451158
22. Jang H, Rempel E, Roth D, Carenini G, Janjua NZ. Tracking COVID-19 discourse on Twitter in North America: infodemiology study using topic modeling and aspect-based sentiment analysis. J Med Internet Res. 2021;23(2):e25431. https://doi.org/10.2196/25431 PMID: 33497352
23. Jiang J, Ren X, Ferrara E. Retweet-BERT: political leaning detection using language features and information diffusion on social networks. ICWSM. 2023;17:459–69. https://doi.org/10.1609/icwsm.v17i1.22160
24. Jing E, Ahn Y-Y. Characterizing partisan political narrative frameworks about COVID-19 on Twitter. EPJ Data Sci. 2021;10(1):53. https://doi.org/10.1140/epjds/s13688-021-00308-4 PMID: 34745825
25. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. Advances in Neural Information Processing Systems. 2020;33:18661–73.
26. Kim, N, Mosallanezhad, D, Cheng, L, Mancenido, MV, Liu, H.: Robust stance detection: understanding public perceptions in social media. In: International Conference on Advances in Social Networks Analysis and Mining. Springer; 2024. p. 21–37.
27. Küçük D, Can F. Stance detection. ACM Comput Surv. 2020;53(1):1–37. https://doi.org/10.1145/3369026
28. Kwak H, An J, Jing E, Ahn Y-Y. FrameAxis: characterizing microframe bias and intensity with word embedding. PeerJ Comput Sci. 2021;7:e644. https://doi.org/10.7717/peerj-cs.644 PMID: 34395864

29. Lison P, Barnes J, Hubin A, Touileb S. Named entity recognition without labelled data: a weak supervision approach. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. p. 1518–33. https://doi.org/10.18653/v1/2020.acl-main.139
30. Luo, B, Feng, Y, Wang, Z, Zhu, Z, Huang, S, Yan, R, Zhao, D.: Learning with noise: enhance distantly supervised relation extraction with dynamic transition matrix. arXiv preprint 2017. arXiv:1705.03995
31. Lyu JC, Han EL, Luli GK. COVID-19 vaccine-related discussion on twitter: topic modeling and sentiment analysis. J Med Internet Res. 2021;23(6):e24435. https://doi.org/10.2196/24435 PMID: 34115608
32. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. JOSS. 2018;3(29):861. https://doi.org/10.21105/joss.00861
33. Mekala D, Shang J. Contextualized weak supervision for text classification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. p. 323–33. https://doi.org/10.18653/v1/2020.acl-main.30
34. Meng Y, Shen J, Zhang C, Han J. Weakly-supervised neural text classification. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018. p. 983–92. https://doi.org/10.1145/3269206.3271737
35. Metaxas P, Mustafaraj E, Wong K, Zeng L, O'Keefe M, Finn S. What do retweets indicate? results from user survey and meta-review of research. ICWSM. 2021;9(1):658–61. https://doi.org/10.1609/icwsm.v9i1.14661
36. Müller M, Salathé M, Kummervold PE. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arXiv preprint 2020. https://arxiv.org/abs/2005.07503
37. Muric G, Wu Y, Ferrara E. COVID-19 vaccine hesitancy on social media: building a public twitter data set of antivaccine content, vaccine misinformation, and conspiracies. JMIR Public Health Surveill. 2021;7(11):e30642. https://doi.org/10.2196/30642 PMID: 34653016
38. Nerghes A, Lee J-S. Narratives of the refugee crisis: a comparative study of mainstream-media and Twitter. MaC. 2019;7(2):275–88. https://doi.org/10.17645/mac.v7i2.1983
39. Ng KW, Wendt N, Eshun J, Saldanha E. Weakly supervised contrastive representation learning to encode narrative viewpoint of covid-19 tweets. In: Complex Networks & Their Applications XIII: Proceedings of the Thirteenth International Conference on Complex Networks and Their Applications: COMPLEX NETWORKS -Volume 3. Springer; 2024, p. 246.
40. Ng QX, Lee DYX, Ng CX, Yau CE, Lim YL, Liew TM. Examining the negative sentiments related to influenza vaccination from 2017 to 2022: an unsupervised deep learning analysis of 261,613 Twitter Posts. Vaccines (Basel). 2023;11(6):1018. https://doi.org/10.3390/vaccines11061018 PMID: 37376407
41. Ng QX, Lim SR, Yau CE, Liew TM. Examining the prevailing negative sentiments related to COVID-19 vaccination: unsupervised deep learning of twitter posts over a 16 month period. Vaccines (Basel). 2022;10(9):1457. https://doi.org/10.3390/vaccines10091457 PMID: 36146535
42. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. Proceedings VLDB Endowment. 2017;11(3):269–82. https://doi.org/10.14778/3157794.3157797 PMID: 29770249
43. Reimers N, Gurevych I. Sentence-bert: sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019. https://arxiv.org/abs/1908.10084
44. Scott K. The pragmatics of rebroadcasting content on Twitter: how is retweeting relevant?. Journal of Pragmatics. 2021;184:52–60. https://doi.org/10.1016/j.pragma.2021.07.022
45. Shahsavari S, Holur P, Wang T, Tangherlini TR, Roychowdhury V. Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news. J Comput Soc Sci. 2020;3(2):279–317. https://doi.org/10.1007/s42001-020-00086-5 PMID: 33134595
46. Sutter M, Gourru A, Trabelsi A, Largeron C. Unsupervised stance detection for social media discussions: a generic baseline. In: Graham Y, Purver M eds. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). St. Julian's, Malta: Association for Computational Linguistics; 2024, p. 1782–92. https://aclanthology.org/2024.eacl-long.107
47. Tangherlini TR, Shahsavari S, Shahbazi B, Ebrahimzadeh E, Roychowdhury V: An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: bridgegate, pizzagate and storytelling on the web. PLoS One 2020;15(6):e0233879.
48. Xiong L, Xiong C, Li Y, Tang KF, Liu J, Bennett PN, et al. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In: International Conference on Learning Representations; 2021. https://openreview.net/forum?id=zeFrfgyZln
49. Xue J, Chen J, Chen C, Zheng C, Li S, Zhu T. Public discourse and sentiment during the covid 19 pandemic: using latent dirichlet allocation for topic modeling on twitter. PLoS One. 2020;15(9):e0239441.
50. Yang K, Menczer F. Anatomy of an AI-powered malicious social botnet. JQD. 2024;4. https://doi.org/10.51685/jqd.2024.icwsm.7
51. Yin H, Song X, Yang S, Li J. Sentiment analysis and topic modeling for COVID-19 vaccine discussions. World Wide Web. 2022;25(3):1067–83. https://doi.org/10.1007/s11280-022-01029-y PMID: 35250362
52. Zhang C, Zhou Z, Peng X, Xu K. DoubleH: Twitter user stance detection via bipartite graph neural networks. ICWSM. 2024;18:1766–78. https://doi.org/10.1609/icwsm.v18i1.31424
53. Zhou Z, Elejalde E. Stance inference in Twitter through graph convolutional collaborative filtering networks with minimal supervision. In: Companion Proceedings of the ACM Web Conference 2023. 2023. p. 1030–8. https://doi.org/10.1145/3543873.3587640