Unsupervised Stance Detection for Social Media Discussions: A Generic Baseline

Maïa Sutter[1] **Antoine Gourru**[1] **Amine Trabelsi**[2] **Christine Largeron**[1]
Laboratoire Hubert Curien, UMR CNRS 5516, Saint-Etienne, France
Université de Sherbrooke, Department of Computer Science, Canada
maia.d.sutter@gmail.com, antoine.gourru@univ-st-etienne.fr,
amine.trabelsi@usherbrooke.ca, christine.largeron@univ-st-etienne.fr

Abstract

With the ever-growing use of social media to express opinions on the national and international stage, unsupervised methods of stance detection are increasingly important to handle the task without costly annotation of data. The current unsupervised state-of-the-art models are designed for specific network types, either homophilic or heterophilic, and they fail to generalize to both. In this paper, we first analyze the generalization ability of recent baselines to these two very different network types. Then, we conduct extensive experiments with a baseline model based on text embeddings propagated with a graph neural network that generalizes well to heterophilic and homophilic networks. We show that it outperforms, on average, other state-of-the-art methods across the two network types. Additionally, we show that combining textual and network information outperforms using text only, and that the language model size has only a limited impact on the model performance.

1 Introduction

Stance detection is the task of determining the position of a text or person towards a certain target, often split into "for" and "against" with an optional third split being "neutral", "unknown", or "neither". The target can be an entity, a topic, or other subject, such as a claim or event.

While our work focuses on unsupervised methods for stance detection, much of the recent work on this domain has been done with supervised methods (Sun et al., 2023; Zhu et al., 2022; Zhang et al., 2023; Liang et al., 2022; Largeron et al., 2021). In contrast, only a limited range of methods have been proposed for unsupervised contexts (e.g. Trabelsi and Zaiane (2018)). This is important for topics where annotation is costly, or results are time-sensitive. Like with supervised methods, some methods are mainly text-based, (Ghosh et al., 2018; Hardalov et al., 2021; Kawintiranon and Singh, 2021) focusing on the text itself, while others are graph-based, focusing mainly on the users' interaction network (Li and Qi, 2022; Li et al., 2022). Recently, Hofmann et al. (2022) proposed a method integrating both text and graph/network information. Their work focused more on determining polarizing concepts and identifying the source of polarity in communities, which differs from the focus of our work. Additionally, it is noteworthy that this approach builds upon the foundation set by earlier supervised studies (e.g. Mishra et al. (2019)) which have similarly attempted to utilize semantic and social graph information for different applications than stance detection, notably using Graph Neural Networks.

Nevertheless, we show in our experiments that existing methods have the strong disadvantage of being *tailored* to a specific data source, i.e. a particular social media platform such as Twitter, but can hardly generalize. In particular, the phenomena of homophily, where users tend to interact with other users who share their opinions or beliefs, and heterophily, where users interact with those who hold opposing beliefs to their own (McPherson et al., 2001; Albert and Barabási, 2002) are important when considering the source of the data. For instance, platforms such as Twitter tend more towards homophily (Khanam et al., 2023), while other platforms, such as debate forums, fall more on the side of heterophily (Pick et al., 2022). Current methods are designed for one or the other of these network types, raising difficulty when the data does not match the network type that the method was designed for, regardless of data source.

To the best of our knowledge, we are the first to propose an unsupervised stance detection method that leverages both semantic information via text embeddings and network information using a graph neural networks (GNNs) to handle both types of networks (homophilic and heterophilic). This method is trained with a controllable contrastive

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1782–1792

Model	Graph -based?	Uses Text?	Type of Graph	Embedding Method	Clustering Method	Max # Clusters
GUSD	Yes	Yes	Simple	Contrastive learning	K-means	-
InfoVGAE	Yes	Indirectly	Bipartite	VGAE	K-means	-
STEM	Yes	No	Simple	MaxCut SDP	Hyperplane	2
Darwish et al.	No	Indirectly	-	Frequency vectors	Mean Shift	-

Table 1: Characteristics for the four models tested.

setting and, on average, outperforms existing baselines. Note that we focus on predicting the stance at the user level and not on classifying the stances at the document/publication level. Our contributions are three-fold: (1) We provide an analysis of the ability of unsupervised state-of-the-art models to generalize to both network types (heterophilic vs homophilic). (2) We propose a generic baseline for unsupervised stance detection that demonstrates improved resilience to network type variations compared to existing methods. This is achieved by exploiting textual information, along with network information through Graph Neural Networks (GNNs), similar to prior work in NLP detection tasks. (3) We study the effect both of the size of the language model and of the combination of the semantic and network information in this context of unsupervised learning. While not prioritizing outperforming customized approaches, our model performs well on average for both homophilic and heterophilic networks, by exploiting both data modalities (text embeddings and network information).

2 Existing Approaches

Our research focuses on an unsupervised framework, based on textual and/or network data, designed to predict user stance in social media networks. In the following, V denotes the collection of users, the interactions are modeled by a graph G. Additionally, each user is associated to a set of documents they posted online. The nature of G varies according to the framework but the general approach consists of first finding a representation that best describes the behavior of the users based on their interactions and/or the content they published and then to cluster these vector representations of users in such a way that the cluster label assigned to a user corresponds to their predicted stance. As unsupervised approaches are rare, we have chosen three recent methods that can be considered as baselines to test against our approach. Each is built for either homophilic or heterophilic networks, exploits either graph structure or textual data in some way, and has code made available: InfoVGAE (Li et al., 2022), STEM (Pick et al., 2022) and the method from Darwish et al. (2020).

InfoVGAE (Li et al., 2022) utilizes a different graph type than other methods we tested, opting for a bipartite graph $G = (V, T, E)$ with two types of nodes: the users (V) and the tweets (T). There is an edge $(v_i, t_j) \in E$ between a user v_i and a tweet t_j if the user has tweeted t_j. The incidence matrix for this bipartite graph is then the input to a variational graph auto-encoder (VGAE), which aims to recreate this matrix. After training, InfoVGAE uses embeddings from the latent space as input to a K-means algorithm to cluster the users. Thus, InfoVGAE focuses on the user-tweet relationship instead of the textual content of tweets, disregarding semantic information. This model is tailored to perform well on homophilic networks.

STEM (Pick et al., 2022) relies on the assumption that if one user responds to another, they do not hold the same stance. The input graph of STEM is a weighted undirected graph of user interactions where an edge (v_i, v_j) between two nodes v_i and v_j indicates a direct interaction between these two speakers and the weight of this edge corresponds to the number of interactions. The goal of STEM is to find user embeddings that maximize the distance between two users who have interacted and thus have an edge between them. To do so, it first reduces the graph to its 2-core, then solves a relaxation of the max-cut algorithm in order to find the node embeddings. STEM then finds, in this vector space, a random hyperplane that passes through the origin to split the vectors into two opposed groups. The labels obtained for the nodes belonging to the 2-core are then propagated to the nodes that sit outside of the 2-core using the initial assumption that if one user responds to another, they should have opposing stances. Due to its underlying hypothesis, STEM is more particularly dedicated to non-homophilic or even heterophilic networks. In addition, it only applies when the number of stances is limited to two (for and against). Finally, STEM disregards textual information due to its emphasis on structural embeddings.

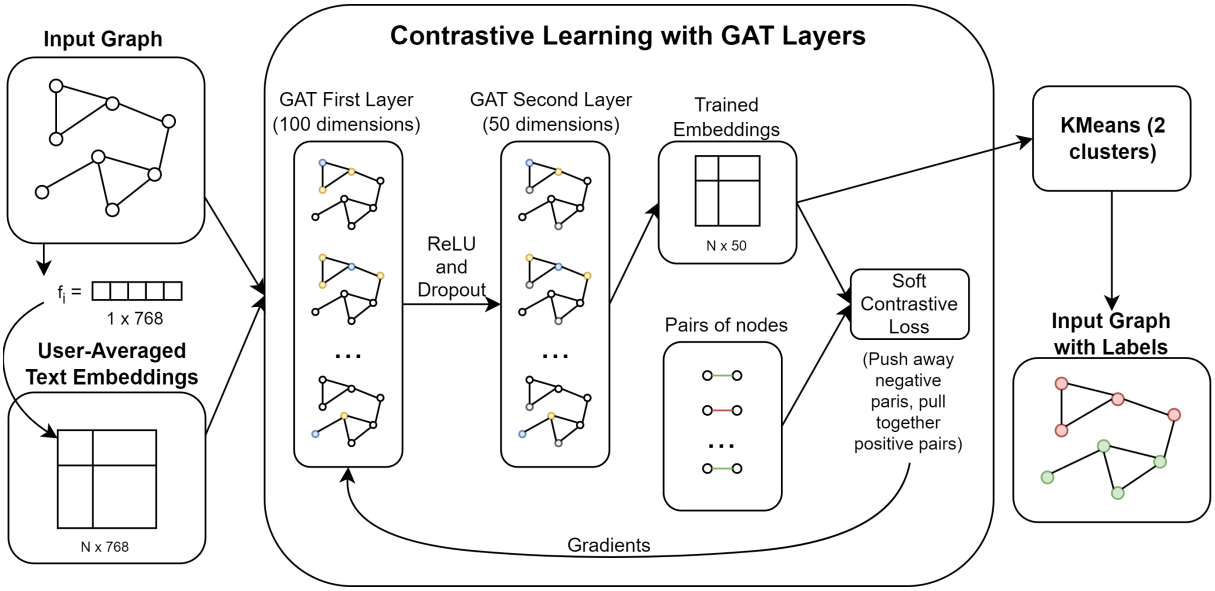

Figure 1: GUSD in one picture.

Darwish et al. (2020) use Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction and Mean Shift for clustering (McInnes et al., 2018). The method generates vectors to describe the users, made up of three sub-vectors: one for the user's tweets, one for the user's mentions of other users, and one for the user's hashtags used. Each vector is created by looking at the corpus of unique tweets, mentions, or hashtags respectively and building a frequency vector for the given user by counting the number of times they have retweeted each tweet or used each mention or hashtag. The feature matrix thus obtained is passed through UMAP to perform dimensionality reduction, then passed to Mean Shift for clustering. It should be noted that while not explicitly built for graph data, this method implicitly uses it via the first sub-vector that contains the same information as the user by tweet portion of InfoVGAE's incidence matrix. It also does not explicitly utilize semantic information, favoring a frequency-based approach to cluster similar uses of hashtags or interactions with tweets and users without specifically analyzing the meaning in the texts or hashtags. Furthermore, Mean Shift being parameter-free, unlike K-means, may not yield the same number of clusters as the number of stances, which is a significant drawback. This method also favors homophilic networks.

3 A Generic Model for Unsupervised Stance Detection (GUSD)

This section presents our model, called GUSD whose architecture is given in Figure 1. We have made the code freely available[1]. In this framework, $G = (V, E)$ is a simple undirected graph and there is a weighted edge $(v_i, v_j) \in E$ between v_i and v_j if they have interacted in some way (retweet, mention, reply, etc) in this network. More precisely, the edge weight is the count of the interactions between the users. Let F be a matrix of size $|V| \times d$ where each row f_i corresponds to a vectorized representation of the user's textual production. We leverage pre-trained language models (encoder Transformers in our experiments) to build F. More precisely, we use an average of the [CLS] token representation of each document (a tweet or a post) produced by a user as done in (Devlin et al., 2019).

We use a graph neural network trained in a self-supervised setting to build informative representations of the users that incorporate both the graph and text information. Specifically, we leverage graph attention networks (GAT) (Veličković et al., 2018). We recall that the computation of the embedding of node v_i in layer l we write $z_i^l \in \mathbb{R}^r$ can be expressed as follows:

$$z_i^l = \sigma\left(\sum_{j=1}^{|V|} \alpha_{ij} W_l z_j^{l-1}\right) \quad (1)$$

with σ an activation function, W_l a learnable

[1] https://github.com/anongusd/GUSD

weight matrix shared across all nodes, and α_{ij} a learned attention coefficient for node v_j with respect to node v_i that captures the importance of neighbors. Finally, the initial representation is the text embedding previously built, so $z_j^0 = f_j$.

As no node annotation is provided in the unsupervised setting, we follow prior works (Hamilton et al., 2017) and use a self-supervised objective to train the node representation. We perform graph reconstruction using a contrastive approach. The aim is to maximize the probability of observed edges and minimize the probability for a set of negative examples (a subset of the unconnected nodes of the graph). We use a soft contrastive loss as presented in (Oh et al., 2019) :

$$\mathcal{L}_{softcon} = \begin{cases} -\log p(m|z_1, z_2) & \text{if } \hat{m} = 1 \\ -\log(1 - p(m|z_1, z_2)) & \text{if } \hat{m} = 0 \end{cases} \quad (2)$$

where $p(m|z_1, z_2) := \sigma(-a||z_1 - z_2||_2 + b)$ is the probability that a pair of nodes has an edge between them, with z the final embeddings (the last GAT layer representation), $a > 0$ and $b \in \mathbb{R}$ provide a soft, trainable threshold for the distance, and $\hat{m}$ is the indicator function being 1 for positive pairs and 0 for negative pairs.

This loss aims to evaluate whether two nodes are likely to be linked. In doing so, the embeddings of linked nodes are pulled towards each other and those of unlinked nodes are pushed away from each other, letting the model integrate information about the interactions between users into the embeddings of their posts.

Note that this self-supervised strategy is quite standard (Oh et al., 2019; Hamilton et al., 2017), therefore not well suited to deal with both homophilic and heterophilic graphs. To circumvent this issue, we propose two versions of this training objective. One fits a homophilic network, while the second can handle a heterophilic network.

For the former, the homophilic version, the positive examples are connected nodes and negative examples are drawn among unconnected pairs of nodes. More precisely, for a positive pair, i.e. an edge between nodes u and v, we draw one random node w that is not connected to u, and add the pair (u,w) to the set of negative pairs. For the latter, the heterophilic version, we build an alternative adjacency matrix A' from the original adjacency A of G. With $c(.)$, the function that changes positive non-zero values to 1 and negative values to 0, we compute $A' = c(AA^T) - c(A)$. In A' there is an edge between two users that interacted with at least one common user (e.g. that debated with the same person in a debate setting). It also removes every direct interaction from the initial graph. This process transforms the debate/heterophilic graph into a homophilic graph. Positive values in A' form the positive examples, while for negative examples we use positive values in $c(A)$ to give a strong signal of users that should be placed apart in the latent space.

After training the model in this self-supervised setting, we perform K-means clustering on the representations z, providing groups of users that we believe cover their opinion proximity and therefore their stance.

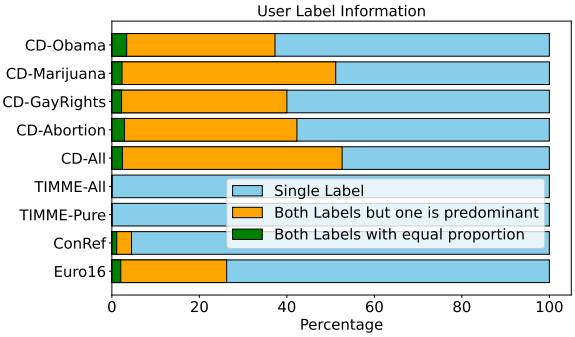

Figure 2: Percent of users with: a single stance in their textual production, several but one predominant, and both stances with equal proportion. There are only a few users with ambiguous positioning w.r.t. the subject at hand.

4 Experimental protocol

The first aim of our experiments is to compare our model, implemented with BERT-large, with the state-of-the-art on various datasets with different characteristics and to show its capacity to adjust to different network types (homophily and heterophily, specifically), confirming that it constitutes a solid baseline for further works. Then, to investigate how the size of a pre-trained language model utilized to represent textual information, affects the performance of our model GUSD, we conducted additional experiments using DistilBERT (Sanh et al., 2020), BERT-base, and BERT-large (Devlin et al., 2019). Finally, we study the interest of combining both text and graph data for unsupervised stance detection. Before presenting the obtained results, we detail our experimental protocol.

Dataset	Source	# Nodes	# Edges	Avg. Tweets Per User	Avg. Mentions Per User	Avg. Hashtags Per User	Dyadicity (Avg Weighted)	Hetero-phillicity
Euro16	Twitter	343	654	5.87	9.14	8.58	1.96	0.11
ConRef	Twitter	178	208	13.40	12.77	16.11	1.66	0.30
TIMME-Pure	Twitter	389	4544	46.76	65.45	36.30	1.91	0.02
TIMME-All	Twitter	942	14558	93.62	123.72	35.84	1.80	0.10
CD-All	CreateDebate	247	724	15.11	0.01	0.15	0.58	1.41
CD-Abortion	CreateDebate	104	245	14.85	0	0.19	0.55	1.43
CD-GayRights	CreateDebate	90	207	10.87	0.03	0.02	0.54	1.62
CD-Marijuana	CreateDebate	43	60	7.47	0	0.02	0.63	1.37
CD-Obama	CreateDebate	59	109	10.83	0	0.22	0.53	1.53

Table 2: Characteristics of the filtered datasets. Nodes and edges are reported for the simple undirected graph.

4.1 Datasets

Our method is devised for datasets with both text produced by users and interactions between them modeled by a graph, demonstrating heterophily or homophily, and for which the users' stance is not available. However its experimental evaluation requires datasets with ground truth i.e. for which the users' stance is known. There are very few publicly available datasets that meet these requirements.

This led us to select commonly used datasets: the homophilic datasets **Euro16** (Li et al., 2022), **ConRef** (Lai et al., 2018) and **TIMME** (Xiao et al., 2020) and one heterophilic dataset **CreateDebate** (CD) (Hasan and Ng, 2014), divided in 5 sub datasets, expanding our data to a total of 9 datasets.

Euro16 (Li et al., 2022) contains Twitter interactions surrounding the controversy over the 2016 Eurovision Song Contest winner, Jamala.

ConRef (Lai et al., 2018) contains data from Twitter, with interactions between users on the 2020 Italian Constitutional Referendum. It is a dataset with large imbalance in favor of the negative stance.

TIMME (Xiao et al., 2020) contains Twitter data from politicians in the United States. TIMME-Pure corresponds to the P_Pure dataset, containing the tweets of only the politicians, while TIMME-All corresponds to the P_All dataset. We use TIMME-All as an augmented version of TIMME-Pure to address the question of noise in the stance label, as it incorporates non-politicians who do not necessarily belong clearly to one of the two primary political parties.

CreateDebate (Hasan and Ng, 2014) contains data from the debate forum CreateDebate on four topics: abortion, gay rights, marijuana, and Obama. We also provide the results for CD-All that mixes all the subjects, which match roughly with political orientation. Therefore, it provides an additional point of view on the ability of the models to separate these orientations even when there is a mix of subjects. All the resulting datasets are heterophilic in nature.

Filtration was done in preprocessing using InfoVGAE's filter that removes users with fewer than three texts and texts with fewer than five keywords. Table 2 presents characteristics of the nine datasets used: number of nodes and edges in the graph, average number of tweets, mentions and hashtags per user.

To evaluate the heterophilic or homophilic tendency of the graph associated to each dataset, we calculated respectively heterophilicity and an average weighted dyadicity (i.e. homophily) from both labels (for and against) (Park and Barabási, 2007). These scores are centered around 1, with scores above 1 indicating respectively heterophilicity or dyadicity (i.e. homophily) and scores below 1 indicating the inverse. According to the scores presented in Table 2, we can consider that the graphs associated to **Euro16**, **ConRef** and **TIMME** datasets are homophilic whereas the graphs generated from **CreateDebate** (CD) are heterophilic.

4.2 Experimental settings

In this section, we describe the various experimental settings used in this paper. Note that there is no proper train/validation/test split, as each method is completely unsupervised.

Comparison to the ground truth We compare our method to baselines in their ability to reconstruct the stance in an unsupervised setting. Thus, the ground truth is only exploited to compute the evaluation metrics: cluster accuracy and F1-scores.

Aggregation of multiple labels As users potentially publish multiple texts and as their posts might be associated with different stances, we determine their overall ground truth stance on the topic by selecting their most frequent stance, as done by (Li et al., 2022; Darwish et al., 2020). This methodological choice is justified by the fact that the majority of users demonstrate a single label

Dataset	GUSD Acc.	GUSD F1	InfoVGAE Acc.	InfoVGAE F1
Euro16	87.75 ± 12.25	87.63 ± 12.59	92.9 ± 1.31	92.69 ± 1.38
ConRef	68.31 ± 6.01	70.5 ± 5.48	53.11 ± 3.11	56.94 ± 2.87
TIMME-Pure	97.89 ± 0.63	97.88 ± 0.63	70.49 ± 1.76	67.69 ± 2.3
TIMME-All	97.26 ± 0.31	97.26 ± 0.3	-	-
CD-All	76.11 ± 1.02	76.23± 1.02	47.3 ± 2.08	41.08 ± 2.89
CD-Abortion	61.34 ± 3.71	61.33 ± 3.78	47.09 ± 2.38	40.49 ± 3.86
CD-GayRights	81.33 ± 2.15	81.99 ± 2.01	43.23 ± 15.6	40.22 ± 16.81
CD-Marijuana	69.76 ± 2.94	70.83 ± 2.82	48.8 ± 14.12	46.64 ± 15.04
CD-Obama	78.31 ± 1.97	78.3 ± 1.96	62.77 ± 2.49	59.13 ± 2.54
Mean score	**79.78** ± 12.68	**80.22** ± 12.35	58.21 ± 16.72	55.61 ± 18.1
Dataset	STEM Acc.	STEM F1	Darwish et al. Acc.	Darwish et al. F1
Euro16	58.89 ± 0.94	59.2 ± 0.99	63.06 ± 3.83	63.18 ± 4.01
ConRef	54.16 ± 0.89	57.77 ± 0.81	94.78 ± 0.27	94.67 ± 0.27
TIMME-Pure	54.32 ± 21.12	53.34 ± 21.31	97.92 ± 0.08	97.92 ± 0.08
TIMME-All	61.21 ± 18.94	56.4 ± 23.19	94.52 ± 0.2	94.53 ± 0.2
CD-All	82.67 ± 1.01	82.73 ± 1.0	56.6 ± 0.46	43.86 ± 0.64
CD-Abortion	75.96 ± 0	75.97 ± 0	54.04 ± 0.88	43.99 ± 0.79
CD-GayRights	85.56 ± 0	85.99 ± 0	-	-
CD-Marijuana	73.02 ± 1.2	73.83 ± 1.15	-	-
CD-Obama	81.19 ± 2.46	81.25 ± 2.38	-	-
Mean score	69.66 ± 12.59	69.61 ± 12.85	76.82 ± 20.97	73.03 ± 25.85

Table 3: Average accuracy and weighted F1 scores and standard deviation (s.d.) with mean score over the datasets given at the bottom. A "-" indicates a technical issue. A zero standard deviation indicates either no change in scores across trials or too small to report. *Homophilic* datasets are on top, *heterophilic* on bottom.

Dataset	DistilBERT Acc.	DistilBERT F1	BERT Base Acc.	BERT Base F1	BERT Large Acc.	BERT Large F1
Euro16	87.58 ± 12.49	87.63 ± 12.50	87.93 ± 11.78	87.84 ± 12.04	87.75 ± 12.25	87.63 ± 12.59
ConRef	71.12 ± 2.06	72.98 ± 1.89	72.58 ± 1.71	74.51 ± 1.43	68.31 ± 6.01	70.50 ± 5.48
TIMME-Pure	95.68 ± 3.01	95.68 ± 3.01	95.52 ± 3.22	95.52 ± 3.22	97.89 ± 0.63	97.88 ± 0.63
TIMME-All	97.24 ± 0.20	97.24 ± 0.20	97.09 ± 0.08	97.08 ± 0.08	97.26 ± 0.31	97.26 ± 0.30
CD-All	73.27 ± 1.71	73.39 ± 1.69	73.68 ± 1.57	73.78 ± 1.55	76.11 ± 1.02	76.23± 1.02
CD-Abortion	63.84 ± 1.88	63.86 ± 1.87	62.30 ± 4.09	62.27 ± 4.05	61.34 ± 3.71	61.33 ± 3.78
CD-GayRights	82.44 ± 3.09	83.04 ± 2.91	77.55 ± 5.32	78.42 ± 5.06	81.33 ± 02.15	81.99 ± 2.01
CD-Marijuana	55.81 ± 2.94	57.40 ± 3.00	61.86 ± 2.94	62.81± 3.32	69.76 ± 2.94	70.83 ± 2.82
CD-Obama	78.30 ± 0.67	78.26 ± 0.69	78.64 ± 0.83	78.61 ± 0.85	78.31 ± 1.97	78.30 ± 1.96
Mean Score	78.36 ± 13.95	78.83 ± 13.53	78.57 ± 12.86	78.98 ± 12.59	79.78 ± 12.68	80.22 ± 12.35

Table 4: GUSD results with text embeddings generated by different sized language models.

in their posts, as shown in Figure 2. Additionally, the percentage of users with the same number of "for" and "against" posts is between 0 and 3.39%. Consequently, even if both labels are present, the user often demonstrates an inclination toward one of the positions.

Settings of baseline methods The baselines were run with their default settings with two exceptions - InfoVGAE was run on 300 epochs instead of 500 as there was no significant difference in results (we observed no changes in accuracy even if the loss tends to slightly decrease). STEM had the option of agreeing propagation as well as opposing and the best result is reported. The K-means algorithm for InfoVGAE and our model was run with $k = 2$ to produce two clusters and the data was filtered to contain only entries labeled with the binary labels to compensate for STEM's binary constraint. Due to Mean Shift's lack of parameter to control the number of clusters, the method from Darwish et al. was run a maximum of 500 times and the first 10 partitions giving 2 clusters were taken as the 10 trials.

We did not apply the adjustment of the adjacency matrix to baseline methods ($A \rightarrow A'$ for heterophilic graphs). These baselines were not conceived to use the adjacency matrix as input. InfoVGAE constructs its own bipartite heterogenous information network between users and posts. STEM's objective function is built under the assumption of opposition in interactions, while the method from Darwish et al. (2020) does not directly use user interaction data in a graph format. Integrating this modified adjacency matrix would

Dataset	A	A'
CD-All	0.52	0.76
CD-Abortion	0.50	0.61
CD-GayRights	0.53	0.81
CD-Marijuana	0.61	0.70
CD-Obama	0.58	0.78

Table 5: Accuracy comparison when using A or A' for GUSD on the CreateDebate datasets.

therefore lead to a significant modification of the baselines.

Settings of GUSD[2] We use a two-layer GAT with standard hyperparameters : hidden dimensions of 100 and 50, ReLU activation function on the first layer, along with a 20% dropout layer between the two GAT layers.

The contrastive self-supervised training phase is run for 10,000 epochs with early stopping based on the graph reconstruction score on a validation set, after which the generated embeddings are passed through the K-means algorithm to generate the two clusters of stances (more precisely, we select the epoch that minimizes the inertia of the K-means output among the 100 last steps).

We used a standard hyperparameter set for the GNN architecture, which is common for all the datasets, so there is no hyperparameter tuning (number of layers, learning rate, etc.). Second, we used early stopping and a test set composed of pairs of nodes, linked or not linked, where the edges have been hidden during the training step as done in the literature. Concretely, considering the self-supervised training of the GNN, we used 5% of the links (that were hidden in the training set) as evaluation set to compute the convergence criterion: when the accuracy of the link prediction stops increasing on this set, we stop the optimization phase. We believe that these will prevent overfitting and provide a fair setting for unsupervised evaluation of the methods (closer to the real life setting).

Table 1 presents a summary of model characteristics as an overview of the four models tested.

5 Results

Table 3 details the clustering accuracy and the weighted F1 score for each model on each dataset, averaged across 10 runs. Furthermore, the overall mean and the standard deviation of these metrics, computed across the nine datasets, are reported. It should be noted that a high standard deviation associated with this overall average means a high variability of the scores obtained by a model over the different datasets.

As shown by the results, the baselines have specific contexts where they perform best, but outside of those contexts they often perform on par with random chance or are unable to produce results. InfoVGAE struggles with heterophilic networks, as it uses graph convolutions on the bipartite graph and thus aggregates up to the two-hop neighbors - users that shared the same content. It performs best on homophilic networks with a higher percentage of retweets because of its use of a bipartite graph and graph convolutions. This is because a graph with many unique texts (meaning texts that have been posted by a single user without any retweets, such as with CreateDebate) will produce a less connected graph and lower the model's effectiveness. STEM, on the other hand, was built for heterophilic interaction networks as confirmed by its poor performance on the homophilic datasets. Some attempts were made to adjust STEM to homophilic data, such as adjusting the max cut objective, however these tests were unsuccessful. The method from Darwish et al. works best on homophilic Twitter data that can build more informative frequency vectors, meaning that the average tweets, hashtags, or mentions per user are relatively high, such as in datasets ConRef and TIMME (see data statistics in Table 2). This helps the UMAP low-dimension vectors to be more effective in clustering the users.

While GUSD does not outperform every model on every dataset, that was not its goal but it does outperform the other models on average across the datasets. Because it can be run on both homophilic and heterophilic interaction networks, it is able to adjust to the needs of these two different types of networks, where assumptions made by other models do not hold across both types. Additionally, unlike the other models GUSD runs without any issues on all the datasets and is less variant to the additional noise that the TIMME-All dataset contains in comparison to TIMME-Pure, which is filtered to only data on politicians.

5.1 Impact of accounting for the heterophilic nature of a network with GUSD

In Table 5, we provide the results of an evaluation of the impact of using the original adjacency matrix A, or the re-weighted one A' for our model GUSD. We recall that the transformation $A \rightarrow A'$

[2]https://github.com/anongusd/GUSD

Dataset	DistilBERT		BERT Base		BERT Large	
	Acc.	F1	Acc.	F1	Acc.	F1
Euro16	54.52	54.96	60.93	61.72	51.60	52.63
ConRef	77.53	68.26	69.10	71.15	55.61	59.11
TIMME-Pure	66.32	65.88	84.57	84.47	87.14	87.13
TIMME-All	65.50	65.64	65.92	66.03	62.63	62.80
CD-All	53.03	53.13	55.06	55.00	55.06	55.28
CD-Abortion	55.76	55.81	54.80	54.53	63.46	63.46
CD-GayRights	60.00	59.01	63.33	62.16	52.22	53.86
CD-Marijuana	58.13	58.13	65.11	62.30	60.46	58.82
CD-Obama	55.93	55.83	50.84	49.38	54.23	39.33
Mean Score	60.75 ± 7.82	59.63 ± 5.54	63.3 ± 9.98	62.97 ± 10.37	60.27 ± 10.97	59.16 ± 12.71

Table 6: Results for text embeddings from different sized language models, without inclusion of graph data.

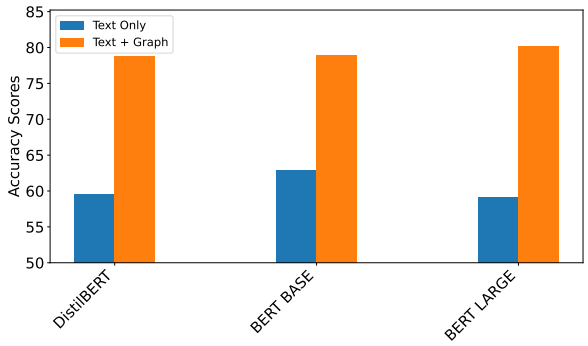

Figure 3: Averaged results (weighted F1 score) over all datasets for different sized language models and impact on using the graph data

Dataset	GUSD	InfoVGAE	STEM	Darwish
Euro16	94	119	161	258
ConRef	75	252	24	4
TIMME-P	891	231	1121	17
TIMME-A	3608	-	5803	345
CD-All	339	340	805	9
CD-AB	129	141	45	4
CD-GR	85	106	15	-
CD-MA	43	67	1	-
CD-OB	39	85	5	-

Table 7: Timing for a single trial on each dataset for each model, in seconds. Results for GUSD are calculated with DistilBERT.

allows for the transformation of a heterophilic adjacency matrix into a homophilic one, as explained in Section 3. The accuracy reported in this experiment clearly demonstrates that this transformation improves the stance detection results by 34% on average on the CreateDebate datasets. Additionally, we noted that before transformation, the average dyadicity of CD datasets is 0.57 (see Table 2). After transformation, the average dyadicity reaches 0.97, demonstrating that this process increases the connectivity between users with similar stance.

5.2 Impact of pre-trained model size for textual representation with GUSD

In Figure 3, we present details on the effect of model size on GUSD. We used three model sizes: BERT-base is two times larger than DistilBERT and two times smaller than BERT-large. When varying the textual representations issued from the different size versions of BERT, we did not observe a substantial improvement in the performances: DistilBERT (smallest) had 78.83% weighted F1-score averaged across all datasets, while BERT-base had 78.98%, and BERT-large a small increase at 80.22%. We provide full results of this experiment in Table 4.

5.3 Impact of text embedding quality and interest of combining text and graph data

Figure 3 compares the results using only the encoder output (text only without interaction data) with those obtained by combining graph and text data. The results are presented with the same varying model sizes as seen in the above subsection.

It is worth noting that when we directly passed the embeddings to the K-means algorithm without utilizing the contrastive graph learning component of the model, BERT-base embeddings achieve the highest averaged F1-score (62.97%). In contrast, using BERT-large resulted in 59.16%, and DistilBERT yielded 59.63% (cf. Table 6). These figures clearly indicate lower performance compared to when we incorporate graph learning and interaction data. Full results can be seen in Table 6.

5.4 Timing of all models on all datasets

Table 7 presents the time (in seconds) of a single trial of each model. All trials to measure the timing of the models were performed on a PC with 4 GB of GPU, an AMD Ryzen 7 5800H CPU with 8 cores and 16 threads, and 16GB of RAM. GUSD computation time seems to be impacted by the den-

sity of the network, similarly to STEM. We observe a tenfold increase in running time when transitioning from Euro16 to TIMME-Pure, which mainly differs in the number of edges. However, excluding TIMME-All, it is faster than InfoVGAE. It is also relatively equivalent to STEM, although GUSD is less affected by an increase in network density.

6 Conclusion

While unsupervised models do exist for stance detection, they struggle to generalize to network types that do not hold to the assumptions the methods are based on. As such, none of them can act as baselines across multiple datasets with opposing characteristics of homophily or heterophily. To that end, we have conducted an analysis of state-of-the-art models on varying datasets and proposed a new baseline model. Unlike existing work, it exploits both text content and graph structure by using text embeddings propagated via graph neural networks, which makes it more generalizable to different network types. GUSD outperforms the other unsupervised models on average and is robust to changes in the number of parameters of the language model used to construct the text embeddings.

Limitations

We tested a wide range of metrics in the final representation space to serve as a surrogate for expert knowledge to determine whether the network is heterophilic or homophilic. The tested metrics included inertia, the Calinski-Harabasz score, the silhouette score, and the Davies-Bouldin index. While some scores showed promise, none were correlated with the final stance detection accuracy for either heterophilic or homophilic interaction networks. This is still an open research question for the community. Fortunately, the nature of the discussion in most broadly used social media platforms is known.

In this work, we only consider binary stance detection. This choice was motivated by the fairness of evaluation compared to competitors. Among the baselines, STEM can only handle the binary case and InfoVGAE has also only been evaluated for this case. As such, we chose to follow these previous works. Moreover, some of the data used, notably CreateDebate, contains only binary annotations. Expanding past the binary case would have required us to collect and annotate additional data to test the heterophilic case. This was not done for experimental reasons, however our model can be used as-is in the case of non-binary data.

Ethics Statement

Media opinions do not necessarily reflect votes (Lai et al., 2018) so the information provided by the model cannot be taken as certainty without considerations. These include which communities are involved, which communities are likely to be vocal on social media versus participate in a vote, and how the data is being collected. The model is only meant to give a rough idea of people's opinions based on the data, so if the data is biased, the model's results will reflect that bias.

This model is meant to be a baseline for further research, not for direct application use. It does not use demographic or identity information and even the identifiers it does use (usernames) can be anonymized without any effect on the model's results. The anonymization could also be important to mitigate malicious use to attack users who are detected to hold a certain belief (though it is important to note that due to the model using publicly posted tweets/texts, this is not dependent on the model itself). All data and code used in this paper has been made publicly available [3] [4] [5] [6]. Table 2 gives characteristics of the data and Table 1 provides characteristics of the models.

Acknowledgements

This work was partially supported by the MANUTECH-Sleight Graduate School thanks to a public grant from Saint-Etienne Métropole, by the CNRS IEA project CODANA, and by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant No. RGPIN-2022-04789.

References

Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised User Stance Detection on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:141–152.

[3] https://github.com/anongusd/GUSD
[4] https://github.com/jinningli/InfoVGAE
[5] https://github.com/NasLabBgu/STEM
[6] https://github.com/elaaf/stance-detect

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Subrata Ghosh, Konjengbam Anand, Sailaja Rajanala, A Bharath Reddy, and Manish Singh. 2018. Unsupervised stance classification in online debates. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 30–36. ACM.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-Domain Label-Adaptive Stance Detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 751–762.

Valentin Hofmann, Xiaowen Dong, Janet Pierrehumbert, and Hinrich Schuetze. 2022. Modeling Ideological Salience and Framing in Polarized Online Groups with Graph Neural Networks and Structured Sparsity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 536–550. Association for Computational Linguistics.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge Enhanced Masked Language Model for Stance Detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735. Association for Computational Linguistics.

Kazi Zainab Khanam, Gautam Srivastava, and Vijay Mago. 2023. The homophily principle in social network analysis: A survey. *Multimedia Tools and Applications*, 82(6):8811–8854.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance evolution and twitter interactions in an italian political debate. In *International Conference on Applications of Natural Language to Data Bases*.

Christine Largeron, Andrei Mardale, and Marian-Andrei Rizoiu. 2021. Linking the dynamics of user stance to the structure of online discussions. In Pedro Henriques Abreu, Pedro Pereira Rodrigues, Alberto Fernández, and João Gama, editors, *Advances in Intelligent Data Analysis XIX*, pages 275–286. Springer International Publishing.

Jinning Li, Huajie Shao, Dachun Sun, Ruijie Wang, Yuchen Yan, Jinyang Li, Shengzhong Liu, Hanghang Tong, and Tarek Abdelzaher. 2022. Unsupervised Belief Representation Learning with Information-Theoretic Variational Graph Auto-Encoders. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1728–1738.

Yang Li and Rui Qi. 2022. Heterogeneous Graph Contrastive Learning for Stance Prediction. *IEICE Transactions on Information and Systems*, E105.D(10):1790–1798.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. JointCL: A Joint Contrastive Learning Framework for Zero-Shot Stance Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 81–91. Association for Computational Linguistics.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29).

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Review of Sociology*, 27:415–444.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Abusive Language Detection with Graph Convolutional Networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2145–2150, Minneapolis, Minnesota. Association for Computational Linguistics.

Seong Joon Oh, Andrew C. Gallagher, Kevin P. Murphy, Florian Schroff, Jiyan Pan, and Joseph Roth. 2019. Modeling uncertainty with hedged instance embeddings. In *International Conference on Learning Representations*.

Juyong Park and Albert-László Barabási. 2007. Distribution of node characteristics in complex networks. *Proceedings of the National Academy of Sciences*, 104(46):17916–17920.

Ron Korenblum Pick, Vladyslav Kozhukhov, Dan Vilenchik, and Oren Tsur. 2022. STEM: Unsupervised STructural EMbedding for Stance Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11174–11182.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *EMC2: 5th Edition Co-located with NeurIPS'19*.

Qingying Sun, Xuefeng Xi, Jiajun Sun, Zhongqing Wang, and Huiyan Xu. 2023. Stance Detection with a Multi-Target Adversarial Attention Network. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–21.

Amine Trabelsi and Osmar Zaiane. 2018. Unsupervised model for topic viewpoint discovery in online debates leveraging author interactions. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

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*, pages 2258–2268. ACM.

Chong Zhang, Zhenkun Zhou, Xingyu Peng, and Ke Xu. 2023. DoubleH: Twitter User Stance Detection via Bipartite Graph Neural Networks. ArXiv:2301.08774 [cs].

Qinglin Zhu, Bin Liang, Jingyi Sun, Jiachen Du, Lanjun Zhou, and Ruifeng Xu. 2022. Enhancing Zero-Shot Stance Detection via Targeted Background Knowledge. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2070–2075. ACM.