Improving (Dis)agreement Detection with Inductive Social Relation Information From Comment-Reply Interactions

Yun Luo
School of Engineering, Westlake University
Hangzhou, China
luoyun@westlake.edu.cn

Zihan Liu
AI Lab, Research Center for Industries of the Future,
Westlake University
Hangzhou, China
liuzihan@westlake.edu.cn

Stan Z. Li
AI Lab, Research Center for Industries of the Future,
Westlake University
Hangzhou, China
stan.zq.li@westlake.edu.cn

Yue Zhang*
School of Engineering, Westlake University
Hangzhou, China
zhangyue@westlake.edu.cn

ABSTRACT

(Dis)agreement detection aims to identify the authors' attitudes or positions (*agree, disagree, neutral*) towards a specific text. It is limited for existing methods merely using textual information for identifying (dis)agreements, especially for cross-domain settings. Social relation information can play an assistant role in the (dis)agreement task besides textual information. We propose a novel method to extract such relation information from (dis)agreement data into an inductive social relation graph, merely using the comment-reply pairs without any additional platform-specific information. The inductive social relation globally considers the historical discussion and the relation between authors. Textual information based on a pre-trained language model and social relation information encoded by pre-trained RGCN are jointly considered for (dis)agreement detection. Experimental results show that our model achieves state-of-the-art performance for both the in-domain and cross-domain tasks on the benchmark – DEBAGREEMENT. We find social relations can boost the performance of the (dis)agreement detection model, especially for the long-token comment-reply pairs, demonstrating the effectiveness of the social relation graph. We also explore the effect of the knowledge graph embedding methods, the information fusing method, and the time interval in constructing the social relation graph, which shows the effectiveness of our model.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; *Information extraction*; *Semi-supervised learning settings*; • **Information systems** → *Social networks.*

*Corresponding author

WWW '23, May 1–5, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583314

KEYWORDS

Stance Detection, Disagreement Detection, Opinion Mining, Social Relation

ACM Reference Format:
Yun Luo, Zihan Liu, Stan Z. Li, and Yue Zhang*. 2023. Improving (Dis)agreement Detection with Inductive Social Relation Information From Comment-Reply Interactions. In *Proceedings of the ACM Web Conference 2023 (WWW '23), May 1–5, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3543507.3583314

1 INTRODUCTION

Automatic elicitation of semantic information has attracted increasing attention in recent years to the widespread social platforms online. The tasks contain sentiment analysis, sarcasm detection, stance detection, etc. We focus on the task of (dis)agreement detection, which aims to identify the authors' attitudes or positions (*agree, disagree, neutral*) towards a specific text [33]. This task falls under the field of stance detection and opinion mining. For example, to the text '*Peace is sometimes a translation of Shalom, which also carries the meaning of wellbeing. It speaks to the heart of what Peace is about.*', the reply '*Someone explains to me how climate activism relates to peace I feel like it's a bit unrelated*' expresses a disagreeing stance. The task of (dis)agreement detection is crucial in understanding the societal polarisation and spread of ideas online [36, 37, 40].

There are several challenges for (dis)agreement detection. One salient challenge is that the textual information is limited for the task [33], and when a human detects the (dis)agreement of a reply to a specific comment, some commonsense knowledge or contextual understanding assists in the identification. Taking the first example in Figure 1, the comment talks about leaders' mathematical and philosophical training. However, the reply is about the importance of scientists, not politicians, which has different textual features. It is difficult for models to correctly identify the (dis)agreement solely based on the textual features. In addition, it remains challenging to detect (dis)agreements in cross-domain settings [1, 33], where the topics or contents of testing data are different from the training data. The language expressions for stance-related text usually vary across different topics. Suppose the (dis)agreement detection model is trained on the data of Republican, like the example in Figure

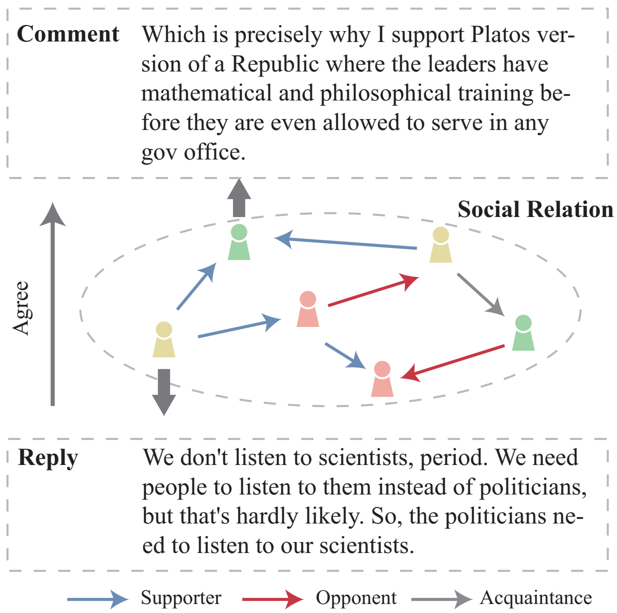

Figure 1: Examples for (dis)agreement detection in the DEBAGREEMENT dataset.

1, but tested on the topic of Climate. In that case, models can be confused in giving identification.

It has been identified in sociology and psychology studies that individuals' opinions are significantly affected by social relations and online contents [9, 18, 27]. Accordingly, some studies use contextual information from Twitter, such as hashtags or retweets, to solve the challenges mentioned above [14, 38]. For example, Dey et al. [14] propose a latent concept space to obtain the stance similarity using twitter hashtags for identifying stance. Samih and Darwish [38] propose a classification method based on the accounts he/she retweeted, computing its similarity to all users in the training set. Nevertheless, the features limit the extensibility of the models, which most datasets or platforms lack. In addition to this specific information, individuals' relations can also be reflected by the interactions between them, which is common, and easily obtained in most scenarios. However, relatively little work applies the information due to the lack of suitable datasets.

Recently, Pougué-Biyong et al. [33] propose a large dataset in real-world online discussions (42,804 comment-reply pairs) on Reddit[1], which contains information of authors and the temporal order (common information on most social platforms). The dataset provides a testbed for investigating general social information's effect and how it enhances (dis)agreement detection models. In particular, the dataset contains contextual information (authorship, post, timestamp, etc.) and comment-reply pairs for (dis)agreement detection. We reform the comment-reply pairs to a social relation graph to detect (dis)agreement with social information, which facilitates the (dis)agreement detection. For the examples in Figure 1, if social relation information is effectively used so that the model knows that the authors of the replies are supporters of comment authors (obtained from previous interactions), it becomes simple to identify the (dis)agreement of the comment-reply pairs. In addition, for the cross-domain setting, the model can be more accurate with the use of the social relation information, where implicit relations such as 'A friend of my enemy is my enemy' [7] can also be effectively used by message passing from graph neural networks.

Individuals tend to maintain their initial beliefs even in the face of evidence that contradicts them, which is called belief perseverance in psychology [3, 16]. Thus, individuals tend to insist on their (dis)agreement with others on a specific issue. Inspired by the effectiveness of graph neural networks in extracting the representation of structured data [19, 26, 44], we propose a novel method to extract social relations from temporal comment-reply interactions to an inductive social relation graph, which gives general information on different social platforms and offline scenarios. We pre-train a graph autoencoder to encode social relation information through a relational graph network (RGCN) [39] encoder and a knowledge graph embedding (KGE) decoder DistMult [45]. The social relations information encoded by the graph autoencoder is fused with textual information from pre-trained language models such as BERT [21], and RoBERTa [25] to identify the (dis)agreement.

Experiments show that our model achieves state-of-the-art performance in in- and cross-domain settings for (dis)agreement detection on the standard benchmark [33]. We prove the effectiveness of social relation features on BiLSTM, BERT, and RoBERTa for (dis)agreement detection. Then we demonstrate that the model performs better for long-token comment-reply pairs. We also show the significance of each module in our model, such as the reconstruction loss of graph features and the pre-training of the graph autoencoder. To the best of our knowledge, we are the first to consider the general inductive social relation information from comment-reply pairs for (dis)agreement detection. The codes and trained models can be found at https://github.com/LuoXiaoHeics/StanceRel.

The contributions of our paper can be summarized as follows:

(1) We propose a novel method to extract relation information from (dis)agreement data into an inductive social relation graph, merely using the comment-reply pairs without any additional platform-specific information.
(2) We propose a (dis)agreement detection model jointly considering the textual information from pre-trained language models and social relation information from pre-trained RGCN.
(3) Experimental results show that our model achieves state-of-the-art performance for both the in-domain and cross-domain tasks. We also show the effectiveness of our models through various analyses.

2 RELATED WORK

(Dis)agreement detection is a sub-task of stance detection [24, 30, 33], (also known as stance classification [42], stance identification [47], stance prediction [34], debate-side classification [2], and debate stance classification [17]). Many models are proposed to solve the task of stance detection or (dis)agreement detection by solely using textual information, while some studies have used graph (or network) features to boost the performance of stance detection or (dis)agreement detection, such as interaction networks, preference networks, and connection networks. Borge-Holthoefer et al. [6], Darwish et al. [12] and Darwish et al. [13] propose to use the

[1]reddit.com: the 20th most visited site globally as of March 2020

relations of retweet data. Dey et al. [14] and Samih and Darwish [38] make use of hashtags to infer Twitter users' stances.

Existing work also considers incorporating social context [22] and structured knowledge [11] into language models to boost the performance on natural language processing tasks. However, previous datasets on stance detection mostly merely provide textual information. Some work that uses graph features is specific to a Twitter discussion, using the hastags or retweets, which limits the extensibility of the model. Unlike previous studies, we propose a simple method to construct the social relation graph using the (dis)agreement data with authors and temporal orders, which are common information in most social platforms or debate situations, boosting the extensibility of using social relation graph for disagreement detection.

We use knowledge graph embedding (KGE) methods to pre-train the node embeddings of the authors, which are widely used for encoding knowledge graph concepts. KGE methods can effectively simplify the manipulation while preserving the knowledge graph's inherent structure and achieving remarkable performance in the downstream tasks such as knowledge graph completion, and relation extraction [5, 31, 45]. Prior work can be divided into translational distance models using distance-based scoring functions and semantic matching models using similarity-based ones [43]. In this paper, we use the idea of knowledge graph embedding to pre-train a graph autoencoder to extract social information, assisting (dis)agreement detection.

3 METHOD

The architecture of our model is illustrated in Figure 3, which contains two components: (1) relation graph encoding, which extracts social relation information (Section 3.2); (2) (dis)agreement detection with relation information (Section 3.3).

3.1 Task Description

We formulate the (dis)agreement detection task as a classification task. Formally, let $D = \{c_i, t_i, y_i, n_i^c, n_i^t\}_{i=1}^N$ be a dataset with N examples, each consisting of a comment c_i from author n_i^c, a reply r_i from author n_i^t, and a stance label y_i from r_i to n_i^c through the comment-reply pair. The task is to predict a stance label $\hat{y} \in \{agree, disagree, neutral\}$ for each comment-pair, based on the definition of Pougué-Biyong et al. [33].

3.2 Relation Graph Autoencoder

We denote the relation graph as a directed graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}, \mathcal{R}\}$, with nodes (authors) $n_i \in \mathcal{N}$ and labeled edges $(n_i, r, n_j) \in \mathcal{E}$, where $r \in \mathcal{R}$ is the relation type of the edge from n_i to n_j. The relation types include {*supporter, opponent, acquaintance, interaction*}.

Social Relation Graph Construction. To construct the relation graph, we first extract the set of all authors in the dataset, corresponding to the node set $\mathcal{N}$. The time interval to aggregate the social relations is a significant factor due to the temporal effects of social relations between individuals. Inspired by [20, 28], we model the temporal network by weighting the links with frequencies to obtain the type of social relation. For the sequence of the graph weighted adjacent matrix (snapshots) $S(w, \tau) = [\mathcal{A}^0, \mathcal{A}^1, ..., \mathcal{A}^{(w-1)}]$ ($\mathcal{A}^k$ is the graph weighted adjacent matrix during time period $[k\tau, (k+1)\tau]$), the inductive graph is drawn from the interactions that appear during the timescale $w\tau$ (Figure 2). For each graph weighted adjacent matrix $\mathcal{A}^k$, if the author n_i expresses an agree/disagree/neutral stance towards n_j in a comment-reply pair, the value $a_{i,j}^k = +1/-1/0$, and if there are multiple interactions between them, the most frequent opinion (agree/disagree/neutral) are considered to determine $a_{i,j}^k$. Then the weighted adjacent matrix of the inductive graph is $\mathcal{A}^* = \mathcal{A}^0 + \mathcal{A}^1 + ... + \mathcal{A}^{(w-1)}$. The triplet (n_i, r, n_j) is in relation graph $\mathcal{G}_R$ as follows:

$$r = \begin{cases} supporter & \text{if } a_{ij}^* > 0, \\ opponent & \text{if } a_{ij}^* < 0, \\ acquaintance & \text{if } a_{ij}^* = 0 \text{ and } a_{ij}^k \neq 0, \end{cases}$$

and (n_i, r, n_j) is not in $\mathcal{G}_R$ in other situations.

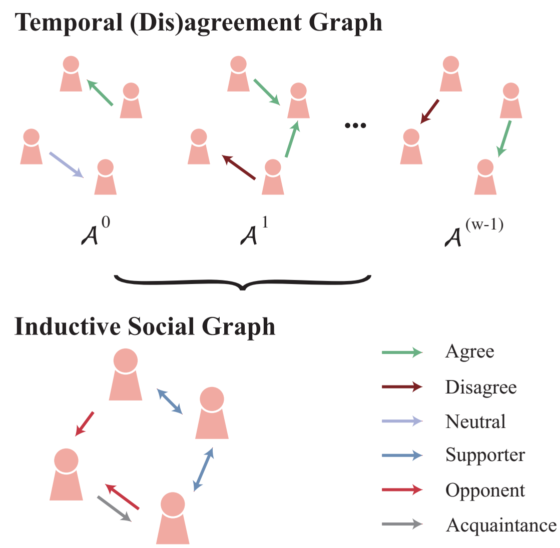

Figure 2: The illustration of the construction of the social relation graph using the temporal order information.

In order to avoid label leaking in development and test sets, we add another type of relation *interaction* for the edges unseen in the training set but appear in the development and test sets. The node feature can be normally observed in the semi-supervised learning on graph neural network tasks [10, 41]. It also aims to solve the issue that node features would be unknown if the nodes are not added to the social relation graph before pre-training. To avoid the over-fitting of training the model, we randomly select edges in $\mathcal{G}_R$ with probability ρ to be *interaction* edges.

To obtain the social relation information, a graph autoencoder is adopted following Schlichtkrull et al. [39]. An incomplete set (randomly sampled with 50% probability) of edges $\hat{\mathcal{E}}$ from $\mathcal{E}$ in $\mathcal{G}_R$ is fed as the graph autoencoder input. The incomplete set $\hat{\mathcal{E}}$ is negatively sampled to a complete set of samples denoted $\mathcal{U}$ (details in Training). Then we assign the possible edges $(n_i, r, n_j) \in \mathcal{U}$ with scores, which are used to determine the probability of whether the edges are true in $\mathcal{E}$. Relational author network (RGCN) [39] is applied to the encoder to obtain the latent feature representations of authors, and a DistMult scoring decoder [45] is used to recover the missing edges.

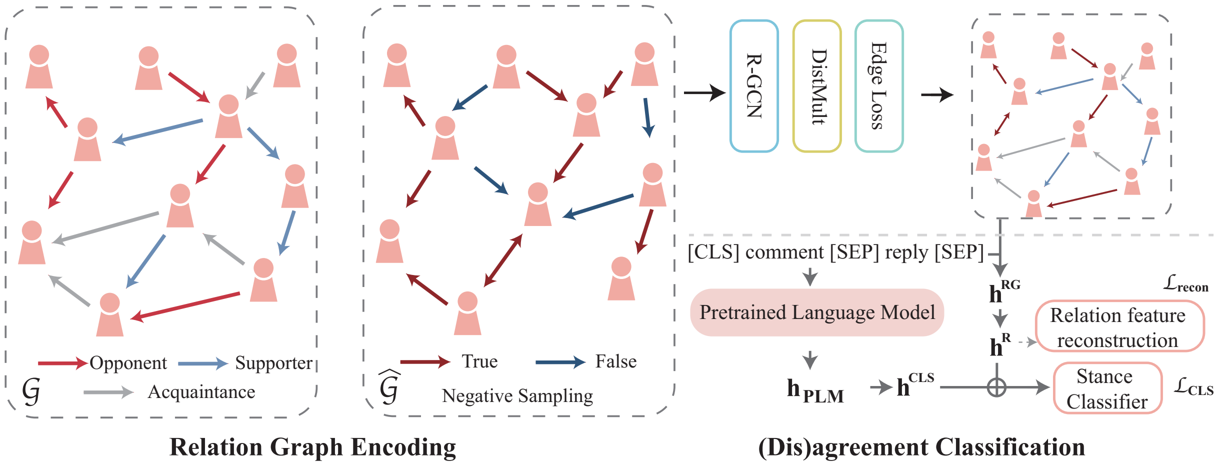

Figure 3: Framework of our proposed model, which contains two components, (1) relation graph encoding, (2) (dis)agreement detection with social relation information.

Encoder. The RGCN module serves to accumulate relational evidence in multiple inference steps. In each step, a neighborhood-based convolutional feature transformation process uses the related authors to induce an enriched author-aggregated feature vector for each author. Two stacked RGCN encoders are applied to encode the social information. The parameters of author feature vectors are initialized with $\mathbf{u}_i$. Then the vectors are transformed into relation-aggregated feature vectors $\mathbf{h}_i \in \mathbb{R}^d$ using the RGCN encoders:

$$f(x_i, l) = \sigma(W_0^{(l)} x_i + \sum_{r\in\mathcal{R}} \sum_{j\in N_i^r} \frac{1}{n_{i,r}} W_r^{(l)} x_j),$$
$$\mathbf{h}_i = \mathbf{h}_i^{(2)} = f(\mathbf{h}_i^{(1)}, 2)\ ;\ \mathbf{h}_i^{(1)} = f(\mathbf{u}_i, 1), \tag{1}$$

where f is the encoder network (requiring inputs of feature vector x_i and the rank of the layer l); N_i^r is the neighbouring authors i with the relation $r \in \mathcal{R}$; $n_{i,r}$ is a normalization constant, set in advance $n_{i,r} = |N_i^r|$ or learned by network learning; σ is the activation function such as ReLU, and $W_r^{(1)}, W_0^{(1)}, W_r^{(2)}, W_0^{(2)}$ are learnable parameters though training.

Training. We use DistMult factorization as the decoder to assign scores. For a given triplet (n_i, r, n_j), the score can be obtain as follows:

$$s(n_i, r, n_j) = \sigma(\mathbf{h}_{n_i}^T R_r \mathbf{h}_{n_j}), \tag{2}$$

where σ is a logistic function; $\mathbf{h}_{n_i}, \mathbf{h}_{n_j} \in \mathbb{R}^n$ are the encoding feature vectors through the graph encoder for author n_i and n_j; every type of relation $r \in \mathcal{R}$ is associated with a diagonal matrix $R_r \in \mathbb{R}^{n\times n}$.

The method of negative sampling [39] is used for training our graph autoencoder module. First, we randomly corrupt the true triplets, i.e., triplets in $\hat{\mathcal{E}}$, to create an equal number of false samples. We corrupt the triplets by randomly modifying the connected authors or relations, creating the overall set of samples $\mathcal{U}$. The training objective is a binary classification between true/false (denoted as y) triplets with a cross-entropy loss function:

$$\mathcal{L}_{\mathcal{G}'} = -\frac{1}{2|\hat{\mathcal{E}}'|} \sum_{(n_i,r,n_j,y)\in\mathcal{U}} (y\, log\, s(n_i, r, n_j) + (1-y) log(1 - s(n_i, r, n_j))). \tag{3}$$

3.3 (Dis)agreement Detection

Relation Feature Encoding. After training the graph autoencoder, in order to extract the author-specific relation graph feature for the comment c_i and the reply t_i, we denote n_c^i and n_t^i as the authors for c_i and t_i, respectively. Then we extract a sub-graph $\mathcal{G}_A$ from $\mathcal{G}_R$, which contains all the authors on the graph within the vicinity of radius 1 from n_c^i and n_t^i. Next, we make a forward pass of $\mathcal{G}_A$ through the encoder of graph autoencoder to obtain the feature vectors $\mathbf{h}_j$ for all unique authors j in $\mathcal{G}_A$. The average of feature vectors $\mathbf{h}_j$ for all unique authors in $\mathcal{G}_A$ is regarded as the relation graph feature vector $\mathbf{h}^{RG}$:

$$\mathbf{h}^{RG} = RGCN(\mathcal{G}_A). \tag{4}$$

The relation graph feature vector $\mathbf{h}^{RG}$ is fed into a linear layer to obtain hidden states $\mathbf{h}^R$:

$$\mathbf{h}^R = W_R\, \mathbf{h}^{RG} + b_R \tag{5}$$

where W_R and b_R are the trained parameters of the linear layer.

Textual Feature Encoding. Pre-trained language models (PLMs), such as BERT[21], RoBERTa [25], and GPT3 [35], have been proven effective in various NLP applications, which are pre-trained on the large-scale unlabelled corpus. Taking BERT, for example, it uses a bidirectional transformer on single or multiple sentences. We take $[CLS]\ c_i\ [SEP]\ \ t_i\ [SEP]$ as the input x_i for our model, where $[CLS]$ refers to the first token of the sequence and $[SEP]$ is used to separate sequences. The input x_i is fed into PLMs such as BERT and RoBERTa to obtain its hidden states:

$$\mathbf{h}_{PLM} = PLM(x_i). \tag{6}$$

The hidden state of $[CLS]$ token is adopted as the representation of the comment-reply pairs.

(Dis)agreement Classification. The hidden states vectors of $\mathbf{h}^R$ and $\mathbf{h}^{CLS}$ are concatenated for classification:

$$p = Softmax(W[\mathbf{h}^{CLS}, \mathbf{h}^R] + b), \tag{7}$$

where W and b are the parameters and p is the probability distribution on the three (dis)agreement labels.

	r/Br	r/Cl	r/BLM	r/Re	r/De
#nodes	722	4,580	2,516	8,832	6,925
#edges	15,745	5,773	1,929	9,823	9,624
Agree	29%	32%	45%	34%	42%
Neutral	29%	28%	22%	25%	22%
Disagree	42%	40%	33%	41%	36%

Table 1: Statistics on DEBAGREEMENT. Br for the subreddit Brexit; Cl for the subreddit Climate; BLM for the subreddit BLM; Re for the subreddit Republican and De for the subreddit Democrats, henceforth.

	r/Br	r/Cl	r/BLM	r/Re	r/De	All
#Supporter	2,159	989	511	1,882	2,299	7,833
#Opponent	3,040	1,304	357	2,170	1,957	8,820
#Interaction	7,613	3,383	1,039	5,723	5,276	23,004
Degree	35.39	2.48	1.51	2.22	2.75	3.43
Betweenness	1.54	0.49	0.01	0.22	0.52	0.53

Table 2: Statistics metrics on the inductive social relation graph and the subgraph of each subreddit. Degree and betweenness are the averaged metrics on each subgraph, which indicate the graph centrality.

Training. The training loss $\mathcal{L}_{train}$ consists of a classification term and a reconstruction term, denoted as:

$$\mathcal{L}_{train} = \mathcal{L}_{stance} + \mathcal{L}_{recon}. \tag{8}$$

Given the input and its golden label (x_i, y_i), the $\mathcal{L}_{stance}$ for classifying (dis)agreement is a cross-entropy loss:

$$\mathcal{L}_{stance} = -\frac{1}{|N|} \sum_{(x_i, y_i)} y_i \, log \, p(y_i), \tag{9}$$

where $|N|$ is the number of data samples. To further ensure stronger author invariance constraints of $\mathbf{h}_{RG}$, we add a shared decoder layer D_{recon} with a reconstruction loss:

$$\mathcal{L}_{recon} = -E_{\mathbf{h}^{RG}}(||D_{recon}(\mathbf{h}^R) - \mathbf{h}^{RG}||_2^2). \tag{10}$$

4 EXPERIMENTS

We verify the effectiveness of social relation information for the in-domain (train the model on all the subreddits and evaluate it on the corresponding test data) in Section 4.3 and cross-domain tasks (train the model on four subreddits and evaluate it on the one subreddit left) in Section 4.4. We also carry out further analysis of our model in Section 4.5.

4.1 Settings

Dataset: We adopt the dataset- DEBAGREEMENT [33] for (dis)agreement detection. The dataset consists of 42,804 comment-reply pairs from the popular discussion website reddit with authorship and temporal information. The data topics include Brexit, Climate, BlackLivesMatter, Republican, and Democrats. The statistics of the dataset are shown in Table 1. As shown in the dataset, the interactions of the dataset are sparse, especially in the subreddits BlackLivesMatter and Republican.

Training Details. We perform experiments using the official pre-trained BERT [21] and RoBERTa [25] models provided by Huggingface [2]. We train our model on 1 GPU (Nvidia GTX2080Ti) using the Adam optimizer [23]. To construct the relation graph, we use the probability $\rho = 0.3$ to select edges in the training set to be *interaction* edges. We show the statistics of the inductive social relations in Table 2. For training the graph autoencoder, the initial learning rate is 1e-2, the epoch is 2e3, the batch size is 1e5, and we take each edge as the temporal graph matrix $\mathcal{A}^i$ for the reason that the interactions of authors in the dataset are sparse (23,101 nodes and 42,804 edges). For the (dis)agreement detection training process, the initial learning rate is 1.5e-5, the max sequence length is 256, the batch size for training is 8 for BERT-based/RoBERTa-based models, and the models are trained for three epochs. We split the data into 80%/10%/10% train/val/test sets while maintaining the temporal order, where testing is done on the latest data. We adopt the macro-F1 score to find the best model configuration, and the main results reported are averaged on five different runs.

Baselines. The standard benchmark [33] does not contain platform-specific information such as hashtags or retweets, we provide several baselines for (dis)agreement detection, such as BiLSTM-based models − BiLSTM, BiLSTM-rel, BERT-based models − BERT-sep, BERT-joint, and RoBERTa-joint.

BiLSTM, we use the same bidirectional LSTM (BiLSTM) to encode both the comment and reply, and the average hidden states of each word are regarded as sentence representations of them. The sentence representations of the comment and reply are then concatenated. We use a linear layer to reduce the dimension, after which a softmax layer is applied to obtain the label's probability distribution. We use Glove-300 as the initial word embedding, a popular word embedding method capturing semantics [32].

BiLSTM-rel, we concatenate the textual information encoded by BiLSTM with the relation feature $\mathbf{h}^R$ and use a linear layer and a softmax layer to identify the (dis)agreement.

BERT-joint, we feed the input of $[CLS]$ *comment* $[SEP]$ *reply* $[SEP]$ into the BERT and apply a linear layer to reduce the dimension of $[CLS]$ hidden states, after which a softmax layer is used to obtain the distributions.

BERT-sep, the comment and reply are encoded by BERT in the format of $[CLS]$ *comment* $[SEP]$ and $[CLS]$ *reply* $[SEP]$ separately. The hidden states of $[CLS]$ tokens are concatenated as the representations of the comment-reply pair for classification.

RoBERTa-joint, we feed the input of $[CLS]$ *comment* $[SEP]$ *reply* $[SEP]$ into the RoBERTa and apply a linear layer to reduce the dimension of $[CLS]$ hidden states, after which a softmax layer is used to obtain the distributions.

4.2 In-domain Results

4.2.1 Overall results. We train our model with all the data from five subreddits, and the results are shown in Table 3. First, BiLSTM achieves 47.56%, 54.00%, and 32.70% macro-F1 scores for the categories, respectively, which are the lowest compared with BERT-based and RoBERTa-based models. It indicates that pre-trained language models such as BERT and RoBERTa can better learn textual representations for (dis)agreement detection. BiLSTM-rel

[2]https://huggingface.co/

Model	Agree			Disagree			Neutral			All	
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Acc	M-F1
BiLSTM	47.29	47.85	47.56	47.86	61.96	54.00	44.44	25.87	32.70	47.11	44.75
BERT-sep	68.92	68.26	68.44	68.79	73.29	70.58	53.29	48.55	50.80	64.68	63.27
BERT-joint	67.88	67.78	66.30	68.84	74.80	70.36	54.44	48.12	50.28	65.50	63.59
RoBERTa-joint	**72.28**	69.18	70.56	74.11	69.80	71.89	51.31	58.67	54.57	66.78	65.67
Ours											
BiLSTM-rel	50.35	57.65	53.75	51.87	55.71	53.77	42.23	28.79	34.17	49.62	47.23
BERT-rel	70.15	70.60	70.35	73.62	71.19	72.34	52.52	54.68	53.51	66.82	65.40
RoBERTa-rel	70.97	**72.01**	**71.44**	**75.62**	**73.01**	**74.27**	**54.16**	**55.95**	**55.02**	**68.38**	**66.91**

Table 3: In-domain testing results. The models are trained on the five subreddits and tested on the corresponding test data. (Prec , Rec, F1, Acc and M-F1 for the metrics of precision, recall, micro-F1 score, accuracy and macro-F1 score, henceforth).

	r/Br	r/Cl	r/BLM	r/Re	r/De
BiLSTM	44.82	43.08	51.81	46.59	52.86
BERT-joint	64.10	64.90	66.90	66.10	67.20
BERT-sep	63.68	65.05	64.24	65.11	66.73
RoBERTa-joint	65.83	66.92	71.23	69.38	67.55
BiLSTM-rel	46.15	44.46	53.89	50.05	53.27
BERT-rel	65.99	66.99	70.17	67.77	67.04
RoBERTa-rel	**66.81**	**68.77**	**71.37**	**70.25**	**68.24**

Table 4: Accuracies of RoBERTa-rel on each subreddit.

achieves 53.75%, 53.77%, and 34.17% macro-F1 scores for the classes, respectively. The macro-F1 score is 47.23% for BiLSTM-rel, which is 2.48% higher than that of BiLSTM. It demonstrates that the graph autoencoder and social relation information can help boost the performance of the randomly initialized model in the disagreement detection task.

In addition, the averaged macro-F1 model of BERT-joint is 0.32% higher than that of BERT-sep, indicating that the joint model can perform better than the pipeline model for the former captures the attention features between comments and replies. Our model BERT-rel achieves 70.35%, 72.34%, and 53.51% macro-F1 scores for the data, respectively. The averaged marco-F1 score of BERT-rel is 65.40%, which is 2.13% higher than that of BERT-joint and 1.89% higher than that of BERT-sep, indicating that social relation information has a significant effect on a pre-trained (dis)agreement detection model.

Moreover, RoBERTa-rel achieves the macro-F1 score of 66.91%, which is a state-of-the-art performance, which is 1.24% higher than that of RoBERTa-joint and 1.51% higher than that of BERT-rel. The results show that the model can achieve stronger performance with more accurate textual information features, and social relation information can still enhance the (dis)agreement detection performance. RoBERTa-rel achieves macro-F1 scores of 71.44%, 74.27%, and 55.95% on the labels, respectively. The improvement is the most significant on the data of disagreeing labels (2.38% higher than RoBERTa-joint), where the relations of *opponent* in the aggregate social graph are also more than those of *supporter*. And the most challenging part of the dataset is still the neutral data due to the small proportion of the neutral data (statistics are in Table 1).

4.2.2 Breakdown results. We test the models (trained on the five subreddits) on the data of each subreddit, and the results are shown in Table 4. The accuracies of BiLSTM-rel are 46.15%, 44.46%, 53.89%, 50.05%, and 53.27% for the subreddits of Brexit, Climate, BlackLivesMatter, Republican, and Democrats, respectively, which are all higher than those of BiLSTM, indicating the effectiveness of social relation information. The same phenomenon can also be observed in the BERT-rel and RoBERTa-rel, indicating our social relation is effective for a different model architecture of (dis)agreement detection. RoBERTa-rel achieves the accuracies of 66.81%, 68.77%, 71.37%, 70.25%, and 68.24% for the subreddits Brexit, Climate, BlackLivesMatter, Republican, and Democrats, respectively, which all achieve the state-of-the-art performance. The accuracy of RoBERTa-rel on BlackLivesMatter is improved the least (0.14%), which results from the sparsity of the edges in the data of the BlackLivesMatter (2,516 nodes, 19.29 edges, and 1.51 averaged degree).

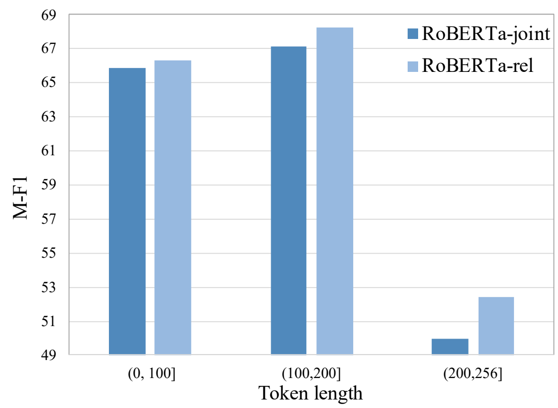

Figure 4: Results of RoBERTa-rel with respect to the different token lengths of the comment-reply pairs.

4.3 Cross-domain Results

We evaluate our model in the cross-domain settings, which aims to evaluate the model generalization ability, reducing the cost and requirement of human annotations for models [4, 8, 29, 46]. In particular, we train our model on the data of four sudreddits and test it on the left subreddit. The results are shown in Table 5. The macro-F1 scores of BiLSTM are 41.90%, 40.24, 39.73%, 41.32%, and 46.79% on each task, which is the worst compared with BERT-based and RoBERTa-based models, indicating that the randomly initialized model are less informative in the features for the task of (dis)agreement detection. The model BiLSTM-rel achieves the 43.19%, 43.14%, 41.05%, 44.13%, and 48.14% on each tasks, which are

Model	r/Br		r/Cl		r/BLM		r/Re		r/De		Average	
	Acc	M-F1	Acc	M-F1	Acc	M-F1	Acc	M-F1	Acc	M-F1	Acc	M-F1
BiLSTM	42.60	41.90	41.52	40.24	46.11	39.73	47.30	41.32	50.88	46.79	45.68	42.00
BERT-sep	61.84	61.73	63.82	63.11	65.80	62.86	64.23	61.51	65.91	63.52	64.32	62.71
BERT-joint	64.12	62.56	64.42	64.34	65.32	62.13	66.64	63.25	66.03	63.21	65.30	63.10
RoBERTa-joint	65.43	64.07	67.64	65.95	69.15	66.06	66.02	64.94	64.80	61.45	66.61	64.46
Ours												
BiLSTM-rel	44.32	43.19	42.33	43.14	46.33	41.05	49.32	44.13	50.78	48.14	46.62	43.93
BERT-rel	**66.49**	**65.13**	65.44	64.05	68.30	65.60	66.57	64.38	64.22	62.43	66.20	64.32
RoBERTa-rel	66.03	64.49	**68.29**	**66.83**	**69.17**	**66.49**	**70.23**	**67.88**	**67.96**	**66.88**	**68.34**	**66.51**

Table 5: Cross-domain testing results. The models are trained on the four subreddits and tested on the left subreddit.

1.29%, 2.90%, 0.32%, 2.81% and 1.45% higher than those of BiLSTM, respectively. The results show that by using social relations, the model can achieve stronger performances.

As in-domain testing results, BERT-joint can still perform better than BERT-sep, but both are less effective than BERT-rel in the cross-domain settings. The averaged precision and macro-F1 scores of BERT-rel on all the subreddit are 66.20% and 64.32%, which are 0.9%, 1.22% higher than BERT-joint, and 1.88%, 1.61% higher than BERT-sep, respectively. The results demonstrate the effectiveness of social relations in assisting (dis)agreement detection. Our model RoBERTa-rel achieves 68.34% accuracy and 66.51% macro F1 score on average, which is the best performance of our model on the (dis)agreement detection task. The performance is the lowest in the Brexit subreddit due to the large averaged degree and betweenness (35.39 and 1.54) in the subreddit while deleting the data from training hinders the model to learn complete social information (shown in Table 2). But in other subreddit, the macro F1 scores show a roughly positive correlation with an averaged degree and betweenness of the subgraphs in each subreddit (i.e., with the increase of averaged degree and betweenness, the improvement margin of macro F1 score increases). In particular, the averaged degrees are 0.01, 0.22, 0.49, and 0.52 for the subgraphs in the subreddits BlackLivesMatter, Republican, Climate, and Democrats, and the corresponding improvement margins are 0.43%, 3.05%, 0.49%, and 5.43%, respectively. The phenomenon demonstrates that with more abundant social relation information, it is simpler to identify the (dis)agreement. Note that the results of climate departure from the positive correlation, which may result from the reason that the authors of the Climate subreddit have less relation to those in other subreddits.

4.4 Further Analysis

4.4.1 Effect of token lengths. We test our model RoBERTa-rel with respect to different token lengths of comment-reply pairs (shown in Figure 4). It shows that RoBERTa-rel boosts the averaged macro-F1 scores of (dis)agreement detection with a large margin compared with RoBERTa-joint, 1.87% and 2.45 for the data with lengths $(100, 200]$ and > 200, respectively, but it outperforms RoBERTa-joint only 0.45% for data with lengths $(0, 100]$. The results show that it becomes challenging to identify the (dis)agreement labels with long sequence lengths merely using textual information. And it demonstrates that social relation information boosts the performance (dis)agreement detection, especially for the data with long lengths, which are difficult for models merely using textual information.

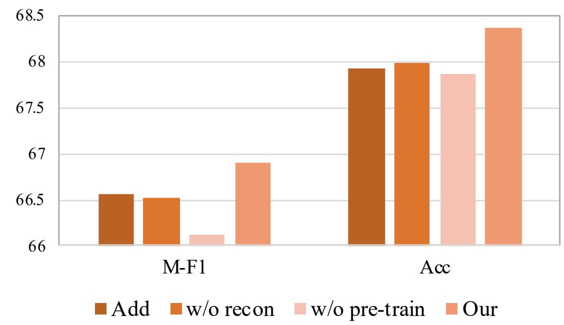

Figure 5: Ablation study on RoBERTa-rel, and different methods of information fusing in the in-domain testing.

4.4.2 Fusing Method. We also test our model with other methods for fusing textual and social relation information. We add the feature of social relation information and textual information, following $p = Softmax(W(\mathbf{h}^{CLS} + \mathbf{h}^R) + b)$. The averaged macro F1 score and accuracy are 66.54% and 67.86%, which are 0.37% and 0.52% lower than those of concatenating method. It demonstrates that although concatenation is intuitive, it is more effective than addition.

4.4.3 Ablation Study. Figure 5 shows the results of ablation studies. First, we show the results without using the reconstruction loss function, but only cross-entropy loss for (dis)agreement classification. The averaged macro F1 score and accuracy are 66.25% and 68.00%, which are 0.39% and 0.38% lower than those of RoBERTa-rel, respectively.

We also test our model without pre-training the RGCN module using KGE methods but solely train it on the (dis)agreement objectives $\mathcal{L}_{train}$. Without pre-training the RGCN module, the model performance decreases with a large margin of 0.78% and 0.80% in averaged macro F1 score and accuracy, respectively. It demonstrates the significance of the pre-training process in the embeddings of the social relation graph.

4.4.4 Scoring Function of Graph Autoencoder. To further analyze the influence of different knowledge graph embeddings (KGE), we compare RoBERTa-rel (using the DistMult method) with several models using other typical scoring functions in the decoder of the graph autoencoder (the encoding method of the textual information is the same), including the translated-based methods **TransE** [5], **TransF** [15], and semantic matching method **HolE** [31]. The results are shown in Table 7.

Comment	Reply	Soci Rel.	Label	Output
By that standard, every person on the internet is hundreds of times more guilty than rural villagers in Africa and India. Why don't you give up your technology?	That wasn't the point. I just read a news article telling people what they can do to stop climate change when he himself has multiple private jets. He can take first class on a normal plane but that would inconvenience him.	Supporter	Agree	Agree
Am I the only person who gets worried when they see a line of only other people! lol. jokes but... actually not joking. It scares me now.	I smile (awkwardly, I'm sure) at poc. I'll knock a person up if anyone were to harass someone who's just minding their own business.	Interaction	Disagree	Disagree

Table 6: Case Study. Soci Rel. is for social relations.

	Agree F1	Disagree F1	Neutral F1	M-F1
RoBERTa-joint	70.56	71.89	54.57	65.67
TransE	70.66	72.10	54.70	65.82
TransF	71.12	72.22	54.75	66.03
HolE	71.03	73.34	54.92	66.43
DistMult(Ours)	**71.44**	**74.27**	**55.02**	**66.91**

Table 7: Results with respect to different scoring functions of the graph autoencoder of the model ReBERTa-rel.

Observed in the results, all the models using the graph autoencoder outperform the model RoBERTa-joint, demonstrating the effectiveness of using social relation information. The translational distance methods TransE (a macro-F1 score of 65.67%) and TransF (a macro-F1 score of 65.82%) perform worse than the semantic matching methods HolE (a macro-F1 score of 66.43%) and DistMult (a macro-F1 score of 66.91%), for the reason TransE and TransF only extract the relation information of entities instead of semantic information. Since we consider the social relation graph in a directed graph, HolE should obtain more substantial expressive power than DistMult in encoding asymmetric relations. However, the model with the HolE method achieves lower performance (66.43%) than that with the DistMult method (66.91%), which may result from the sparsity of the social relation graph.

4.4.5 Effect of interaction selection ρ. We evaluate our model concerning different numbers of selected *interaction* edges in the training set. The results are illustrated in Figure 6. As is observed in the results, when ρ is 0.0, the macro-F1 scores of both the in-domain and cross-domain tests are the lowest, which are 66.22% and 65.77%, respectively, which indicates the importance of adding *interaction* edges with the increase of ρ. The macro-F1 scores increase and reach the peak (66.91% and 66.51%, respectively) when ρ is 0.3, which means the selection of edges to be *interaction* boost the performance of the (dis)agreement models. However, as ρ continues increasing to 0.4, the macro-F1 scores of both the tasks decrease (66.54% and 66.25%), which indicates that excessive *interaction* relations can also introduce noise and spurious features to social relation information.

4.5 Case Study

Some cases are shown in Table 6. The first case shows that the reply 'That was not the point.' implies a disagreeing stance towards the comment, which results in incorrect identification of RoBERTa-joint. Benefiting from the social relation *supporter* between them, RoBERTa-rel outputs a correct stance. For the second case, for the

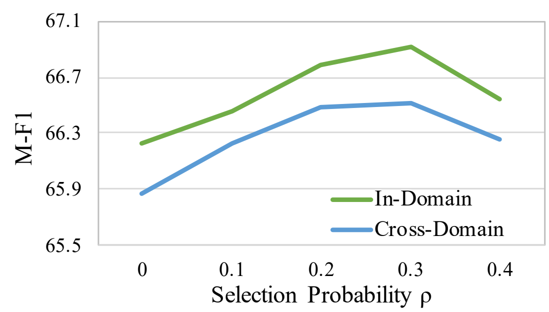

Figure 6: Results of RoBERTa-rel with respect to different rates of selected *interaction* edges in the training set. For the in-domain task, the model is trained in all subreddits and evaluated on the test data. For the cross-domain task, the metric is the averaged macro-F1 score of the five tasks in Section 4.4.

complex textual information limited without textual context, BERT-joint fails to give the correct output. Although there is only one *interaction* relation between the authors of the comment (n_c) and the reply (n_t), the authors have the same neighbor in the relation graph who n_c supporting but n_t opposing. It implies the authors of this comment-reply pair may be opposing, and this information assists BERT-rel in outputting the correct stance. The cases show the effectiveness and reasonableness of using social relations.

5 CONCLUSION

We proposed a method to construct an inductive social relation graph from the comment-reply data to assist (dis)agreement detection. The model used a graph autoencoder to extract relation information, consisting of an RGCN encoder and a DistMult decoder for pre-training. Our model achieves state-of-the-art performance in the standard dataset DEBAGREEMENT for in-domain and cross-domain settings, showing social relations' effectiveness. We found social relation boosts the performance, especially for the long-token comment-reply pairs. Ablation studies showed the significance of each module. The study shows that general external information can boost the (dis)agreement detection. It is promising to model the opinions of the authors on different topics and further analyze how social relations form and how opinions spread on social platforms. For future work, it is a promising direction to consider leveraging the effective temporal information in the sparse social graph network, and in this way, it becomes feasible to study how public opinions spread and evolve on the social platform in more realistic settings.

ACKNOWLEDGEMENT

We would like to thank the anonymous reviewers for the detailed and thoughtful reviews. The work is funded by the Pioneer and "Leading Goose" R&D Program of Zhejiang under Grant Number 2022SDXHDX0003 and State Key Laboratory of Media Convergence Production Technology and System 2020 Annual Research Project (SKLMCPTS2020006).

REFERENCES

[1] Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management* 58, 4 (2021), 102597.

[2] Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*. 1–9.

[3] Craig A Anderson and Kathryn L Kellam. 1992. Belief perseverance, biased assimilation, and covariation detection: The effects of hypothetical social theories and new data. *Personality and Social Psychology Bulletin* 18, 5 (1992), 555–565.

[4] Xuefeng Bai, Seng Yang, Leyang Cui, Linfeng Song, and Yue Zhang. 2022. Cross-domain Generalization for AMR Parsing. *arXiv preprint arXiv:2210.12445* (2022).

[5] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* 26 (2013).

[6] Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. Content and network dynamics behind Egyptian political polarization on Twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 700–711.

[7] Dorwin Cartwright and Frank Harary. 1956. Structural balance: a generalization of Heider's theory. *Psychological review* 63, 5 (1956), 277.

[8] Yulong Chen, Yang Liu, Ruochen Xu, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2022. UniSumm: Unified Few-shot Summarization with Multi-Task Pre-Training and Prefix-Tuning. *arXiv preprint arXiv:2211.09783* (2022).

[9] Chun Cheng, Yun Luo, and Changbin Yu. 2020. Dynamic mechanism of social bots interfering with public opinion in network. *Physica A: Statistical Mechanics and its Applications* 551 (2020), 124163.

[10] Yanwen Chong, Yun Ding, Qing Yan, and Shaoming Pan. 2020. Graph-based semi-supervised learning: A review. *Neurocomputing* 408 (2020), 216–230.

[11] Pedro Colon-Hernandez, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. 2021. Combining pre-trained language models and structured knowledge. *arXiv preprint arXiv:2101.12294* (2021).

[12] Kareem Darwish, Walid Magdy, Afshin Rahimi, Timothy Baldwin, and Norah Abokhodair. 2018. Predicting online islamophopic behavior after# parisattacks. *The Journal of Web Science* 4 (2018).

[13] Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 141–152.

[14] Kuntal Dey, Ritvik Shrivastava, Saroj Kaushik, and Vaibhav Mathur. 2017. Assessing the effects of social familiarity and stance similarity in interaction dynamics. In *International Conference on Complex Networks and their Applications*. Springer, 843–855.

[15] Jun Feng, Minlie Huang, Mingdong Wang, Mantong Zhou, Yu Hao, and Xiaoyan Zhu. 2016. Knowledge graph embedding by flexible translation. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

[16] Corey L Guenther and Mark D Alicke. 2008. Self-enhancement and belief perseverance. *Journal of Experimental Social Psychology* 44, 3 (2008), 706–712.

[17] Kazi Saidul Hasan and Vincent Ng. 2013. Stance Classification of Ideological Debates: Data, Models, Features, and Constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Nagoya, Japan, 1348–1356. https://aclanthology.org/I13-1191

[18] Yiru Jiao and Yongli Li. 2021. An active opinion dynamics model: The gap between the voting result and group opinion. *Information Fusion* 65 (2021), 128–146.

[19] Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. 2020. Graph structure learning for robust graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 66–74.

[20] Márton Karsai, Mikko Kivelä, Raj Kumar Pan, Kimmo Kaski, János Kertész, A-L Barabási, and Jari Saramäki. 2011. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E* 83, 2 (2011), 025102.

[21] Jacob Devlin Kenton, Chang Ming-Wei, and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[22] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* (2019).

[23] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

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

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[26] Zihan Liu, Yun Luo, Lirong Wu, Zicheng Liu, and Stan Z Li. 2022. Towards Reasonable Budget Allocation in Untargeted Graph Structure Attacks via Gradient Debias. In *Advances in Neural Information Processing Systems*.

[27] Philipp Lorenz-Spreen, Stephan Lewandowsky, Cass R Sunstein, and Ralph Hertwig. 2020. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature human behaviour* 4, 11 (2020), 1102–1109.

[28] Yun Luo, Chun Cheng, Yuke Li, and Changbin Yu. 2021. Opinion formation with zealots on temporal network. *Communications in Nonlinear Science and Numerical Simulation* 98 (2021), 105772.

[29] Yun Luo, Fang Guo, Zihan Liu, and Yue Zhang. 2022. Mere Contrastive Learning for Cross-Domain Sentiment Analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*. 7099–7111.

[30] Yun Luo, Zihan Liu, Stan Z Li, Yuefeng Shi, and Yue Zhang. 2022. Exploiting Sentiment and Common Sense for Zero-shot Stance Detection. *arXiv preprint arXiv:2208.08797* (2022).

[31] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2016. Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[32] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[33] John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and Doyne Farmer. 2021. DEBAGREEMENT: A comment-reply dataset for (dis) agreement detection in online debates. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[34] Minghui Qiu, Yanchuan Sim, Noah A Smith, and Jing Jiang. 2015. Modeling user arguments, interactions, and attributes for stance prediction in online debate forums. In *Proceedings of the 2015 SIAM international conference on data mining*. SIAM, 855–863.

[35] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[36] Manoel Horta Ribeiro, Pedro H Calais, Virgílio AF Almeida, and Wagner Meira Jr. 2017. " Everything I disagree with is# FakeNews": Correlating political polarization and spread of misinformation. *arXiv preprint arXiv:1706.05924* (2017).

[37] Sara Rosenthal and Kathleen McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 168–177.

[38] Younes Samih and Kareem Darwish. 2021. A few topical tweets are enough for effective user stance detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2637–2646.

[39] M. Schlichtkrull, Thomas Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. *ArXiv* abs/1703.06103 (2018).

[40] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*. 613–624.

[41] Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine Learning* 109, 2 (2020), 373–440.

[42] Marilyn A. Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance Classification Using Dialogic Properties of Persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Montreal, Canada) *(NAACL HLT '12)*. Association for Computational Linguistics, USA, 592–596.

[43] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering* 29, 12 (2017), 2724–2743. https://doi.org/10.1109/TKDE.2017.2754499

[44] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* 32, 1 (2020), 4–24.

Comment	Reply	Soci Rel.	Label	Output
Gates is promoting Exxon fantasy air carbon capture and "new" nuclear that is not in anyway close to being useful and would take way too long to build when we need cheap, fast and safe renewable energy to replace fossil fuels right now. He is promoting his own book and wants a position on Biden's climate team.	I read his book and in it he actually says that the air carbon capture in no way is scaleable enough, but whatever you say man.	Opponent	Disagree	Disagree
What other countries are experiencing this? Needing to give up towns/big areas to water due to rising sea levels?	Greenland is ground zero for climate change, and everybody who lives in Greenland lives right on or very near the coast.	Interaction	Neutral	Neutral
Think they'll do it? sounds like a cry for attention, surely they know this will render them politically incompetent.	This is the GOP splitting in 2 before our eyes. A lot of conservatives were horrified by the events on Jan. 6, and never bought the big lie.	Interaction	Agree	Agree
A president isn't supposed to be impeached for failing to respond to the most pressing issues in your opinion. He should've been impeached very early on for breaking a handful of other guidelines of the presidency, abusing the power of the office, and violating the Constitution. His climate policy is not something impeachable.	That's absolutely ridiculous. His climate policy should be impeachable. Stupid rules and precedent aren't as important as preventing extinction.	Opponent	Disagree	Disagree

Table 8: Case Study. Soci Rel. is for social relations.

[45] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575* (2014).

[46] Jiali Zeng, Yang Liu, Jinsong Su, Yubin Ge, Yaojie Lu, Yongjing Yin, and Jiebo Luo. 2019. Iterative Dual Domain Adaptation for Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019.* Association for Computational Linguistics, 845–855. https://doi.org/10.18653/v1/D19-1078

[47] Shaodian Zhang, Lin Qiu, Frank Chen, Weinan Zhang, Yong Yu, and Noémie Elhadad. 2017. We make choices we think are going to save us: Debate and stance identification for online breast cancer CAM discussions. In *Proceedings of the 26th International Conference on World Wide Web Companion.* 1073–1081.

A TIME PERIOD OF THE CONSTRUCTION OF THE SOCIAL RELATIONS

We also design the test with the time period τ in the construction of the social relation graph. The results are shown in Figure 7. With the increase of the time period, the number of edges in different types varies in a small range (due to the sparsity of comment-reply pairs). We can still observe that with the increase of the time interval, the change of relations is more drastic, with the decrease of the model performance (from 66.91%, 66.83%, to 66.68%). A suitable time interval is a significant part of the model due to the change in the effectiveness of inductive social relations. The results go against our common sense that the social relation has temporal effectiveness, while it may result from the sparsity of the comment-reply pairs. More deep analysis requires comprehensive datasets with multiple interactions between authors and dense graphs, which can be future work in such an area.

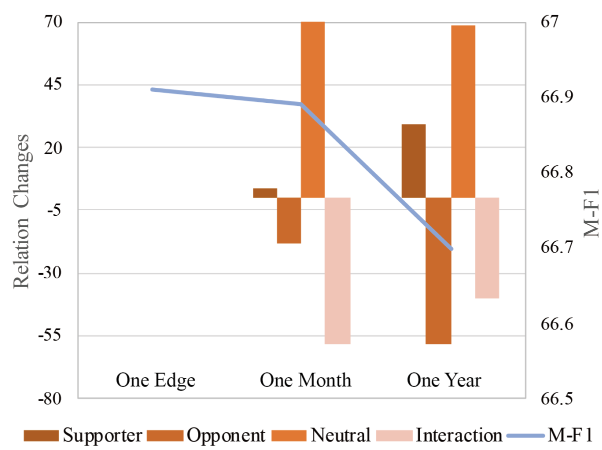

Figure 7: Results of RoBERTa-rel and the changes of relation numbers with respect to different time intervals.

B CASE STUDY

We show more case studies in Table 8.