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 fusion 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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WWW '23, May 1–5, 2023, Austin, TX, USA
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9416-1/23/04...
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
We don't listen to scientists, period. We need people to listen to them instead of politicians, but that's hardly likely. So, the politicians need to listen to our scientists.

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, 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 how it enhances (dis)agreement detection 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, 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 how it enhances (dis)agreement detection 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, 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 how it enhances (dis)agreement detection 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.

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
We formulate the (dis)agreement detection task as a classification task. The architecture of our model is illustrated in Figure 3, which conveys 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 trans-lational 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 $D = \{e_i, t_i, y_i, n_i^r, n_i^n\}_{i=1}^N$ be a dataset with $N$ examples, each consisting of a comment $e_i$ from author $n_i^r$, a reply $t_i$ from author $n_i^n$, and a stance label $y_i$ from $r_i$ to $n_i^n$ through the comment-reply pair. The task is to predict a stance label $y_i \in \{agree, disagree, neutral\}$ 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 $G = (N, E, R)$, with nodes (authors) $n_i \in N$ and labeled edges $(n_i, r, n_j) \in E$, where $r \in R$ 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 $N$. 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 graph of the weighted adjacent matrix (snapshots) $S(w, r) = \{A^0, A^1, ..., A^{(w-1)}\}$ $A^k$ is the graph weighted adjacent matrix during time period $[kr, (k + 1)r]$, the inductive graph is drawn from the interactions that appear during the timescale wr (Figure 2). For each graph weighted adjacent matrix $A^k$, if the author $n_i$ expresses an agree/disagree-neutral stance towards $n_j$ in a comment-reply pair, the value $a^k_{ij} = +1/-1/0$, and if there are multiple interactions between them, the most frequent opinion (agree/disagree-neutral) are considered to determine $a^k_{ij}$. Then the weighted adjacent matrix of the inductive graph is $A^r = A^0 + A^1 + ... + A^{(w-1)}$. The triplet $(n_i, r, n_j)$ is in relation graph $G_R$ as follows:

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

and $(n_i, r, n_j)$ is not in $G_R$ in other situations.

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 $G_R$ with probability $p$ 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 $\tilde{E}$ from $E$ in $G_R$ is fed as the graph autoencoder input. The incomplete set $\tilde{E}$ is negatively sampled to a complete set of samples denoted $U$ (details in training). Then we assign the possible edges $(n_i, n_j) \in U$ with scores, which are used to determine the probability of whether the edges are true in $E$. 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 2: The illustration of the construction of the social relation graph using the temporal order 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 $u_i$. Then the vectors are transformed into relation-aggregated feature vectors $h_i \in \mathbb{R}^d$ using the RGCN encoders:

$$f(x_i, l) = \sigma(W_0^{(l)} x_i + \sum_{r \in R} \sum_{j \in N_i^r} \frac{1}{n_{i,r}} W_r^{(l)} x_j),$$

$$h_i = h_i^{(2)} = f(h_i^{(1)}, 2) ; h_i^{(1)} = f(u_i, 1),$$

(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 neighboring authors $i$ with the relation $r \in R$; $n_{i,r}$ is a normalization constant, set in advance $n_{i,r} = |N_i^r|$ or learned by network learning; $\sigma$ is the activation function such as ReLU, and $W_r^{(1)}, W_r^{(2)}$ 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:

$$s(n_i, r, n_j) = \sigma(h_n^T R_h n_j),$$

(2)

where $\sigma$ is a logistic function; $h_n, h_j \in \mathbb{R}^n$ are the encoding feature vectors through the graph encoder for author $n_i$ and $n_j$; every type of relation $r \in R$ 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 $\mathcal{E}$, 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 $g$) triplets with a cross-entropy loss function:

$$L_G^+ = -\frac{1}{2|\mathcal{E}|} \sum_{(n_i, r, n_j) \in \mathcal{U}} (y \log s(n_i, r, n_j) + (1 - y) \log (1 - s(n_i, r, n_j))).$$

(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 $n_i^C$ and $n_i^R$ as the authors for $c_i$ and $t_i$, respectively. Then we extract a sub-graph $G_A$ from $G_R$, which contains all the authors on the graph within the vicinity of radius 1 from $n_i^C$ and $n_i^R$. Next, we make a forward pass of $G_A$ through the encoder of graph autoencoder to obtain the feature vectors $h_j$ for all unique authors $j$ in $G_A$. The average of feature vectors $h_j$ for all unique authors in $G_A$ is regarded as the relation graph feature vector $h^{RG}$:

$$h^{RG} = \text{RGCN}(G_A).$$

(4)

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

$$h^R = W_R h^{RG} + b_R$$

(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 GPTs [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:

$$h_{PLM} = \text{PLM}(x_i).$$

(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 $h^R$ and $h^{CLS}$ are concatenated for classification:

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

(7)

where $W$ and $b$ are the parameters and $p$ is the probability distribution on the three (dis)agreement labels.
We verify the effectiveness of social relation information for the

Table 2: Statistics metrics on the inductive social relation
detection. The dataset consists of 42,804 comment-reply pairs
on the corresponding test data) in Section 4.3 and cross-domain

<table>
<thead>
<tr>
<th>r/Br</th>
<th>r/Cl</th>
<th>r/BLM</th>
<th>r/Re</th>
<th>r/De</th>
</tr>
</thead>
<tbody>
<tr>
<td>#nodes</td>
<td>722</td>
<td>4,580</td>
<td>2,516</td>
<td>8,832</td>
</tr>
<tr>
<td>#edges</td>
<td>15,745</td>
<td>5,773</td>
<td>1,929</td>
<td>9,823</td>
</tr>
<tr>
<td>Agree</td>
<td>29%</td>
<td>32%</td>
<td>45%</td>
<td>34%</td>
</tr>
<tr>
<td>Neutral</td>
<td>29%</td>
<td>28%</td>
<td>22%</td>
<td>25%</td>
</tr>
<tr>
<td>Disagree</td>
<td>42%</td>
<td>40%</td>
<td>33%</td>
<td>41%</td>
</tr>
</tbody>
</table>

Table 1: Statistics on DEBAGREEMENT. Br for the subreddit
BlackLivesMatter and Republican.

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

<table>
<thead>
<tr>
<th>r/Br</th>
<th>r/Cl</th>
<th>r/BLM</th>
<th>r/Re</th>
<th>r/De</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Supporter</td>
<td>2,159</td>
<td>989</td>
<td>511</td>
<td>1,882</td>
</tr>
<tr>
<td>#Opponent</td>
<td>3,040</td>
<td>1,304</td>
<td>357</td>
<td>2,170</td>
</tr>
<tr>
<td>#Interaction</td>
<td>7,613</td>
<td>3,383</td>
<td>1,039</td>
<td>5,723</td>
</tr>
<tr>
<td>Degree</td>
<td>35.39</td>
<td>2.48</td>
<td>1.51</td>
<td>2.22</td>
</tr>
<tr>
<td>Betweenness</td>
<td>1.54</td>
<td>0.49</td>
<td>0.01</td>
<td>0.22</td>
</tr>
</tbody>
</table>

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

$$L_{\text{train}} = L_{\text{stance}} + L_{\text{recon}}.$$  (8)

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

$$L_{\text{stance}} = -\frac{1}{|N|} \sum_{(x_i, y_i)} y_i \log p(y_i),$$  (9)

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

$$L_{\text{recon}} = -E_{h_{RG}} (\| D_{\text{recon}}(h^R) - h_{RG}^2 \|).$$  (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.

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 $A^t$ 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 sev-
eral 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 en-
code 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 $h^R$ 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
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, BiL-
STM 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

https://huggingface.co/
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).

<table>
<thead>
<tr>
<th>Model</th>
<th>Agree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
<td>F1</td>
<td>Prec</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>47.29</td>
<td>47.85</td>
<td>47.56</td>
<td>47.86</td>
</tr>
<tr>
<td>BERT-sep</td>
<td>68.92</td>
<td>68.26</td>
<td>68.44</td>
<td>68.79</td>
</tr>
<tr>
<td>BERT-joint</td>
<td>67.88</td>
<td>67.78</td>
<td>66.30</td>
<td>68.84</td>
</tr>
<tr>
<td>RoBERTa-joint</td>
<td>72.28</td>
<td>69.18</td>
<td>70.56</td>
<td>74.11</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM-rel</td>
<td>50.35</td>
<td>57.65</td>
<td>53.75</td>
<td>51.87</td>
</tr>
<tr>
<td>BERT-rel</td>
<td>70.15</td>
<td>70.60</td>
<td>70.35</td>
<td>73.62</td>
</tr>
<tr>
<td>RoBERTa-rel</td>
<td>70.97</td>
<td>72.01</td>
<td>71.44</td>
<td>75.62</td>
</tr>
</tbody>
</table>

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

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

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

<table>
<thead>
<tr>
<th>Token length</th>
<th>M-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 100]</td>
<td>69.18</td>
</tr>
<tr>
<td>(100, 200]</td>
<td>67.23</td>
</tr>
<tr>
<td>(200, 256]</td>
<td>65.67</td>
</tr>
</tbody>
</table>

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 of 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).

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

With long sequence lengths merely using textual information. And 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 betweeness of the subgraphs in each subreddit (i.e., with the increase of averaged degree and betweeness, 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.](image)

#### 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 \( p = \text{Softmax}(W(h_{CLS} + h_R) + b) \). 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 \( \mathcal{L}_{\text{train}} \). 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.
which means the selection of edges to be interaction with the HolE method achieves lower performance (66.43%) than interaction (66.54% and 66.25%), which indicates that excessive RoBERTa-rel outputs a correct stance. For the second case, for the joint. Benefiting from the social relation supporter the comment, which results in incorrect identification of RoBERTa-reply ‘That was not 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.

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

<table>
<thead>
<tr>
<th>Comment</th>
<th>Reply</th>
<th>Soci Rel.</th>
<th>Label</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>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?</td>
<td>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.</td>
<td>Supporter</td>
<td>Agree</td>
<td>Agree</td>
</tr>
<tr>
<td>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.</td>
<td>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.</td>
<td>Interaction</td>
<td>Disagree</td>
<td>Disagree</td>
</tr>
</tbody>
</table>

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

![Figure 6: Results of RoBERTa-rel with respect to different rates of selected interaction edges in the training set.](image)

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 2022SDXHD0003 and State Key Laboratory of Media Convergence Production Technology and System 2020 Annual Research Project (SKLMCPTS2020006).

REFERENCES


<table>
<thead>
<tr>
<th>Comment</th>
<th>Reply</th>
<th>Soci Rel.</th>
<th>Label</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gates is promoting Exxon fantasy air carbon capture and &quot;new&quot; 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.</td>
<td>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.</td>
<td>Opponent</td>
<td>Disagree</td>
<td>Disagree</td>
</tr>
<tr>
<td>What other countries are experiencing this? Needing to give up towns/big areas to water due to rising sea levels?</td>
<td>Greenland is ground zero for climate change, and everybody who lives in Greenland lives right on or very near the coast.</td>
<td>Interaction</td>
<td>Neutral</td>
<td>Neutral</td>
</tr>
<tr>
<td>Think they’ll do it? sounds like a cry for attention, surely they know this will render them politically incompetent.</td>
<td>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.</td>
<td>Interaction</td>
<td>Agree</td>
<td>Agree</td>
</tr>
<tr>
<td>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.</td>
<td>That's absolutely ridiculous. His climate policy should be impeachable. Stupid rules and precedent aren’t as important as preventing extinction.</td>
<td>Opponent</td>
<td>Disagree</td>
<td>Disagree</td>
</tr>
</tbody>
</table>

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

---

**A TIME PERIOD OF THE CONSTRUCTION OF THE SOCIAL RELATIONS**

We also design the test with the time period $\tau$ 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.

**B CASE STUDY**

We show more case studies in Table 8.

---

