Information Research - Vol. 30 No. iConf (2025)

A more advanced group polarization measurement approach based on LLM-based agents and graphs

Zixin Liu, Ji Zhang, and Yiran Ding

DOI: https://doi.org/10.47989/ir30iConf47578

Abstract

Introduction. Group polarization is an important research direction in social media content analysis, attracting many researchers to explore this field. Therefore, how to effectively measure group polarization has become a critical topic. Measuring group polarization on social media presents several challenges that have not yet been addressed by existing solutions. First, social media group polarization measurement involves processing vast amounts of text, which poses a significant challenge for information extraction. Second, social media texts often contain hard-to-understand content, including sarcasm, emojis, and internet slang. Additionally, group polarization research focuses on holistic analysis, while texts is typically fragmented. These challenges indicate that a new solution needs to be proposed.

Method. To address these challenges, we designed a solution based on a multi-agent system and used a graph-structured community sentiment network (CSN) to represent polarization states. Furthermore, we developed a metric called community opposition index (COI) based on the CSN to quantify polarization.

Conclusion. We tested our multi-agent system through a zero-shot stance detection task and achieved outstanding results, which proved its significant value in terms of usability and accuracy.

Introduction

With the development of internet technology, social media has gained widespread popularity. *GLOBAL DIGITAL REPORT* (https://datareportal.com/global-digital-overview) indicates that platforms such as Facebook, YouTube, and TikTok boast billions of users worldwide. Social media has become a key avenue for the public to express opinions and engage in discussions. Its anonymity and convenience enable users to freely express their true views, thereby shaping social media public opinion. As a result, research in this field has also flourished.

In these studies, research from the perspective of group polarization holds a significant position. The concept of group polarization was first introduced by Stoner, who observed that group decisions tend to be more extreme compared to individual decisions (Stoner, 1961; Isenberg, 1986). In the internet era, group polarization is broadly defined as the divergence of public opinions or stances. Building on this definition, researchers have conducted extensive and comprehensive studies on various issues related to group polarization. One of the fundamental research problems in the field of social media group polarization is its measurement. Early measurement methods based on statistical approaches suffered from issues such as overly simplistic for the complexity of social media dynamics (Bilal et al., 2019; Gaurav et al., 2013; Hart et al., 2020; Jaidka et al., 2018; Tumasjan et al., 2010). Current mainstream methods, such as text clustering or sentiment classification (Belcastro et al., 2020; Jiang et al., 2018; Ribeiro et al., 2017; Tyagi et al., 2020), struggle to balance efficiency and interpretability. While some researchers have made significant progress by focusing on the relationships between different viewpoints, these studies still fall short in understanding the deeper nuances of opinion stances and their evolution (Boxell et al., 2020; Iyengar et al., 2019; Jiang et al., 2023; Lelkes et al., 2024; Maia et al., 2023).

To address the existing issues in measuring group polarization and improve efficiency, accuracy, and interpretability, we propose a new group polarization measurement approach based on LLM-based agents and graphs. This approach draws inspiration from the stance detection task in natural language processing (NLP) and the earlier *sentiment thermometer* method. We use a community sentiment network (CSN) represented by a graph structure to model the polarization state, where LLM-based agents are employed to construct the network. Additionally, we design polarization measurement metrics based on CSN. To validate the effectiveness of our approach, we tested the module responsible for constructing CSN on zero-shot stance detection tasks, and the results demonstrated its superiority in capturing the nuances of group polarization.

In summary, our contributions are as follows: (1) we propose a temporal community sentiment network (CSN) to represent the polarization state over time. (2) we introduce LLM-based agents for stance detection into group polarization measurement, significantly enhancing both efficiency and accuracy. (3) we propose a more robust metric based on the CSN, community opposition index (COI).

Related work

Opinion polarization measurement

As research on group polarization deepens, extensive exploration has also been conducted on the measurement of group polarization, leading to the development of a relatively comprehensive system of measurement approaches. Existing research suggests that the current mainstream group polarization measurement schemes can be primarily divided into three categories: volume-based, sentiment-based, and network-based (Bilal et al., 2019; Jaidka et al., 2018). We will discuss the characteristics of these three measurement schemes and their shortcomings when applied to the task of measuring group polarization on social media.

Volume-based approaches primarily rely on statistical methods and were widely applied in the early research of group polarization. Early researchers collected data through surveys and

experiments and used statistical analysis to obtain relevant polarization results. In current trend of exploring group polarization via social media, volume-based schemes focus more on various data metrics and employ statistical methods in research. Notable examples include Gaurav et al.'s political polarization study based on the moving average aggregate probability method, Tumasjan et al.'s analysis of political polarization using the LIWC tool, and Hart et al.'s use of multidimensional statistical analysis to analyse polarization during COVID-19 (Gaurav et al., 2013; Tumasjan et al., 2010; Hart et al., 2020).

However, existing studies suggest that volume-based schemes have limitations in terms of capturing information and analysing large datasets. They fail to accurately understand the opinions and sentiments conveyed in the text and generally rely on broad statistical metrics (e.g., word frequency, likes, bookmarks, etc) to gather limited information. The lack of rapid information capture and in-depth understanding makes these techniques less effective for tracking and real-time analysis of group polarization, and they also fall short in terms of precision in measuring polarization.

Compared to volume-based approaches, sentiment-based approaches place greater emphasis on the meaning and emotions conveyed in the text. Typically, sentiment-based approaches are grounded in natural language processing (NLP) and analyse social media text from the perspectives of opinions and emotions. These methods generally follow two main strategies. The first involves clustering texts based on the similarity of sentiments, such as the IOM-NN method proposed by Belcastro for accurately detecting emotional information in political polarization (Belcastro et al., 2020). The second strategy leverages deep learning for direct sentiment classification, exemplified by Tyagi et al.'s research on polarization driven by climate change, and the explorations by Ribeiro et al. and Jiang et al. on the relationship between misinformation and polarization (Tyagi et al., 2020; Ribeiro et al., 2017; Jiang et al., 2018).

Unfortunately, both strategies face notable challenges in practice. For text clustering, current clustering algorithms are relatively coarse and simplistic, making it difficult to distinguish between disruptive information (such as advertisements or neutral statements) and significant content. Moreover, they do not account for the specific relationships between subgroups or their contribution to polarization, resulting in outcomes that lack precision and interpretability. In sentiment classification, current methods often rely on binary classifications, failing to capture the intensity of emotions. This oversimplified method negatively impacts both the interpretability and accuracy of polarization measurement.

It is also worth noting that in political polarization research, a method called the *sentiment thermometer* has been widely adopted (Boxell et al., 2020; Iyengar et al., 2012; Iyengar et al., 2019; Lelkes et al., 2017; Wakefield et al., 2023). This approach uses surveys to gather voters' emotional scores toward various political parties, thus enhancing the precision and interpretability of the analysis. However, this method is costly, limited by small sample sizes, and not well-suited for measuring polarization in the context of social media, as it generally considers only two opposing groups, whereas social media often involves more complex, multi-group dynamics (Iyengar et al., 2012).

Network-based approaches represent an approach that evaluates group polarization by considering social positions and relationships among groups, focusing on emotional direction and the stance between subgroups to better measure polarization levels. Traditional social network analysis in group polarization studies explores peripheral connections around core opinions to assess subgroup positions and emotions (Bravo et al., 2015; Conover et al., 2011; Garcia et al., 2015; Guerra et al., 2013; Medaglia et al., 2017; Vicario et al., 2017). Some researchers have further developed this by dynamically simulating the process of group polarization to explore its evolutionary pathways (Maia et al., 2023; Santos et al., 2021). With the advancement of graph neural

networks (GNNs), network-based approaches have been enhanced, as demonstrated by valuable explorations from researchers such as Xiao et al., Zhang et al., and Jiang et al., who utilized sentiment networks and GNNs in their studies (Xiao et al., 2020; Zhang et al., 2019; Jiang et al., 2023). While these studies have achieved promising results in improving the scientific rigor and interpretability of group polarization research, they generally lack detailed analysis of the textual content. Moreover, social division within purely social networks does not necessarily result from group polarization, raising questions about the accuracy of some conclusions drawn from these methods.

LLM-based agents

With the introduction of OpenAI's GPT series of large language models (LLMs), numerous research fields have incorporated or examined GPT's capabilities (Brown et al., 2020). In the realm of group polarization research, some researchers have also explored its applications. For instance, Lu et al. used agent-based simulations to model group polarization dynamics, while Zhang et al. employed LLM-Based Agents to detect stances within polarized groups (Lu, 2024; Zhang, 2022). These studies have yielded promising results, demonstrating the feasibility and value of applying large language models in this field.

Method

As we mentioned in the first section, to achieve an accurate measurement of group polarization, our proposed method consists of three parts, specifically: (1) a community sentiment network (CSN) used to represent subgroups, the emotions between and within subgroups. (2) an efficient multi-agent system for CSN construction. (3) a group polarization metric based on the CSN, community opposition index (COI).

The set of opinion texts will be used to identify subgroups and conduct sentiment analysis through the multi-agent system, forming a CSN. The current polarization measurement result for the time slice can then be calculated using the COI.

Community sentiment network

The community sentiment network (CSN) is an extension of the *sentiment thermometer* method. The traditional *sentiment thermometer* could only be applied to two subgroups (Iyengar, 2012). CSN extends the *sentiment thermometer* to a directed cyclic graph that involves emotions between multiple subgroups (see Figure 1). Let $G = (V, E)$ be a graph, where V is the vertex set containing subgroups and E is the set of edges representing the sentimental relationships between subgroups. Each edge $e \in E$ is defined as (u, v, s) where $u, v \in V$ and s is the sentiment score. It should be noted that nodes v and u are allowed to be the same node, meaning self-loops are permitted. The score s can be either positive or negative, reflecting the positive or negative nature of the sentiment.

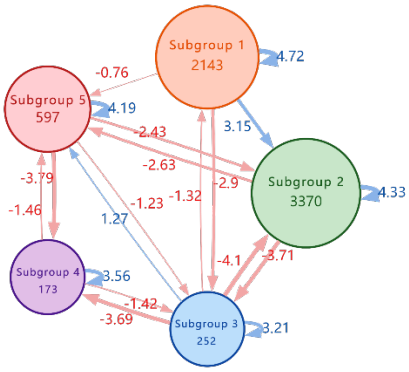

Figure 1. An example of a CSN generated by Graphviz based on comments scraped from Weibo under topics related to the Russia-Ukraine conflict, within a random time window

Unlike traditional social networks, CSN uses sentiment rather than interactions as the basis for constructing connections. Also, CSN clearly illustrates the various subgroups with different stances within the target time period and reveals the emotional states between subgroups as well as the internal cohesion of each subgroup. Therefore, compared to clustering results or social networks, CSN significantly highlights the contributions of different subgroups to polarization, providing greater interpretability of the polarization state.

The construction of the CSN involves multiple issues, such as subgroup identification, stance detection of opinion information, and sentiment recognition. However, existing research in the field of group polarization is insufficient to provide effective solutions to these issues. Therefore, inspired by the advancements in stance detection tasks (Xiao et al., 2020), we designed a multi-agent system based on LLM-based agents (see Figure 2).

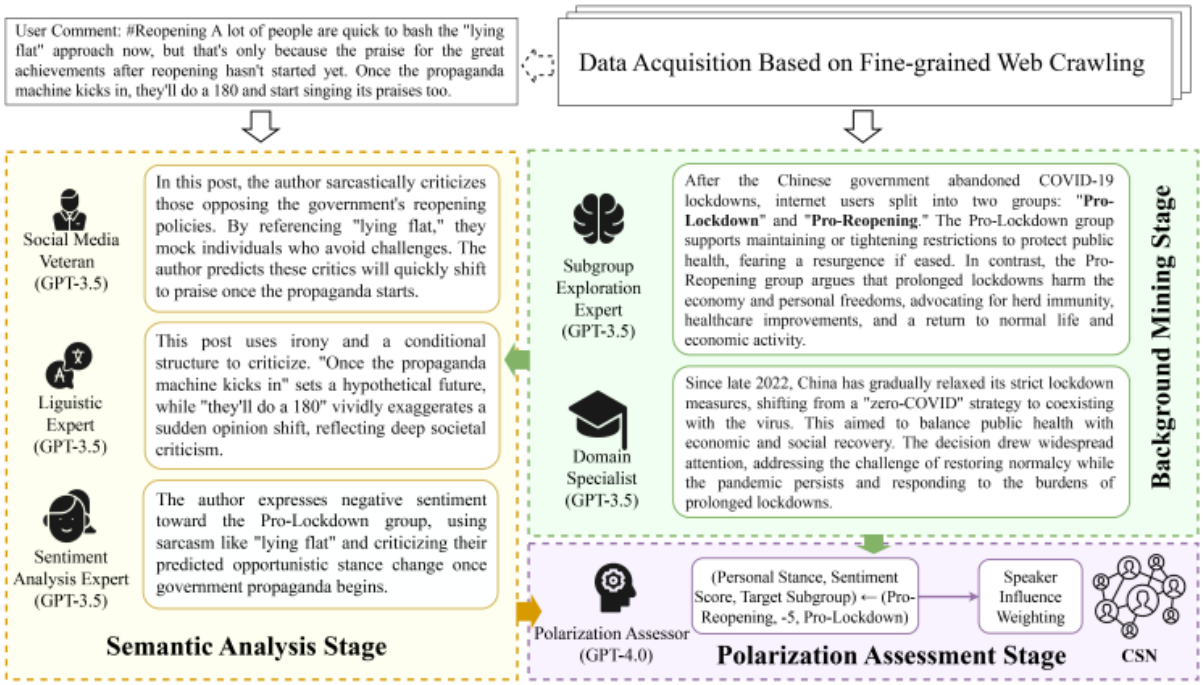

Figure 2. The structure of multi-agent system for CSN construction, containing background mining stage, semantic analysis stage and polarization assessment stage

In summary, our multi-agent system is composed of three stages: the background mining stage, the semantic analysis stage, and the polarization assessment stage. The background mining stage

consists of subgroup exploration expert and domain specialist, the semantic analysis stage includes social media veteran, linguistic expert, and sentiment analysis expert. The polarization assessment stage comprises polarization assessor. The agents collaborate through a series of interactions to provide accurate subgroup division results and reliable sentiment scores (see Figure 3). We will provide a detailed explanation of each stage and the functions of the respective agents in the following paragraphs.

```
Input: Weibo comments data within a specified time range
Output: Set of triplets as analysis results
procedure AnalyzeComments(comments)
    bg ← DomainExpert(comments)                    ▷ bg for background
    sg ← ∅                                         ▷ sg for subgroups
    uncertainComments ← []
    SubgroupMiningExpert(bg)
    for each comment in comments do
        grp ← SubgroupMiningExpert(comment)
        if not grp then
            uncertainComments.add(comment)
            if length(uncertainComments) reaches threshold then
                grp ← HumanExpertHelp(uncertainComments)
                sg.add(grp)
                uncertainComments ← []
            end if
        else
            sg.add(grp)
        end if
    end for
    Initialize Experts:
    SocialMediaVeteran(sg, bg)
    LinguisticExpert(bg)
    AffectAnalysisExpert(sg)
    PolarizationJudge(sg, bg)
    triplets ← ∅
    for each comment in comments do
        out1 ← SocialMediaVeteran(comment)
        out2 ← LinguisticExpert(comment)
        out3 ← AffectAnalysisExpert(comment, out1, out2)
        out4 ← PolarizationJudge(comment, out3)
        if out4.group then
            triplet ← (out4.group, out4.score, out4.targetGroup)
        else
            triplet ← (null, out4.score, out4.targetGroup)
        end if
        triplets.add(triplet)
    end for
    return triplets
end procedure
```

Figure 3. Construction of group affect network

Background mining stage

For the construction of the CSN, the first issue we need to address is how the multi-agent system understands the overall event within the context of the event. To solve this, we propose the Background Mining Stage, which uses textual information to understand the event and identify potential subgroups. Its functionality can be described as follows:

Input: all the comment texts related to the target topic (sampled if necessary).

Output: a description of the event's background, all potential subgroups present in the topic, and detailed descriptions of each subgroup.

Domain specialist. Domain specialist is primarily responsible for extracting the event background described in the comment texts. Their main tasks include exploring the core event of the topic and key stakeholders. The domain specialist develops a comprehensive description of the event's

timeline and related parties, providing this background information to the subgroup exploration expert and other subsequent stages of the process.

Subgroup exploration expert. The task of the subgroup exploration expert is to use the background information provided by the domain specialist along with the source texts to identify potential subgroups involved in the event and summarize the possible speaking patterns of each subgroup's members. This requires the expert to explore the organizations, stances, religions, and other social identities referenced in the texts and form subgroups based on the similarity of their expressions. It is important to note that if there is a significant amount of unclassifiable content, the expert is permitted to consult human experts for clarification.

Semantic analysis stage

Building on the background information, semantic analysis stage needs to address the primary challenge of accurately interpreting the emotions conveyed in the texts, especially when slang, homophones, sarcasm, and other nuanced expressions are present. To achieve better results on this complex task, we design the system with the specific attributes of social media in mind. Its functionality can be described as follows:

Input: comment texts under the target topic and results from background mining stage.

Output: sentiment analysis result.

Social media veteran. The social media veteran is one of the key agents responsible for semantic understanding of social media content. Its main role is to explore the patterns and characteristics of language expression on social media platforms. The agent needs to interpret the actual meanings of hashtags, internet slang, emojis, and other unique forms of expression commonly used on social media. After this analysis, the social media veteran passes the comprehended information to the sentiment analysis expert for further processing.

Linguistic expert. The linguistic expert is another key agent responsible for the semantic understanding of social media content. Unlike the social media veteran, the linguistic expert focuses on analysing the text from a linguistic perspective, examining aspects such as grammatical structure, rhetorical devices, word choice, and tense. The analysis results are also passed to the sentiment analysis expert. It is important to note that the linguistic expert's analysis is not conducted in isolation; the background information provided by the domain expert supports and informs the linguistic analysis.

Sentiment analysis expert. The sentiment analysis expert is the agent responsible for synthesizing various inputs to determine the final sentiment and its direction. It combines the emotional language present in the text with the semantic analysis results provided by other agents, such as the social media veteran and the linguistic expert, to derive the sentiment of the text. Additionally, it utilizes the subgroup information provided by the subgroup exploration expert to identify the potential target of the sentiment. The results of this sentiment analysis will serve as the output of the sentiment analysis stage and will be passed to the next stage.

Polarization assessment stage

The primary task of the polarization assessment stage is to utilize the information from the background mining stage and the semantic analysis stage to generate CSN in the form of triplets. Its functionality can be described as follows:

Input: all information from the background mining stage (including potential subgroups and event background), each text from the target topic, and their sentiment analysis results.

Output: sentiment expressed in the form of triplets for each comment: (personal stance, sentiment score, target subgroup).

Polarization assessor. The polarization assessor is the core agent of the polarization assessment stage. It is responsible for analysing the author's stance, sentiment score, and the target group of the sentiment for each comment, based on the information passed from the other stages. The polarization assessor integrates this information into triplets. With the multi-dimensional and in-depth analysis provided by the other stages, the polarization assessor can make precise judgments and give credible sentiment score.

CSN construction. To construct the final community sentiment network (CSN) based on the sentiment triplets from the existing comments, we designed the relevant algorithm (see Figure 4, Table 1 explains some variables). We use an adjacency matrix, adjMatrix, to represent the CSN, where adjMatrix[i][j] represents the sentiment score of subgroup i towards subgroup j. We first use all the triplets to build an initial network and then merge the nodes that belong to the same subgroup. During the merging process, we use the number of likes on the comments as a weighting factor, applying a weighted calculation to the sentiment scores between subgroups involved in the sentiment. This results in a total sentiment score between the subgroups. Since not all triplets have a personal stance, we complete them by approximating the occurrence frequency of all known personal stances as probabilities and use these probabilities to fill in the incomplete triplets. These operations result in the final CSN.

Initialize:
Initialize 10x10 matrices $adjMatrix$, $weightSum$ and $countMatrix$ to zero
Initialize 1x10 vector $commentCount$ to zero
$incompleteTriplets \leftarrow []$
Process Complete Triplets:
for each $triplet$ in $triplets$ **do**
    if $triplet.group$ is not null **then**
        $src \leftarrow triplet.group$
        $tgt \leftarrow triplet.targetGroup$
        $weightedScore \leftarrow triplet.score \times \max(triplet.likes, 1)$
        $adjMatrix[src][tgt] \leftarrow adjMatrix[src][tgt] + weightedScore$
        $weightSum[src][tgt] \leftarrow weightSum[src][tgt] + \max(triplet.likes, 1)$
        $countMatrix[src][tgt] \leftarrow countMatrix[src][tgt] + 1$
        $commentCount[src] \leftarrow commentCount[src] + 1$
    else
        $incompleteTriplets.add(triplet)$
    end if
end for
Process Incomplete Triplets:
for each $triplet$ in $incompleteTriplets$ **do**
    $tgt \leftarrow triplet.targetGroup$
    $probabilities \leftarrow []$
    for $i \leftarrow 1$ to 10 **do**
        $probabilities[i] \leftarrow countMatrix[i][tgt] / \sum_{j=1}^{10} countMatrix[j][tgt]$
    end for
    $src \leftarrow$ sample a group based on $probabilities$
    $weightedScore \leftarrow triplet.score \times \max(triplet.likes, 1)$
    $adjMatrix[src][tgt] \leftarrow adjMatrix[src][tgt] + weightedScore$
    $weightSum[src][tgt] \leftarrow weightSum[src][tgt] + \max(triplet.likes, 1)$
    $countMatrix[src][tgt] \leftarrow countMatrix[src][tgt] + 1$
    $commentCount[src] \leftarrow commentCount[src] + 1$
end for
Compute Averages:
for $i \leftarrow 1$ to 10 **do**
    for $j \leftarrow 1$ to 10 **do**
        if $weightSum[i][j] > 0$ **then**
            $adjMatrix[i][j] \leftarrow adjMatrix[i][j] / weightSum[i][j]$
        end if
    end for
end for
Output: Draw the group affect network using $adjMatrix$

Figure 4. Construction of community sentiment network

Variable	Explanation
weightSum	triplets' weight
countMatrix[i][j]	number of triplets whose person stance is i and target subgroup is j
incompleteTriplets	variable for storing incomplete triplets
commentCount	number of comments in every subgroup

Table 1. The explanation of some variables in Figure 4

Community opposition index (COI)

We have designed a dedicated group polarization metric for CSN to derive an interpretable polarization index from its complex graph structure. In previous research on sentiment-based polarization measurement, a widely accepted viewpoint is that the stronger the internal cohesion within subgroups and the greater the hostility between subgroups, the higher the level of polarization (Iyengar et al., 2012; Iyengar et al., 2015; Lelkes et al.,2017; Wakefield et al., 2023; Yarchi et al., 2021). The *sentiment thermometer* was developed based on this perspective, and its approach of calculating sentiment temperature differences has gained widespread recognition and practical use (Boxell et al., 2020; Iyengar et al., 2019; Iyengar et al., 2012; Lelkes et al.,2017; Wakefield et al., 2023). However, as we mentioned earlier, this method is only applicable when there are exactly two subgroups involved in group polarization. Therefore, we extended this calculation method to the multi-group domain and proposed the community opposition index (COI).

Firstly, we calculate the sentiment score of a subgroup towards the other subgroups:

$$\left(-e_{ij}\right) \cdot 1_{e_{ij} \leq 0}.$$

Here, e_{ij} represents the sentiment score of subgroup i towards subgroup j. $1_{eij \leqslant 0}$ means that we consider friendly subgroups as not contributing to the overall group polarization.

Subsequently, we sum the sentiment scores of subgroup i towards all other related subgroups and take into account the internal cohesion within each subgroup. Therefore, we get the polarization score of subgroup i. Here, t_i represents the internal cohesion of the subgroup i:

$$t_i \cdot \sum_j \left(-e_{ij}\right) \cdot 1_{e_{ij} \leq 0}.$$

Finally, we weight the overall sentiment score of each subgroup according to its size and calculate the final polarization score:

$$\sum_i \left(\frac{n_i}{N} \cdot t_i \cdot \sum_j \left(-e_{ij}\right) \cdot 1_{e_{ij} \leq 0} \right).$$

Here N represents the total number of comments on the target topic and n_i represents the number of comments of subgroup i within this topic.

It is important to emphasize that, since our metric is a relative indicator, it can avoid interference in the polarization measurement results caused by differences in the number of comments. Additionally, this metric, by focusing on both the internal and external sentiments of subgroups, offers better interpretability.

Zero-shot experiments

Our experiments will focus on the multi-agent system. Since there is no established benchmark in the field of group polarization research, we have chosen to test the system using stance detection

tasks, which share a similar nature. However, unlike stance detection, where subgroups are predefined, our system autonomously identifies and extracts subgroups, making our task distinct from stance detection. We describe the specific setup of our experiments as follows.

Datasets

Based on existing work in the field of stance detection, we will conduct our experiments on the following three datasets (Augenstein et al., 2016; Liang et al., 2022):

SEM16 (Mohammad et al., 2016). This dataset includes six different targets selected from various domains, namely Donald Trump (DT), Hillary Clinton (HC), Feminist Movement (FM), Legalization of Abortion (LA), Atheism (A), and Climate Change is a Real Concern (CC). It includes three types of stances: Favour, Against, and None.

P-Stance (Li et al., 2021). This dataset includes six different targets selected from political domains, namely Donald Trump (Trump), Joe Biden (Biden), Bernie Sanders (Sanders). It includes three types of stances: Favour and Against.

VAST (Allaway et al., 2020). This dataset includes large number of varying targets, and it includes three types of stances: Pro, Con and Neutral.

The statistics of our utilized datasets are shown in Table 2. Since our model's use case is almost zero-shot, we will utilize these three datasets to conduct zero-shot stance detection. We will strictly adhere to the licensing requirements of the respective datasets.

To better evaluate the model's performance, we selected appropriate metrics based on existing literature (Allaway et al., 2021; Lan et al., 2024; Liu et al., 2021). For the SEM16 and P-Stance datasets, we chose F_{avg}, which represents the average of F1 scores for Favour and Against. For the VAST dataset, we opted for Macro-F1 as the evaluation metric.

Dataset	Target	Pro	Con	Neutral
SEM16	DT	148 (20.9%)	299 (42.3%)	260 (36.8%)
	HC	163 (16.6%)	565 (57.4%)	256 (26.0%)
	FM	268 (28.2%)	511 (53.8%)	170 (17.9%)
	LA	167 (17.9%)	544 (58.3%)	222 (23.8%)
	A	124 (16.9%)	464 (63.3%)	145 (19.8%)
	CC	335 (59.4%)	26 (4.6%)	203 (36.0%)
P-Stance	Biden	3217 (44.1%)	4079 (55.9%)	-
	Sanders	3551 (56.1%)	2774 (43.9%)	-
	Trump	3663 (46.1%)	4290 (53.9%)	-
VAST	-	6952 (37.5%)	7297 (39.3%)	4296 (23.2%)

Table 2. Statistics of our utilized datasets

Model adjustment

Since the primary purpose of our designed multi-agent system is to construct the CSN, we need to adjust the model for the experiments. We removed the subgroup exploration expert and eliminated the subgroup exploration process. Instead, we input texts with predefined target groups into the remaining five agents and obtained the output from the polarization assessor. The adjusted model's output only includes the sentiment score and target group, making it suitable for performing stance detection tasks.

Regarding the specific details of the model configuration, we use multiple GPT-3.5 Turbo models provided by OpenAI to serve as agents in the background mining stage and semantic analysis stage,

while GPT-4 is employed as the polarization assessor. This selection was primarily based on a balance between cost and the desired final performance.

Comparison methods

We compare our method with various methods in zero-shot stance detection. This includes adversarial learning method: TOAD, contrastive learning methods: PT-HCL, Bert-based techniques: TGANet and Bert-GCN, LLM-based techniques: GPT-3.5 Turbo, GPT-3.5 Turbo + Chain of thought (COT) and COLA (Allaway et al., 2021; Liang et al., 2022; Allaway et al., 2020; Liu et al., 2021; Zhang et al., 2022; Zhang et al., 2023; Lan et al., 2024).

Zero-shot stance detection results

In Table 3, we present the performance of our method on the zero-shot stance detection task, along with a comparison to other baselines. The results demonstrate that our method exhibits excellent performance in this task, with a performance improvement of 8.4% over the current best result on the VAST dataset. Although our method did not achieve state-of-the-art (SOTA) results across all metrics, its ability to come close to or surpass current SOTA algorithms indicates its significant value when applied to group polarization research.

Model	**SEM16**						**P-Stance**			**VAST**
-	DT	HC	FM	LA	A	CC	Trump	Biden	Sanders	All
TOAD	49.5	51.2	54.1	46.2	46.1	30.9	53	68.4	62.9	41
TGA Net	40.7	49.3	46.6	45.2	52.7	36.6	-	-	-	65.7
BERT-GCN	42.3	50	44.3	44.2	53.6	35.5	-	-	-	68.6
PT-HCL	50.1	54.5	54.6	50.9	56.5	38.9	-	-	-	71.6
GPT-3.5	62.5	68.7	44.7	51.5	9.1	31.1	62.9	80	71.5	62.3
GPT-3.5+COT	63.3	70.9	47.7	53.4	13.3	34	63.9	81.2	73.2	68.9
COLA	68.5	81.7	63.4	71	70.8	65.5	86.6	**84**	79.7	73
Ours	**74.4***	**81.9**	**70.3***	**75.8***	**76.9***	**70.7**	**87.9**	83.2	**86.2***	**81.4***

Table 3. Comparison of our method and baselines in zero-shot stance detection task, all values are percentages. Bold refer to the best performance. '*' denotes our method improves the best baseline at $p < 0.05$ with paired t-test

Limitations and future works

A major limitation of this study is the lack of a widely accepted quantitative definition of group polarization, which hinders the development of standardized benchmarks and affects the rigor of our experiments. Additionally, while our proposed CSN framework, as a temporal dynamic graph, has potential for modelling the evolution of group polarization using graph neural networks (GNNs), the absence of benchmarks limits further exploration. Future work should prioritize establishing standardized definitions and benchmarks to enable more rigorous experimentation and predictive modelling.

Conclusion

In this paper, we discussed the shortcomings of current group polarization measurement approaches and proposed our multi-agent and graph-based measurement scheme. Our solution innovatively introduces a large language model-based multi-agent system into the measurement of group polarization and utilizes the community sentiment network (CSN) to represent the polarization state. Additionally, we provided a metric (community opposition index) for calculating polarization levels using the CSN, allowing the polarization state to be quantified. Finally, we tested our multi-agent system through a zero-shot stance detection task, and the results demonstrated its usability and value.

Acknowledgements

We appreciate the support from the School of Information Management at Wuhan University, which enabled us to complete this paper as students.

About the authors

Zixin Liu is an undergraduate student at the School of Information Management, Wuhan University, PRC, majoring in Information Management and Information Systems. His primary research interests include knowledge graphs and graph learning. Contact him at lzx2562521178@outlook.com

Ji Zhang is an undergraduate student at the School of Information Management, Wuhan University, PRC, majoring in Information Management and Information Systems. He is the co-first author.

Yiran Ding is an undergraduate student at the School of Information Management, Wuhan University, PRC, majoring in Information Management and Information Systems.

References

Allaway, E., & McKeown, K. (2020). Zero-shot stance detection: A dataset and model using generalized topic representations. arXiv preprint arXiv:2010.03640. https://arxiv.org/abs/2010.03640

Allaway, E., Srikanth, M., & McKeown, K. (2021). Adversarial learning for zero-shot stance detection on social media. arXiv preprint arXiv:2105.06603. https://arxiv.org/abs/2105.06603

Augenstein, I., Rocktäschel, T., Vlachos, A., & Bontcheva, K. (2016). Stance detection with bidirectional conditional encoding. arXiv preprint arXiv:1606.05464. https://arxiv.org/abs/1606.05464

Belcastro, L., Cantini, R., Marozzo, F., Talia, D., & Trunfio, P. (2020). Learning political polarization on social media using neural networks. *IEEE Access*, 8, 47177–47187.https://doi.org/10.1109/ACCESS.2020.2972932

Bilal, M., Gani, A., Marjani, M., & Malik, N. (2019). Predicting elections: Social media data and techniques. In 2019 *International Conference on Engineering and Emerging Technologies* (ICEET) (pp. 1-6). Springer. https://doi.org/10.1109/CEET1.2019.8711854

Bomsdorf, E., & Otto, C. (2007). A new approach to the measurement of polarization for grouped data. AStA *Advances in Statistical Analysis*, 91(2), 181–196. https://doi.org/10.1007/s10182-007-0047-x

Boxell, L., Conway, J., Druckman, J. N., & Gentzkow, M. (2020). Affective polarization did not increase during the coronavirus pandemic. National Bureau of Economic Research. https://www.nber.org/papers/w27133

Bravo, R. B., Del Valle, M. E., & Gavidia, À. R. (2015). A multilayered analysis of polarization and leaderships in the Catalan Parliamentarians' Twitter Network. In 2015 *Fifteenth International Conference on Advances in ICT for Emerging Regions* (ICTer) (pp. 200–206). IEEE. https://doi.org/10.1109/ICTER.2015.7377689

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv*:2005.14165. https://arxiv.org/abs/2005.14165

Conover, M. D., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on Twitter. In *Proceedings of the international AAAI conference on web and social media* (Vol. 5, No. 1, pp. 89-96). https://doi.org/10.1609/icwsm.v5i1.14126

Garcia, D., Abisheva, A., Schweighofer, S., Serdült, U., & Schweitzer, F. (2015). Ideological and temporal components of network polarization in online political participatory media. *Policy & Internet*, 7(1), 46–79. https://doi.org/10.1007/s12132-014-0185-7

Gaurav, M., Srivastava, A., Kumar, A., & Miller, S. (2013). Leveraging candidate popularity on Twitter to predict election outcome. In *Proceedings of the 7th workshop on social network mining and analysis* (pp. 1-8). Association for Computing Machinery. https://doi.org/10.1145/2501025.2501038

Guerra, P., Meira Jr, W., Cardie, C., & Kleinberg, R. (2013). A measure of polarization on social media networks based on community boundaries. In *Proceedings of the international AAAI conference on web and social media* (Vol. 7, No. 1, pp. 215-224). https://doi.org/10.1609/icwsm.v7i1.14421

Hart, P. S., Chinn, S., & Soroka, S. (2020). Politicization and polarization in COVID-19 news coverage. *Science Communication*, 42(5), 679–697. https://doi.org/10.1177/1075547020926489

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, 50(6), 1141–1151. https://doi.org/10.1037/0022-3514.50.6.1141

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, 22(1), 129–146. https://doi.org/10.1146/annurev-polisci-050518-030121

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, 76(3), 405–431. https://doi.org/10.1093/poq/nfs069

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3), 690–707. https://doi.org/10.1111/ajps.12165

Jaidka, K., Ahmed, S., Skoric, M., & Hilbert, M. (2018). Predicting elections from social media: A three-country, three-method comparative study. *Asian Journal of Communication*, 29(3), 252–273. https://doi.org/10.1080/01292986.2018.1455448

Jiang, J., Ren, X., & Ferrara, E. (2023). Retweet-BERT: Political leaning detection using language features and information diffusion on social networks. In *Proceedings of the international AAAI conference on web and social media* (Vol. 17, pp. 459-469). https://doi.org/10.1609/icwsm.v17i1.22160

Jiang, S., & Wilson, C. (2018). Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1–23. https://doi.org/10.1145/3274357

Knox, R. E., & Inkster, J. A. (1968). Postdecision dissonance at post time. *Journal of Personality and Social Psychology*, 8(4, Pt. 1), 319. https://doi.org/10.1037/h0025557

Kogan, N., & Wallach, M. A. (1967). Risky-shift phenomenon in small decision-making groups: A test of the information-exchange hypothesis. *Journal of Experimental Social Psychology*, 3(1), 75–84. https://doi.org/10.1016/0022-1031(67)90025-4

Lan, X., Gao, C., Jin, D., & Li, Y. (2024). Stance detection with collaborative role-infused LLM-based agents. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 18, pp. 891-903). https://doi.org/10.1609/icwsm.v18i1.31360

Lelkes, Y., & Westwood, S. J. (2017). The limits of partisan prejudice. The Journal of Politics, 79(2), 485-501. https://doi.org/10.1086/688223

Li, A., Liang, B., Zhao, J., Zhang, B., Yang, M., & Xu, R. (2023). Stance detection on social media with background knowledge. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 15703-15717). https://doi.org/10.18653/v1/2023.emnlp-main.972

Liang, B., Chen, Z., Gui, L., He, Y., Yang, M., & Xu, R. (2022). Zero-shot stance detection via contrastive learning. In Proceedings of the ACM Web Conference 2022 (pp. 2738-2747). https://doi.org/10.1145/3485447.3511994

Liu, R., Lin, Z., Tan, Y., et al. (2021). Enhancing zero-shot and few-shot stance detection with common sense knowledge graph. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 3152-3157). https://doi.org/10.18653/v1/2021.findings-acl.278

Li, Y., Sosea, T., Sawant, A., Nair, A. J., Inkpen, D., & Caragea, C. (2021). P-stance: A large dataset for stance detection in political domain. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 2355-2365). https://doi.org/10.18653/v1/2021.findings-acl.208

Lu, H. C., & Lee, H. W. (2024). Agents of Discord: Modeling the Impact of Political Bots on Opinion Polarization in Social Networks. Social Science Computer Review. Advance online publication. https://doi.org/10.1177/08944393241270382

Maia, H. P., Ferreira, S. C., & Martins, M. L. (2023). Controversy-seeking fuels rumor-telling activity in polarized opinion networks. Chaos, Solitons & Fractals, 169, 113287. https://doi.org/10.1016/j.chaos.2023.113287

Medaglia, R., & Zhu, D. (2017). Public deliberation on government-managed social media: A study on Weibo users in China. Government Information Quarterly, 34(3), 533-544. https://doi.org/10.1016/j.giq.2017.02.002

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). SemEval-2016 Task 6: Detecting stance in tweets. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 31-41). https://doi.org/10.18653/v1/S16-1003

Ribeiro, M. H., Calais, P. H., Almeida, V. A., & Meira, W., Jr. (2017). "Everything I disagree with is #FakeNews": Correlating political polarization and spread of misinformation. arXiv preprint arXiv:1706.05924. https://arxiv.org/abs/1706.05924

Santos, F. P., Lelkes, Y., & Levin, S. A. (2021). Link recommendation algorithms and dynamics of polarization in online social networks. Proceedings of the National Academy of Sciences, 118(50), e2102141118. https://doi.org/10.1073/pnas.2102141118

Stoner, J. A. F. (1961). A comparison of individual and group decisions involving risk (Doctoral dissertation, Massachusetts Institute of Technology).

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 4, No. 1, pp. 178-185). https://doi.org/10.1609/icwsm.v4i1.14009

Tyagi, A., Uyheng, J., & Carley, K. M. (2020). Affective Polarization in Online Climate Change Discourse on Twitter. In 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 443-447). IEEE. https://doi.org/10.1109/ASONAM49781.2020.9381419

Vicario, M. D., Gaito, S., Quattrociocchi, W., Zignani, M., & Zollo, F. (2017). News Consumption during the Italian Referendum: A Cross-Platform Analysis on Facebook and Twitter. In 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 648-657). IEEE. https://doi.org/10.1109/DSAA.2017.33

Wakefield, R. L., & Wakefield, K. (2023). The antecedents and consequences of intergroup affective polarization on social media. Information Systems Journal, 33(3), 640-668 https://doi.org/10.1111/isj.12319

Xiao, Z., Song, W., Xu, H., Ren, Z., & Sun, Y. (2020). TIMME: Twitter ideology-detection via multi-task multi-relational embedding. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2258-2268). https://doi.org/10.1145/3394486.3403275

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. Political Communication, 38(1-2), 98-139. https://doi.org/10.1080/10584609.2020.1847458

Zhang, B., Ding, D., & Jing, L. (2022). How would stance detection techniques evolve after the launch of ChatGPT? arXiv preprint arXiv:2212.14548. https://arxiv.org/abs/2212.14548

Zhang, B., Fu, X., Ding, D., Huang, H., Li, Y., & Jing, L. (2023). Investigating chain-of-thought with ChatGPT for stance detection on social media. arXiv preprint arXiv:2304.03087. https://arxiv.org/abs/2304.03087

Zhang, C., Song, D., Huang, C., Swami, A., & Chawla, N. V. (2019). Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 793-803). Association for Computing Machinery. https://doi.org/10.1145/3292500.3330961