Context-Aware Sentiment Forecasting via LLM-based Multi-Perspective Role-Playing Agents

Fanhang Man[1], **Huandong Wang**[2], **Jianjie Fang**[3], **Zhaoyi Deng**[4],
Baining Zhao[1], **Xinlei Chen**[1*], **Yong Li**[2]
[1]Shenzhen International Graduate School, Tsinghua University,
[2]Department of Electronic Engineering, Tsinghua University,
[3]Northeastern University at Qinghuangdao,
[4]Department of Computer Science, University of California, Irvine
mfh21@mails.tsinghua.edu.cn, wanghuandong@tsinghua.edu.cn,
chen.xinlei@sz.tsinghua.edu.cn, liyong07@tsinghua.edu.cn

Abstract

User sentiment on social media reveals underlying social trends, crises, and needs. Researchers have analyzed users' past messages to track the evolution of sentiments and reconstruct sentiment dynamics. However, predicting the imminent sentiment response of users to ongoing events remains understudied. In this paper, we address the problem of **sentiment forecasting** on social media to predict users' future sentiment based on event developments. We extract sentiment-related features to enhance modeling and propose a multi-perspective role-playing framework to simulate human response processes. Our preliminary results show significant improvements in sentiment forecasting at both microscopic and macroscopic levels.

1 Introduction

The study of sentiments on social media has facilitated social research (Levy et al., 2022), marketing (Zhang et al., 2021), and public management (Solovev and Pröllochs, 2022). For instance, during public emergencies, measuring negative sentiment can help disaster relief organizations prioritize areas experiencing heightened collective distress (Zhang et al., 2020). The task of **sentiment forecasting** timely anticipates the sentiment of a person or a crowd with the available information on social media. While distinct from sentiment analysis and other related tasks, forecasting complements these approaches by providing forward-looking insights that enhance our understanding of sentiment dynamics and enable proactive decision-making.

Retrospective sentiment analysis methods typically assign a discrete or linear sentiment score to a sentence, paragraph, or document (Mousavi et al., 2022). Researchers leveraged deep neural networks (Hu and Flaxman, 2018), Transformer (Zhong et al., 2021), and Bert (Islam and Bhattacharya, 2022) to extract users' sentiment states. The evolution of sentiments is also heavily studied (Tu and Neumann, 2022). Okawa and Iwata (2022) leveraged a sociologically informed neural network (SINN) to deduce the users' sentiments as they evolve. Liu and Yang (2022) integrated the DeGroot Model with a probabilistic linguistic method to simulate people's sentiments. However, these methods merely consider the reciprocal influences among social media users. In the real-world scenario, the nature of human sentiment is highly context-dependent (Kuppens and Verduyn, 2017; Halim et al., 2020; Hu and Flaxman, 2018). The development of ongoing social events highly affects social media users' sentiments. However, the ongoing events have complex semantic information, which is too intricate to be formulated as an input for the above methods.

The advent of large language models (LLMs) sheds light on comprehensive and prospective sentiment reasoning in real-world scenarios (Chang et al., 2023; Zha et al., 2025). LLMs understand complex external event contexts with embedded common-sense knowledge (Zhang et al., 2025; Zhao et al., 2025). Zhang et al. (2023) fine-tuned GPT to analyze the sentiment with rich contexts. Moreover, LLMs can capture the nuanced tone of voice in texts including sarcasm, humor, and rhetorical questions (Wang et al., 2023b; Safdari et al., 2023). Deng et al. (2023) applied such an advantage to analyze sentiment for Reddit. These works formalized sentiment analysis as a reasoning problem to integrate semantic information and subjectivity in human opinion (Hou et al., 2024). Such advantages provide an opportunity to pivot from traditional retrospective studies of sentiment analysis to prospective studies of sentiment forecasting.

However, accurately forecasting sentiment for diverse social media users remains challenging. First, it requires comprehensive user features for behav-

*Corresponding Author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2687–2703

ioral modeling (Iosifidis and Ntoutsi, 2017). Even though some social media users report their own attributes like locations, personality, religion, etc. Other important user-specific features are difficult to obtain due to anonymity and privacy regulations. Second, even with sufficient labels, the evolution of human sentiments is too intricate and subtle to model (Kuppens and Verduyn, 2017; Wang et al., 2022). Individuals in different circumstances exhibit diverse sentiment responses to the same events (Hu et al., 2024). Yet it is challenging to model these subtle sentimental clues.

This paper addresses **Sentiment Forecasting** to predict people's future sentiments on social media in response to real-world events. To incorporate context comprehensively, we leverage LLMs to understand the complex semantic information of the context. To enrich the features for user-specific sentiment modeling, we extract features from users' social media comments, specifically the textual tone of voice and attitude toward the event. To address the complexity and variety of sentiment evolution, we develop a multiperspective role-playing framework to forecast the users' social media comments. Specifically, the subjective role-playing agents simulate the social media user to express oneself on social media. A fine-tuned objective role-playing LLM with expertise in behavioral psychology analyzes the response and provides feedback for reflections, ensuring consistency. The proposed framework addresses the complexity of user-specific sentiment forecasting, enabling precise and nuanced predictions of both individual and collective sentiments.

Our contributions lie in the following aspects:

- We focus on **Sentiment Forecasting** to predict users' future sentiment responses to ongoing real-world events, formulating it as a reasoning problem incorporating external context.
- We implement an LLM-based feature extraction method targeting implicit features (tone of voice, attitude) to simulate user sentiment responses.
- We propose a multi-perspective role-playing framework for predicting future sentiment responses. The subjective role-playing agent simulates users with extracted features to generate responses, while the objective role-playing agent ensures behavioral consistency.

2 Related Work

2.1 The Study of Sentiment

The study of sentiments is crucial in understanding the public opinion in diverse event settings, including public events (Solovev and Pröllochs, 2022; Li et al., 2022), natural disasters (Li et al., 2024a), and livelihood (Li et al., 2024b; Liu et al., 2024). Researchers use agent-based modeling and network theory to model the changes in collective sentiment (Castellano et al., 2009). For example, Okawa and Iwata (2022) leveraged a sociologically-informed neural network to track and predict the evolution of user sentiments over time. Liu and Yang (2022) integrated the DeGroot Model with a probabilistic linguistic method to forecast people's decisions during the COVID-19 pandemic. (Monti et al., 2020) fits the agent model with real-world social traces to recover the real-world dynamics. The above methods merely construct the evolution based on mutual interactions of the social media users without any event context, compromising applicability in real-world scenarios. Our proposed framework adopted the development of ongoing real-world events as context and user-specific circumstances as sentiment clues to tackle the problem of timely real-world deployment.

2.2 Role-Play with LLMs

LLMs exhibit sophisticated dialogue behavior (Abbasiantaeb et al., 2024). Shanahan et al. (2023) introduced the idea of role-play to characterize the phenomena for LLMs to perform the part of a person or a superposition of simulacra within a multiverse of possible characters. Role-play methods let the LLM act as the user themselves to convey the dialogue. Several researchers therefore fine-tuned the LLMs to role-play characters from entertainment works, including animations (Li et al., 2023), TV series (Zhou et al., 2023; Wang et al., 2023a), and Movies (Chen et al., 2022). Others also applied such a technique to recreate notable historical figures (Zhou et al., 2023; Shao et al., 2023). However, existing methods incorporated massive character-centered data, e.g., personality, social status, and relationships, to train the LLMs to role-play the character. In the online social media scenario, it is unrealistic and unethical to obtain such data from social media users. Though inspiring, such methods could only be used to predict some specific fabricated character's reaction to events with limited generalizability in real-life social media.

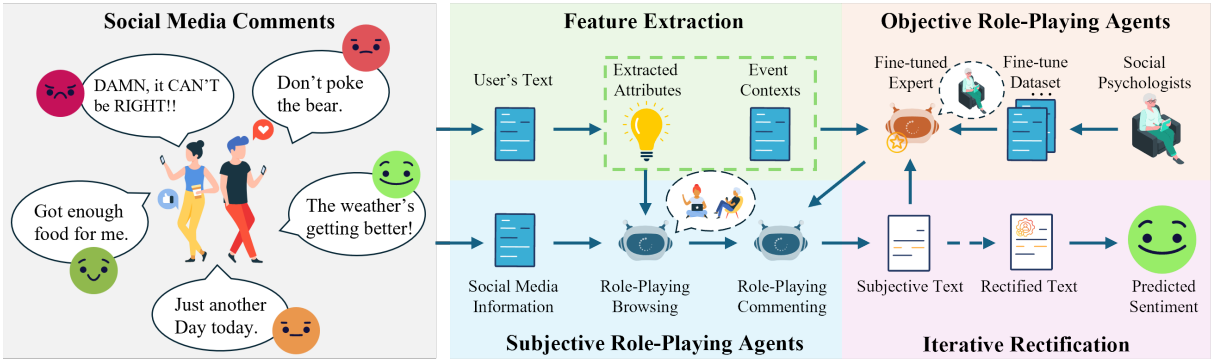

Figure 1: Architecture of the proposed LLM-based multi-perspective role-playing framework.

3 Methodology

In this section, we first formally define the problem of sentiment forecasting. Then we outline the proposed context-aware sentiment forecasting framework (Figure 1), followed by detailed descriptions of each core component: feature extraction, subjective role-playing agent, objective role-playing agent, and iterative rectification. Formal notation definitions are summarized in Table 1.

3.1 Problem Formulation

For social media, traditional sentiment analysis operates retrospectively, mapping an existing user comment c to a sentiment space via:

$$\sigma = F_{\text{SA}}(c), \tag{1}$$

where F_{SA} denotes the sentiment analysis function.

However, sentiment forecasting constitutes a prospective temporal reasoning task that requires systematic logical inferences, including identification of historical patterns, contextual cues, and future sentiment trends. Formally, sentiment forecasting aggregates historical semantic information up until time t to predict the future sentiment at the time of interest t', where $t' \geq t$. With the ability to gather social media comments, sentiment forecasting can be timely conducted as the ongoing event evolves. The task of sentiment forecasting is formalized as follows:

$$\sigma_{t'} = F_{\text{SF}}(\bigcup_{\tau \leq t} \cdot_\tau), \tag{2}$$

where $\cdot_\tau$ indicates the information at time τ and F_{SF} is the function of sentiment forecasting.

3.2 Overall Framework

Sentiment forecasting aims to conjecture the future sentiment of a social media user or crowd. To effectively integrate the semantic information from user attributes, user comments, and event context for sentiment forecasting, we develop an LLM-based multi-perspective role-playing framework. The subjective agent generates future comments where the sentiment lies. The objective agent applies knowledge in behavioral psychology to discriminate abnormality in the generated comments to restrict stochasticity.

As shown in Figure 1, the framework comprises four components: feature extraction, subjective role-playing agent, objective role-playing agent, and iterative rectification. Feature extraction aims to identify implicit features derived from existing social media comments. This process is designed to capture the user's habitual textual tone of voice and infer the user's attitude towards the ongoing event. With the extracted features, the subjective role-play agent simulates the behavior of a social media user to comprehend the specific contexts and skim through the followees' social media comments. Subsequently, the agent is instructed to generate a new comment at the future time of interest t' regarding the event. To ensure consistency in the user's textual tone of voice and attitude flow, a fine-tuned objective role-play agent serves as a behavioral psychologist to review this generated comment to filter potential behavioral inconsistency. During iterative rectification, the analysis from the objective agent is fed back to the subjective role as a guide to iteratively regenerate a rectified comment. Finally, the forecasted sentiment is retrieved with the state-of-the-art sentiment analysis method. Further details are provided in the subsections below.

3.3 Feature Extraction

The feature extraction model aims to retrieve social media users' sentiment-centered elements. In ad-

dition to the self-reported user attributes $\mathcal{A}$, which includes gender, religion, location, etc., the implicit sentiment-centered aspects underlined in the social media comments are also critical. We define the set of all (attainable) social media comments of a user u to be $\mathcal{C}^u = \{c^u_\tau | \tau \in T^u\}$, where c^u_τ is a specific comment made by u at time τ and T^u is the set of all time points at which user u has made comments. Since the comment is predicted for the future time of interest $t' \geq t$, we are specifically interested in comments $\mathcal{C}^u_t = \{c^u_\tau | \tau \in T^u, \tau \leq t\}$ before t during the feature extraction to avoid spoiled information.

In case of social media comments, a user's intentional choice of lexical, syntactic, and paralinguistic elements reflects the unique way of expression and persona. Such a choice is defined as the **textual tone of voice** ν. Studies demonstrate that people tend to maintain a consistent textual tone of voice according to their social image on social media (Cingi et al., 2023; Rettberg et al., 2017). The textual tone of voice is extracted via an LLM, which is formalized as follows:

$$\nu^u_t = \text{LLM}(\mathcal{C}^u_t, i_\nu), \tag{3}$$

where ν^u_t indicates the extracted tone of voice with information before t. i_ν denotes the instruction of analyzing the textual tone of voice. The output is structured as three descriptive adjectives. Examples are demonstrated in Appendix A.1.

In addition, attitude is a psychological construct representing an individual's enduring evaluative stance towards certain aspects. In our case, we address the user attitude toward public events. The event context $\mathcal{E}^u_t$, experienced or witnessed by user u before t, serves as an important factor when inferring user attitude. Unlike the high consistency in the textual tone of voice, user attitude may reasonably shift as the event evolves. Yet the textual tone of voice highly affects the expression of user attitude. Therefore, user attitude is extracted as follows:

$$\alpha^u_t = \text{LLM}(\mathcal{C}^u_t, \mathcal{E}^u_t, \nu^u_t, i_\alpha), \tag{4}$$

where α^u_t represents user attitude of user u before time t. i_α represents the designed instruction to analyze user attitude toward the event.

3.4 Subjective Role-playing Agent

The subjective role-play strategy instructs the LLM to role-play a social media user based on the extracted user features. We strictly limit the input to before time t to role-play the user. The role-playing LLM is instructed to maintain the textual tone of voice and avoid unreasonable attitude shifts. The user's social media comments $\mathcal{C}^u_t$ before t are provided as few-shot learning samples. The subjective role-playing agent LLM^s_t is formalized as:

$$\text{LLM}^s_t = \text{RP}(\mathcal{A}^u_t, \mathcal{C}^u_t, \mathcal{E}^u_t, \nu^u_t, \alpha^u_t, i_r), \tag{5}$$

where RP indicates the role-playing process for LLM^s_t to role-play a subjective social media user with the information before time t. i_r is the instruction for role-play. $\mathcal{A}^u_t$ is the attainable self-reported user attribute of user u on social media.

With the role-playing LLM, we could simulate the reactive process with two stages: browsing social media and commenting. A sample of the followees' social media comments $\mathcal{F}^u_t$ is extracted based on empirical media influence factors, such as relevance to the topic, frequency of interaction, number of followers, etc. The role-playing LLM browses through these comments and predicts a comment to be sent at the time of interest t'. These processes can be integratively formalized as follows:

$$\phi^u_{t'} = \text{LLM}^s_t(\mathcal{F}^u_t, \mathcal{E}^u_t, t', i_s), \tag{6}$$

where $\phi^u_{t'}$ is the generated future comment to be posed at time t' from a user regarding the ongoing event. i_s is the instruction for the act of browsing and commenting.

3.5 Objective Role-playing Agent

To prevent unusual stochastic behaviors, we fine-tune an objective psychologist LLM to critically analyze the generated comments.

Symbol	Description of Notations
u	Social media user indicator
$\mathcal{C}^u$	Social media comments of u
$\mathcal{T}^u$	The set of time points u made comments
$\mathcal{A}^u$	Self-reported attributes of u
$\mathcal{F}^u$	u's Followees' social media comments
$\mathcal{E}^u$	Context of the event experienced by u
i	Instruction for a specific task
ν	Textual tone of voice
α	Attitude towards the event
ϕ	Predicted social media comment
θ	Behavioral psychological analysis
σ	Sentiment indicator

Table 1: Definition of Notations

Constructing fine-tune dataset: We first gathered ten sets of social media comments from Twitter/X. Noted that these samples are carefully gathered to avoid overlap among events during the testing. The event context spans natural disasters, political events, social events, etc. With the method detailed in Sections 3.3 and 3.4, we extracted ν_t^u and α_t^u from the comments and generated the predicted social media comment $\phi_{t'}^u$. Then, we enlisted three experts in behavioral psychology. The experts were asked to independently assess whether the LLM_t^s-generated comments maintained a consistent textual tone of voice and reasonably coherent attitude flow with the original user comments C_t^u. Their "yes" or "no" judgments were followed by brief written analyses of observed consistencies/inconsistencies. Agreement analysis revealed a 70% inter-expert percentage agreement and a Fleiss' Kappa (Fleiss et al., 1981) of 0.796, indicating substantial agreement. Both metrics demonstrate strong consensus, supporting the reliability of our annotations. With their analyses as few-shot samples, we construct over 25,000 sets of such reviews with GPT 4o.

Finetuning psychologist LLM: The reviews are utilized for low-rank adaptation (LoRA) supervised fine-tuning, wherein the pre-trained model weights are kept intact and the weight matrices are modified using low-rank decomposition (Hu et al., 2021). This approach increases the task-specific parameter gains, embedding expert behavioral psychology knowledge into the fine-tuned model, functioning as an objective behavioral psychologist. The fine-tuning process is demonstrated as follows:

$$\lambda^{(m+1)} = \lambda^{(m)} - \eta \nabla_\lambda L(\lambda^{(m)}), \tag{7}$$

where $\lambda^{(m)}$ represents the parameters at the m-th iteration, and $L(\lambda)$ is the loss function that measures the prediction error of the model on the given task. By computing the gradient $\nabla_\lambda L(\lambda^{(t)})$, we determine how the parameters should be updated to reduce the error. The learning rate η determines the size of each update step. Further details of the fine-tuning process are illustrated in Appendix A.2.

Consistency analysis: The fine-tuned model is defined as LLM_t^o, role-playing a behavioral psychologist with event context untill time t. LLM_t^o is assigned to analyze the textual tone of voice consistency between the LLM_t^s generated social media comment $\phi_{t'}^u$ and the user's previous social media comment $\mathcal{C}_t^u$ and oversee potential unreasonable attitude shift.

$$\theta_{t'}^u = \text{LLM}_t^o(\mathcal{C}_t^u, \nu_t^u, \alpha_t^u, \phi_{t'}^u, t', i_o), \tag{8}$$

where $\theta_{t'}^u$ is the analysis from the fine-tuned psychologist LLM for the generated comment $\phi_{t'}^u$, and i_o is the instruction for consistency analysis.

3.6 Iterative Rectification

The objective LLM would determine if the generated comment demonstrates consistency. A comment that passes the consistency test is treated as the comment to be sent at time t'. For inconsistent comments, the analysis $\theta_{t'}^u$ from psychologist LLM_t^o is fed to the subjective role LLM^s along with its generated comment $\phi_{t'}^u$ for a more consistent comment regeneration, which is formalized as follows:

$$\phi_{t'}^u = \text{LLM}_t^s(\mathcal{F}_t^u, \mathcal{E}_t^u, \theta_{t'}^u, \phi_{t'}^u, t', i_g), \tag{9}$$

where $\phi_{t'}^u$ on the left side of the equation is the regenerated comment. $\phi_{t'}^u$ on the right side of the equation is the previously generated comment, which is determined to be inconsistent by the objective agent LLM_t^o. i_g indicates the instruction for rectification.

The regenerated comments would be iteratively analyzed by the psychologist LLM_t^o with a limited number of iterations

4 Experiment

4.1 Experiment Settings

Datasets: We conducted extensive experiments on two datasets corresponding to two large-scale socially-aware events, i.e., the 2012 Hurricane Sandy and the 2020 U.S. Presidential Election. To help understand the course of events, brief summaries of both datasets and events are shown in Appendix A.3. Both datasets were collected from Twitter (renamed to X in 2023).

Both datasets allow in-depth analysis not only of the text itself but also of the temporal-spatial attributes, providing valuable insights into social media behaviors during the event. The sentiment label is obtained using the state-of-the-art supervised learning model developed by NLPtown, bert-base-multilingual-uncased-sentiment[1], which achieves an accuracy as high as 87% on multiple datasets (Sahoo et al., 2023; Contreras Hernández et al., 2023).

[1] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

Table 2: Macroscopic performance comparison of two datasets, evaluated using JSD. Bold indicates the best (lowest) JSD score, and underlining denotes the second-best result.

Dataset	2012 Hurricane Sandy								2020 U.S. Election			
Metrics	Distribution of Sentiment				Distribution of Polarity				Dis. of Sen.		Dis. of Pol.	
Location	New Jersey		New York		New Jersey		New York		/			
Time	T1	T2	T1	T2	T1	T2	T1	T2	T3	T4	T3	T4
Voter	0.2822	0.3234	0.2245	0.2581	0.2543	0.2423	0.2454	0.2341	0.1289	0.1129	0.1030	0.1006
DeGroot	0.2376	0.1869	0.2511	0.2243	0.2082	0.1744	0.1310	0.1213	0.1854	0.1738	0.1204	0.1183
SLANT+	0.3630	0.30355	0.2912	0.2874	0.2516	0.2681	0.2905	0.2734	0.1490	0.1408	0.1331	0.1243
NN	0.1904	0.2161	0.1733	0.1804	0.1517	0.1379	0.1270	0.1355	0.0482	0.0441	0.0267	0.0228
SINN	0.1673	0.1359	0.1504	0.1377	0.1201	0.1312	0.1022	0.1216	0.0554	0.0625	0.0294	0.0363
MPR_G	0.0243	**0.0105**	0.0456	0.0396	0.0211	**0.0064**	0.0290	0.0239	**0.0097**	**0.0053**	**0.0024**	**0.0013**
MPR_M	**0.0148**	0.0192	**0.0220**	**0.0313**	**0.0114**	0.0100	**0.0116**	**0.0204**	0.0106	0.0068	0.0038	0.0017

In addition, for each event, we picked two distinct time points to perform sentiment forecasting (T1 and T2 for 2012 Hurricane Sandy, T3 and T4 for the 2020 U.S. Election). For the 2012 Hurricane Sandy Dataset, T1 is the time immediately after Sandy hit New Jersey on Oct. 29, 2012. T2 is one week later on Nov. 5, 2012, with significant post-disaster relief. For the 2020 U.S. Election dataset, T3 represents the time after the second presidential debate on Oct. 29, 2020. T4 represents the time after President-elect Biden claimed victory on Nov. 7, 2020. These markers are chosen to align with major developments, ensuring rich contextual relevance, and are separated by at least a week to avoid overlap and isolate distinct phases of the events. Details of data preprocessing are elaborated in Appendix A.4.

Implementation Details: For the subjective role-playing agent, we applied Gemma 2 9B (Team et al., 2024) and Mistral NeMo 12B (Jiang et al., 2023) to role-play the social media users and generate future comments at the time of interest. Such choices are because the popular models like the GPT series are designed not to process or generate inappropriate or offensive content, which is prevalent on social media. Moreover, testing both models shows the robustness of our proposed frameworks. For the objective role-playing agent, we leveraged Llama 3 8B Instruct. The learning rate η is set to be 1×10^{-4} to prevent acute update while maintaining an accessible rate of convergence (Hu et al., 2021). For iterative rectification, the limit of iterations is set to be 3 to balance the computational efficiency and the effectiveness of the module.

Baselines: The dynamics of information and sentiment propagation of social media have been a heavily researched topic. We compare it against state-of-the-art methods, including social model-based and neural network-based methods. (1) **Voter** adapts the users' sentiment from their followees (Muslim et al., 2024). (2) **DeGroot** assumes users update their sentiment iteratively to the weighted average of the followee's sentiment (Degroot, 1974; Wu et al., 2023). (3) **SLANT+** is a non-linear generative model applying a recurrent neural network (RNN) with the point process model to learn the non-linear evolution of the user's sentiment (Kulkarni et al., 2017). (4) **NN** is a pure neural network method to learn the evolution of users' sentiments based on previous patterns (De et al., 2016). (5) **SINN** is the Sociologically-informed Neural Network model, which harnesses sociological models to guide neural networks to approach users' sentiment evolution (Okawa and Iwata, 2022).

Metrics: We evaluate the performance of sentiment forecasting with real-world data on both microscopic and macroscopic levels. The microscopic evaluation measures if the prediction is accurate for each individual user with accuracy and Macro F1 scores. The macroscopic evaluation focused on the distribution of a general crowd sharing similar characteristics. We adopted the Jensen-Shannon divergence (JSD) to measure the similarity between the distribution of the forecasted sentiment and the ground truths for a crowd, which can be represented as follows:

$$\text{JSD}(\mathbf{p}||\mathbf{q}) = \frac{1}{2}\text{KL}(\mathbf{p}||\frac{\mathbf{p}+\mathbf{q}}{2}) + \frac{1}{2}\text{KL}(\mathbf{q}||\frac{\mathbf{p}+\mathbf{q}}{2}), \tag{10}$$

where $\mathbf{p}$ and $\mathbf{q}$ are two distributions and $\text{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence. The lower the JSD, the closer the distribution between the prediction and the ground truth.

Inspired by sentiment analysis works with fine granularity, we used the distribution of sentiment and the distribution of sentiment polarity p =

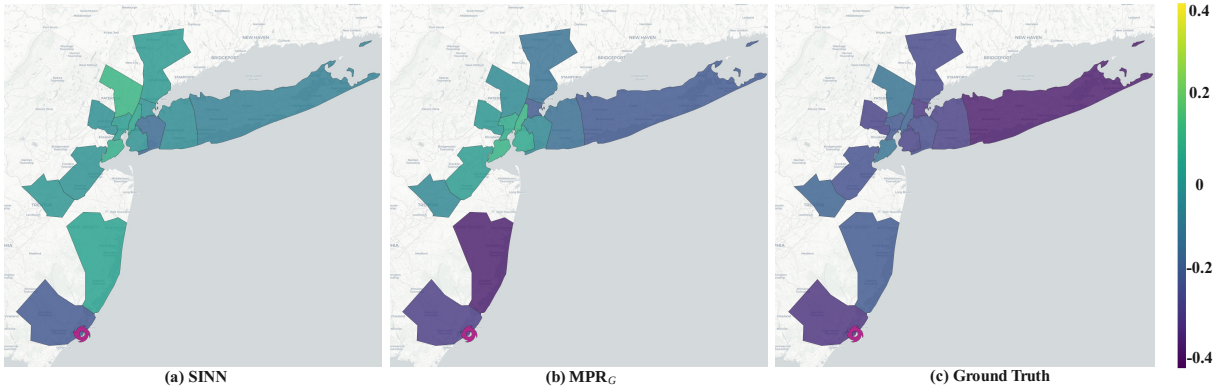

Figure 2: The collective sentiment level for selected New York and New Jersey counties. The red circle indicates the location of the Landfall of Hurricane Sandy. Figures (a) and (b) are the reconstructed sentiment distribution maps for SINN and our proposed MPR$_G$, respectively. Figure (c) is the ground truth distribution.

sgn(s) to measure the fidelity of the forecast. Sentiment has five categories {-2, -1, 0, 1, 2} ranging from strongly negative to strongly positive. Sentiment polarity is on a scale of {-1, 0, 1} where only the polarity of the sentiment is considered.

4.2 Experiment Results

Macroscopic Performance: Table 2 presents the performance of our models. The Gemma and Mistral variants of our multi-perspective role-play (MRP) framework are denoted as MPR$_G$ and MPR$_M$, respectively. For the macroscopic JSD metric, our proposed framework is an order of magnitude better than the baselines, showing a significantly closer distribution of sentiment w.r.t. the ground truth.

The substantial performance improvement stems from a better understanding of context and better user modeling. The model-based methods tend to converge in the long run, with a practically fixed distribution. Learning-based methods also rely heavily on past sentiment scores, especially initialization. It is practically unrealistic for a user in SINN to shift drastically from -2 to 2 or vice versa. However, it is a normal behavior on social media. Among the baselines, SINN achieves the best performance since it adopts social models, i.e. Stochastic bound confidence model, to mitigate the purely data-driven neural networks. A social model could only focus on the dynamics among the social media users, but not the environmental context, which evolves as the development of the event. The proposed MPR framework, however, considers the context information through reasoning. As shown in Table 3, our integration of context information enhances the accuracy of sentiment forecasting when events shift drastically. Hurricane Sandy's landfall significantly altered the living environment for social media users. In the case of the U.S. election, when the swing states were claimed by a president-elect, it would deterministically change the outcome, dramatically affecting people's sentiment.

Moreover, for the 2012 Hurricane Sandy dataset, we analyzed users from different geographical regions to assess group-specific performance. As shown in Figure 2, we compared collective sentiment across 15 selected counties in New York and New Jersey. While SINN predicted more positive collective sentiments, our approach better captured the sentiment shifts before and after Hurricane Sandy's landfall, leading to more accurate collective sentiment forecasting.

Microscopic Performance: In addition to the crowd-level analysis, we also heavily tested the performance of our framework at an individual level. As shown in Table 3, our proposed framework outperformed the baselines for both Gemma 2 and Mistral NeMo. With Gemma 2, our proposed framework achieved an average of 6.23% improvement in accuracy and 14.7% improvement in Macro F1 compared to the best baseline SINN on the 2012 Hurricane Sandy Dataset. For the 2020 U.S. election dataset, the accuracy improved 12.5% and Macro F1 improved 19.3%. Mistral NeMo demonstrates a slightly better overall performance, with an increase of 9.13%, 10.7% in accuracy and 14.15%, 17.6% in Macro F1, for the hurricane and election datasets, respectively. With the proposed framework, we have a 45% chance on average to correctly forecast the users' sentiments based solely on information available from news

Table 3: The microscopic performance for different models for datasets from two different states during the landfall of Hurricane Sandy. Bold denotes the best (highest) score and underline denotes the second-best score.

Dataset	2012 Hurricane Sandy								2020 U.S. Election			
Location	New Jersey				New York				/			
Time	T1		T2		T1		T2		T3		T4	
Metrics	Acc.	Ma. F1	Accu.	Ma. F1	Acc.	Ma. F1	Acc.	Ma. F1	Acc.	Ma. F1	Acc.	Ma. F1
Voter	0.199	0.133	0.169	0.128	0.194	0.137	0.156	0.125	0.347	0.189	0.387	0.198
DeGroot	0.183	0.143	0.310	0.217	0.187	0.150	0.288	0.196	0.238	0.185	0.217	0.184
SLANT+	0.213	0.168	0.227	0.187	0.190	0.131	0.202	0.137	0.332	0.169	0.345	0.170
NN	0.285	0.146	0.302	0.155	0.253	0.187	0.239	0.125	0.426	0.200	0.491	0.186
SINN	0.353	0.179	0.364	0.167	0.385	0.168	0.327	0.152	0.476	0.193	0.485	0.183
MPR$_G$	0.413	0.302	**0.453**	**0.331**	0.396	0.292	0.418	**0.329**	**0.615**	**0.374**	**0.596**	**0.397**
MPR$_M$	**0.445**	**0.312**	0.438	0.309	**0.482**	**0.310**	**0.429**	0.301	0.593	0.368	0.581	0.370

and social media.

Sentiment forecasting for a social media user or a group of users with more specific attributes and contextual information would significantly enhance accuracy. The prompt can be further refined with a higher granularity for specific users, allowing the event context to be tailored w.r.t. the user's circumstances. This would enable the MPR framework to generate social media comments that more closely reflect the user's actual situation.

4.3 Ablation Studies

To demonstrate the effectiveness of our framework, we conducted extensive ablation studies using different LLMs. As shown in Table 4, we performed three sets of experiments, each removing a key component from the full framework. The "MRP-RP" configuration directly forecasts the sentiment score during role-playing, removing the module to predict the next social media comment. "MRP-FE" removes the feature extraction module and directly predicts the social media comment through role-play. "MRP-OB" removes the objective fine-tuned "psychology" LLM, where role-playing is conducted solely based on the extracted features of the users.

Our proposed framework consistently outperforms the variants without important modules. The MRP-RP variant, comparable to random guessing, reflects the challenges humans face when predicting sentiment scores without proper context simulation. This result highlights the importance of replicating human action and thought processes in LLMs for effective human-oriented studies. Directly predicting sentences without feature extraction (MRP-FE) introduces high stochasticity. The MRP-OB variant, while slightly less effective than the full framework, underscores the value of the fine-tuned behavioral psychologist LLM, which was refined with just 25,000 Q&A instances. For applications focusing solely on sentiment polarity, our proposed framework can achieve an accuracy rate as high as 63.9%, demonstrating its capability to comprehend, reason, and forecast sentiment for social media users.

4.4 Discussions

We study the difference between the performance of both datasets and the error cases. First, in our experiments, the subjective role-playing agent only has access to limited information compared to real-life users who have diverse information sources like friends, families, local news, etc. Future comprehensive studies might require access to diverse sources to better simulate the user's information gain. In the case of Hurricane Sandy, more than 15% of people posted about their circumstances enduring Sandy without relying on information on social media, which makes their emotions hard to predict. On the other hand, the topic of the U.S. election dataset relies more extensively on news and social media output, which partially contributes to the better performance. Moreover,

The study of sentiment forecasting has a variety of potential applications. For large-scale natural disaster events, it could be used to detect and predict the areas with the worst mental conditions. In large social events, detecting and predicting extreme sentiment and emotion could help prevent potential chaos. In finance, our framework could be adapted to analyze finance-related social media comments to infer underlying market trends and correlations in a timely manner. Moreover, the ability to interpret and infer future sentiment is also crucial in the development of artificial intelligence

Table 4: The ablation study was performed by removing components across three models. Mistral represents an experiment performed with Mistral NeMo. Bold denotes the best (highest) results.

Dataset			2012 Hurricane Sandy							
Grainularity			Sentiment				Sentiment Polarity			
Time			T1		T2		T1		T2	
Metrics			Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
Gemma 2	NJ	MPR-RP	0.212	0.186	0.206	0.197	0.461	0.357	0.385	0.346
		MPR-FE	0.343	0.266	0.380	0.285	0.504	0.385	0.513	0.415
		MPR-OB	0.408	0.294	0.449	0.323	0.508	0.422	0.582	0.476
		MPR	**0.413**	**0.342**	**0.453**	**0.331**	**0.523**	**0.436**	**0.585**	**0.477**
	NY	MPR-RP	0.181	0.166	0.173	0.164	0.408	0.340	0.475	0.371
		MPR-FE	0.280	0.234	0.392	0.295	0.495	0.371	0.526	0.421
		MPR-OB	**0.393**	0.286	**0.418**	0.307	0.491	0.405	0.547	0.447
		MPR	0.380	**0.292**	0.412	**0.329**	**0.492**	**0.413**	**0.557**	**0.453**
Mistral	NJ	MPR-RP	0.261	0.211	0.225	0.209	0.394	0.342	0.381	0.374
		MPR-FE	0.420	0.268	0.415	0.293	0.532	0.371	0.533	0.425
		MPR-OB	**0.447**	0.299	0.427	0.303	**0.567**	0.434	0.570	0.448
		MPR	0.445	**0.312**	**0.438**	**0.309**	0.563	**0.440**	**0.576**	**0.450**
	NY	MPR-RP	0.286	0.241	0.289	0.234	0.417	0.375	0.442	0.374
		MPR-FE	0.475	0.289	0.434	0.305	0.550	0.425	0.615	0.426
		MPR-OB	0.480	0.291	**0.439**	**0.305**	0.557	0.418	0.622	0.434
		MPR	**0.482**	**0.310**	0.429	0.301	**0.569**	**0.426**	**0.639**	**0.452**
Dataset			2020 U.S. Election							
Grainularity			Sentiment				Sentiment Polarity			
Time			T3		T4		T3		T4	
Metrics			Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
Gemma 2		MPR-RP	0.415	0.196	0.354	0.104	0.477	0.323	0.391	0.173
		MPR-FE	0.516	0.213	0.407	0.190	0.551	0.356	0.434	0.320
		MPR-OB	0.594	**0.376**	0.588	0.378	0.685	**0.521**	0.662	0.509
		MPR	**0.615**	0.374	**0.596**	**0.397**	**0.692**	0.513	**0.669**	**0.533**
Mistral		MPR-RP	0.457	0.221	0.431	0.202	0.567	0.370	0.588	0.352
		MPR-FE	0.500	0.203	0.404	0.193	0.578	0.325	0.455	0.296
		MPR-OB	0.571	0.362	0.578	**0.377**	0.657	0.486	0.651	**0.520**
		MPR	**0.593**	**0.369**	**0.581**	0.370	**0.668**	**0.495**	**0.654**	0.508

with the ability to understand of theory of mind. An artificial agent with such an ability would also help various labor-intensive customer service scenarios.

5 Conclusion

In this paper, we target the problem of sentiment forecasting in social media to predict a user's future sentiment towards a given event. We proposed the context-aware multi-perspective role-playing framework to integrate the social media information up to time t to predict the sentiment at time $t' \geq t$. Experiments show that our proposed framework outperforms the state-of-the-art methods on both macroscopic and microscopic levels. The implementation is currently available at https://github.com/ManFanhang/Context-Aware-Sentiment-Forecasting-via-LLM-based-Multi-Perspective-Role-Playing-Agents.

Limitations

Although our proposed multi-perspective role-playing framework achieved state-of-the-art performance for the sentiment forecasting task, there remain several limitations:

1. **Model Constraints**: Popular models (e.g., GPT series) are explicitly trained to avoid inappropriate/negative content, limiting their use in the task of sentiment forecasting. Our framework uses less constrained LLMs, but performance depends on model choice. A sample performance of the sanctioned model Llama 3.1 can be found in the Appendix.

2. **Modality**: The current implementation only processes textual data, missing multimodal social media information (images, videos). Future work will explore efficient multimodal LLMs (Peng et al., 2025; Li et al., 2025).

3. **Scope**: This work focuses on sentiment, which primarily measures the valence of positive and negative. Emotion encompasses specific states (fear, anger, joy) (Hu et al., 2024). Forecasting specific emotions is a meaningful future direction.

Acknowledgements

This research project was supported by the National Key Research and Development Program of China 2024YFC3307603, Natural Science Foundation of China under Grant 62371269, U23B2030, and Tsinghua Shenzhen International Graduate School Cross-disciplinary Research and Innovation Fund Research Plan (JC20220011).

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 8–17.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Michael Caballero. 2021. Predicting the 2020 us presidential election with twitter. *arXiv preprint arXiv:2107.09640*.

Claudio Castellano, Santo Fortunato, and Vittorio Loreto. 2009. Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591–646.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Ziyang Chen, and Jia Li. 2022. Large language models meet harry potter: A bilingual dataset for aligning dialogue agents with characters.

Can Cemal Cingi, Nuray Bayar Muluk, and Cemal Cingi. 2023. Tone of voice, word choice, dress, behaviour and timing. In *Improving Online Presentations: A Guide for Healthcare Professionals*, pages 133–148. Springer.

Salvador Contreras Hernández, María Patricia Tzili Cruz, José Martín Espínola Sánchez, and Angélica Pérez Tzili. 2023. Deep learning model for covid-19 sentiment analysis on twitter. *New Generation Computing*, 41(2):189–212.

Abir De, Isabel Valera, Niloy Ganguly, Sourangshu Bhattacharya, and Manuel Gomez Rodriguez. 2016. Learning and forecasting opinion dynamics in social networks. *Advances in neural information processing systems*, 29.

Morris H. Degroot. 1974. Reaching a consensus. *Journal of the American Statistical Association*, 69:118–121.

Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2023. Llms to the moon? reddit market sentiment analysis with large language models. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1014–1019.

Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

Zahid Halim, Mehwish Waqar, and Madiha Tahir. 2020. A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email. *Knowledge-based systems*, 208:106443.

Guiyang Hou, Yongliang Shen, and Weiming Lu. 2024. Progressive tuning: Towards generic sentiment abilities for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14392–14402.

Anthony Hu and Seth Flaxman. 2018. Multimodal sentiment analysis to explore the structure of emotions. In *proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, pages 350–358.

Bo Hu, Meng Zhang, Chenfei Xie, Yuanhe Tian, Yan Song, and Zhendong Mao. 2024. Resemo: A benchmark chinese dataset for studying responsive emotion from social media content. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16375–16387.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Vasileios Iosifidis and Eirini Ntoutsi. 2017. Large scale sentiment learning with limited labels. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1823–1832.

Sk Mainul Islam and Sourangshu Bhattacharya. 2022. Ar-bert: Aspect-relation enhanced aspect-level sentiment classification with multi-modal explanations. In *Proceedings of the ACM Web Conference 2022*, pages 987–998.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yury Kryvasheyeu, Haohui Chen, Esteban Moro, Pascal Van Hentenryck, and Manuel Cebrian. 2015. Performance of social network sensors during hurricane sandy. *PLoS one*, 10(2):e0117288.

Bhushan Kulkarni, Sumit Kumar Agarwal, Abir De, Sourangshu Bhattacharya, and Niloy Ganguly. 2017. Slant+: A nonlinear model for opinion dynamics in social networks. *2017 IEEE International Conference on Data Mining (ICDM)*, pages 931–936.

Peter Kuppens and Philippe Verduyn. 2017. Emotion dynamics. *Current Opinion in Psychology*, 17:22–26.

Shilpa Lakhanpal, Ajay Gupta, and Rajeev Agrawal. 2023. Leveraging explainable ai to analyze researchers' aspect-based sentiment about chatgpt. In *International Conference on Intelligent Human Computer Interaction*, pages 281–290. Springer.

Sharon Levy, Robert E Kraut, Jane A Yu, Kristen M Altenburger, and Yi-Chia Wang. 2022. Understanding conflicts in online conversations. In *Proceedings of the ACM Web Conference 2022*, pages 2592–2602.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *ArXiv*, abs/2308.09597.

Jiahao Li, Huandong Wang, and Xinlei Chen. 2024a. Physics-informed neural ode for post-disaster mobility recovery. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1587–1598.

Kaiyuan Li, Xiaoyue Chen, Chen Gao, Yong Li, and Xinlei Chen. 2025. Balanced token pruning: Accelerating vision language models beyond local optimization. *arXiv preprint arXiv:2505.22038*.

Zuxin Li, Fanhang Man, Xuecheng Chen, Susu Xu, Fan Dang, Xiao-Ping Zhang, and Xinlei Chen. 2024b. Quest: Quality-informed multi-agent dispatching system for optimal mobile crowdsensing. In *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*, pages 1811–1820. IEEE.

Zuxin Li, Fanhang Man, Xuecheng Chen, Baining Zhao, Chenye Wu, and Xinlei Chen. 2022. Tract: Towards large-scale crowdsensing with high-efficiency swarm path planning. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, pages 409–414.

Yuanyuan Liu and Youlong Yang. 2022. A probabilistic linguistic opinion dynamics method based on the degroot model for emergency decision-making in response to covid-19. *Computers & Industrial Engineering*, 173:108677 – 108677.

Yuxuan Liu, Haoyang Wang, Fanhang Man, Jingao Xu, Fan Dang, Yunhao Liu, Xiao-Ping Zhang, and Xinlei Chen. 2024. Mobiair: Unleashing sensor mobility for city-scale and fine-grained air-quality monitoring with airbert. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*, pages 223–236.

Manchun, Hui. 2020. US Election 2020 Tweets. https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets.

Corrado Monti, Gianmarco De Francisci Morales, and Francesco Bonchi. 2020. Learning opinion dynamics from social traces. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 764–773.

Maryam Mousavi, Hasan Davulcu, Mohsen Ahmadi, Robert Axelrod, Richard Davis, and Scott Atran. 2022. Effective messaging on social media: What makes online content go viral? In *Proceedings of the ACM Web Conference 2022*, pages 2957–2966.

Roni Muslim, Rinto Anugraha Nqz, and Muhammad Ardhi Khalif. 2024. Mass media and its impact on opinion dynamics of the nonlinear q-voter model. *Physica A: Statistical Mechanics and its Applications*, 633:129358.

Maya Okawa and Tomoharu Iwata. 2022. Predicting opinion dynamics via sociologically-informed neural networks. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Ruiying Peng, Kaiyuan Li, Weichen Zhang, Chen Gao, Xinlei Chen, and Yong Li. 2025. Understanding and evaluating hallucinations in 3d visual language models. *arXiv preprint arXiv:2502.15888*.

Sergey Pletenev. 2024. Somethingawful at pan 2024 textdetox: Uncensored llama 3 helps to censor better.

Jill Walker Rettberg et al. 2017. Self-representation in social media. *SAGE handbook of social media*, pages 429–443.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Arghya Sahoo, Ritaban Chanda, Nabanita Das, and Bikash Sadhukhan. 2023. Comparative analysis of bert models for sentiment analysis on twitter data. In *2023 9th International Conference on Smart Computing and Communications (ICSCC)*, pages 658–663. IEEE.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *ArXiv*, abs/2310.10158.

Kirill Solovev and Nicolas Pröllochs. 2022. Moral emotions shape the virality of covid-19 misinformation on social media. In *Proceedings of the ACM web conference 2022*, pages 3706–3717.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Sijing Tu and Stefan Neumann. 2022. A viral marketing-based model for opinion dynamics in online social networks. In *Proceedings of the ACM Web Conference 2022*, pages 1570–1578.

Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. 2022. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83:19–52.

Zekun Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. 2023a. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *ArXiv*, abs/2310.00746.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023b. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.

Zhibin Wu, Qinyue Zhou, Yucheng Dong, Jiuping Xu, Abdulrahman H. Altalhi, and Francisco Herrera. 2023. Mixed opinion dynamics based on degroot model and hegselmann–krause model in social networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53:296–308.

Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. 2025. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*.

Wei Zhang, Meng Wang, and Yan-chun Zhu. 2020. Does government information release really matter in regulating contagion-evolution of negative emotion during public emergencies? from the perspective of cognitive big data analytics. *International Journal of Information Management*, 50:498–514.

Weichen Zhang, Chen Gao, Shiquan Yu, Ruiying Peng, Baining Zhao, Qian Zhang, Jinqiang Cui, Xinlei Chen, and Yong Li. 2025. Citynavagent: Aerial vision-and-language navigation with hierarchical semantic planning and global memory. *arXiv preprint arXiv:2505.05622*.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476.

Baining Zhao, Ziyou Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. 2025. Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. *arXiv preprint arXiv:2504.12680*.

Wanjun Zhong, Duyu Tang, Jiahai Wang, Jian Yin, and Nan Duan. 2021. Useradapter: Few-shot user learning in sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1484–1488.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *ArXiv*, abs/2311.16832.

A Appendix

A.1 Instruction Demonstration

Our proposed instruction and sample results are demonstrated in Figure 3. We took a social media user in New Jersey as an example. The goal is to predict the sentiment of this user immediately after the landfall. Since the LLMs are trained with a massive collection of corpora data, they contain retrospective knowledge of Hurricane Sandy. To verify the ability to predict the sentiment of a large crowd, the instruction should be designed to omit the effect of retrospective knowledge by constructing a hypothetically identical environment as if it were in a parallel universe. With the instruction, we generated only one social media content. Pre-experiment shows that multiple generations with Gemma 2 9B and Mistral NeMo 12B under the same prompt result in a 4.2% difference in the sentiment of the generated comment, which is not statistically significant.

A.2 Detailed Procedure of Fine-Tuning

The fine-tuning process was carried out using the Llama 3 8B Instruct model. The embedding layer includes `embed_tokens` with dimensions

Feature Extraction

A social media user has sent the following tweets:*{tweets}* Describe this user's overall tone of voice on the social media with three words. Only output in the exact format: 'xxx, xxx, xxx'

Observant, dismissive, and anxious

Suppose it is currently *{time}*, Hurricane Oscar is about to make landfall at {location} as a category 1 hurricane. *{context_description}*, A Twitter user who is currently at {address} had sent the following tweets:'*{tweets}*' before the landfall. The above tweets has a tone of voice of *{tone_of_voice}*. Based on the above information, please use three words to describe this user's overall attitude towards Hurricane Oscar. Only output these three words in the exact format: 'xxx, xxx, xxx.'

Resigned, humorous, and contemptuous

Objective Role-play

You are a behavioral psychologist with an expertise in social media behavioral analysis.

Hurricane Oscar just made landfall near *{location}*. *{new_context}*. A resident in *{state}* who is currently at *{address}* with a tone of voice of *{tone_of_voice}* on social media is *{attitude}* about this hurricane. This user compsed a new comment: *{generated_tweet}*. With your expertise, please analyze if the language use and the enunciated attitude of this new comment conforms with that of the previous ones. The previous social media comments from this user are provided as follows: *{tweets}*

while the observation part aligns with the previous comments, the dismissiveness and humorous might not be consistent with the overall pattern of behavior. However, the reaction to the event still reflects a level of preparedness and acknowledgment, which could be seen as maintaining a balance between being aware and not getting overwhelmed.

Subjective Role-play

You are a resident in *{state}* who is currently at *{address}*. Your attitude towards Hurricane Oscar is *{attitude}* with the past tone of voice of *{tone_of_voice}* on social media.

Suppose it is currently *{time}*. *{context_description}*. *{tweets}* are your previous tweets sent before the landfall of Hurricane Oscar. Now, Hurricane Oscar just made landfall at *{location}* as a category 1 hurricane. *{new_context}* You see your followees' tweets *{followee_tweets}* on Twitter. Based on the above information, you would like to send an immediate post-landfall comment. Only output the tweet.

Oscar just made landfall here in Central Jersey and it's wild out! Heavy rain, crazy wind, trees down everywhere. Stay safe everyone! #HurricaneOscar #NewJersey

Regeneration

You are a resident in *{state}* who is currently at *{address}*. Your attitude towards Hurricane Oscar is *{attitude}* with the past tone of voice of *{tone_of_voice}* on social media.

Suppose it is currently *{time}*. *{context_description}*. *{tweets}* are your previous tweets sent before the landfall of Hurricane Oscar. Now, Hurricane Oscar just made landfall at *{location}* as a category 1 hurricane. *{new_context}* on Twitter. Based on the above information, you've composed a new tweet *{generated_tweet}*. However, it is not consistent with your previous tweets *{tweets}* with the following reasons: *{reviews}*. Based on the above reflection. Please compose a tweet that is consistent with your tone of vice and attitude. Only output your tweet. Only output the tweet.

Ugh, power flickering here in Hamilton. Guess this whole "Category 1" thing is just wind for my neighbor's trees? #HurricaneOscar #NJlife

Figure 3: The demonstration of our constructed instructions for feature extraction, subjective role-play, and objective role-play.

(128256, 4096). The decoder comprises 32 `LlamaDecoderLayer` instances, each with self-attention (`LlamaSdpaAttention`), including `q_proj`, `k_proj`, `v_proj`, `o_proj`, and `rotary_emb`, as well as an MLP (`LlamaMLP`) with `gate_proj`, `up_proj`, `down_proj`, and activation function SiLU. Additionally, it includes `input_layernorm` and `post_attention_layernorm` layers, and a final normalization layer (`LlamaRMSNorm`). The output layer is a linear transformation (`lm_head`) from 4096 to 128256 dimensions. The model uses `torch.bfloat16` data type for processing.

The LoRA configuration was crucial for our fine-tuning approach. The configuration parameters include `LoraConfig` with `peft_type` set to LORA, `task_type` as CAUSAL_LM, and `r` value of 8. The target modules affected by LoRA are `down_proj`, `v_proj`, `up_proj`, `q_proj`, `k_proj`, `o_proj`, and `gate_proj`, with `lora_alpha` set to 32 and `lora_dropout` at 0.1. The model was configured with 20,971,520 trainable parameters out of 8,051,232,768 parameters, making the trainable percentage approximately 0.26%.

The training loss over the steps is visualized in 5, which illustrates the progressive decrease in loss, indicating the model's improvement over the training period. This appendix provides a detailed description of the Llama 3 8B Instruct model architecture, configuration, and training process utilizing the LoRA fine-tuning method. The visualization of the training loss demonstrates the model's convergence and the effectiveness of the fine-tuning approach. The sample Q&A pairs for fine-tuning are shown in Figure 4.

A.3 Event Context with Timeline

Hurricane Sandy is one of the most destructive and widely recognized disasters in U.S. history (Kryvasheyeu et al., 2015). The corresponding dataset comprises a comprehensive collection of Twitter messages from Oct. 15 to Nov. 12, 2012, spanning the period a week before the hurricane's formation and ten days after its dissipation. The metadata includes followee and friend counts, retweet statuses, follower-followee relationships, locations (self-reported or automatically detected), and timestamps. It comprises a total of 52.55 million messages from 13.75 million unique users, offering a detailed view of public sentiment and communication patterns during this significant natural disaster. Hurricane Sandy was formed southwest of Kingston, Jamaica on Oct. 22, 2012. It made landfall in Jamaica as a C1 hurricane on Oct. 24 and in Cuba as a C3 hurricane 10 hours later. On Oct. 29, Hurricane Sandy hit Brigantine, New Jersey as a C1 hurricane. The storm surge was as high as 3.85m, with prevalent levels between 0.8 and 2.6m along the coast of New Jersey and New York. New Jersey bore the brunt of the storm. New York, particularly Long Island, was heavily affected due to its terrain, but was rather farther from the storm's center. We mainly chose densely populated areas from New York and strongly affected areas from New Jersey to conduct the experiments. By Nov. 5, Hurricane Sandy was no longer active but had transitioned into a post-tropical cyclone. The government strived to perform post-disaster relief. However, many areas are still without power.

U.S. presidential election was one of the most influential events. The second dataset corresponds to the 2020 U.S. presidential election, where both candidates directly used Twitter to help express themselves and political opinions (Manchun, Hui, 2020; Caballero, 2021). It contains 1.7 million tweets spanning from Oct. 15 to Nov. 8, 2020, the period of the climax of the presidential race. The metadata contains users' descriptions of themselves, followee and friend counts, locations, and timestamps. On October 22, 2020, the final presidential debate between Donald Trump and Joe Biden took place in Nashville, Tennessee, focusing on various policy issues. On Nov. 3, 2020, Election Day saw a historic voter turnout, with both in-person and absentee voting contributing to the largest voter participation in over a century. The votes were counted from Nov. 4 to Nov. 7. However, the key states (swing states) experienced delays due to absentee votes. On Nov.7, Major news outlets projected Joe Biden as the winner of the election after Pennsylvania's results gave him enough electoral votes. However, President Trump contested the results. Claim of vote fraud.

A.4 Data Preprocessing

Our reasoning problem focuses on generating the next OSN comment for a user. However, corpora related to Hurricane Sandy were used to train the LLMs. Referring to Hurricane Sandy could lead to the retrieval of posterior knowledge, compromising the integrity of a prospective study. To avoid this interference, we renamed the event from Hurricane Sandy to Hurricane Oscar, a candidate hurricane name for 2024, and adjusted the timeline to

Positive Fine-tune Sample

An earthquake has happened near Atlantic City, New Jersey. A user who lives in Atlantic City with a tone of voice of confident, energetic, and assertive on social media is prepared, confident, and resilient about this earthquake. He now send a new comment: "This is not going well, but we will get through it." With your professional knowledge in psychology and social psychology, please analyze if the language use of the new comment conforms with the previous comments regarding tone of voice. Moreover, determine if their reaction aligns with their attitude towards the event. Their previous social media comments are as follows:"The way people supported each other during the power outage was truly amazing.", "We faced challenges before and always came back stronger.", "Community spirit is our greatest strength."

The tone of voice of this tweet is consistent with the previous tweets as it maintains a confident, energetic, and assertive tone. Despite acknowledging the difficulties ('This is not going well'), the user's use of 'but we will get through it' conveys resilience, demonstrating consistency with their previous attitude of being prepared, confident, and resilient."

Negative Fine-tune Sample

A political rally has happened near Washington D.C. A user who lives in Washington D.C. with a tone of voice of passionate, enthusiastic, and persuasive on social media is enthusiastically supportive of this political rally. She now send a new comment: "This rally was a waste of time. I'm disappointed in our community's lack of action. "With your professional knowledge in psychology and social psychology, please analyze if the language use of the new comment conforms with the previous comments regarding tone of voice. Moreover, determine if their reaction aligns with their attitude towards the event. Their previous social media comments are as follows: "The rally last week was electrifying!", "Our community's passion for change and justice is what gives us power.", "Together, we will make a difference."

The tone of voice of this tweet does not conform with the previous tweets, which were passionate, enthusiastic, and persuasive. The new comment expresses disappointment and lack of action, which contrasts with the user's previous attitude of being enthusiastically supportive. This shows inconsistency with their previous enthusiastic and supportive stance towards the event.

Figure 4: The expert example of the tone of voice and attitude consistent analysis.

2024. This ensures that knowledge about Hurricane Sandy is vaguely correlated. For the Hurricane Sandy dataset, keywords and hashtags such as "Hurricane," "Sandy," "storm," "power," etc. For the US 2020 election dataset, keywords and hashtags included "presidential election," "Biden," "Trump," etc. Noted that the context of the event, $\mathcal{E}$, captures the broader context of the event, which includes background knowledge, news updates, and evolving information surrounding the event. The source spans from official news websites, TV, and official announcements. On the other hand, the Followees comments, $\mathcal{F}$, consist of the posts and social media outputs from the followees, which are filtered and analyzed within our experiments. If a user u follows an official news account, it would also be formalized as $\mathcal{F}^u$. $\mathcal{E}^u$ and $\mathcal{F}^u$ could have overlapping information.

Moreover, we mapped OSN comments to a unified sentiment metric. Sentiment analysis is a complex field, encompassing lexicon-based, machine learning-based, and, most recently, LLM-based methods. Although LLMs can perform sentiment analysis on OSN comments, they require extensive resources and can exhibit high stochasticity (Lakhanpal et al., 2023). The same instruction might produce different sentiment labels for the same sentence, introducing instability into the proposed methods. After careful consideration,

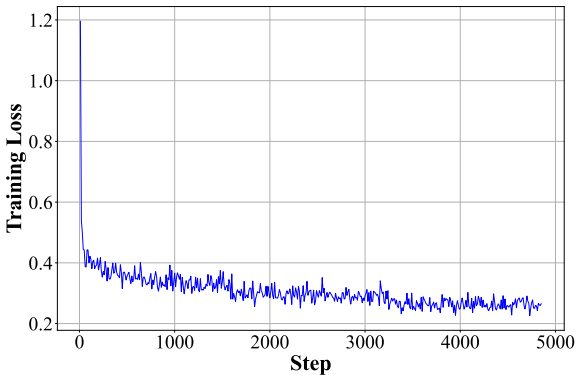

Figure 5: Fine-tune training loss for Llama 3 8b Instruct.

we selected the state-of-the-art supervised learning model bert-base-multilingual-uncased-sentiment.

A.4.1 Ground Truth

We first used this model to perform text-sentiment mapping on corpora of the original dataset, establishing the ground truth. Our problem is designed to predict the future expressed sentiments on OSN. We can either choose the next exact tweet or a collection of OSN users given a certain timeframe to make up the ground truth. The first scheme involves directly selecting the subsequent message from each user as the ground truth for scoring. Conversely, the second scheme computes the average value of all messages from each user during the

specified period, in our case, 24 hours.

To verify the consistency between two distinct design schemes for determining ground truth in our study, we employed the Kolmogorov-Smirnov (K-S) test. The K-S test results indicated a K-S Statistic of 0.0045, suggesting a minimal distributional difference between the two samples. Additionally, the P-value was found to be 0.9999, which implies that there is no significant difference between the distributions of the two schemes.

Based on these findings, we opted to utilize the second method for characterizing the ground truth. The rationale behind this choice lies in the stability of the second scheme. By averaging all messages from each user over the given period, this method mitigates potential anomalies and provides a more robust representation of user behavior.

A.4.2 User Selection

Then, we applied the model to analyze the sentiment of our OSN comments, ensuring uniformity across experiments. The model assigns a discrete sentiment score ranging from -2 to 2, where -2 indicates strongly negative sentiment and 2 indicates strongly positive sentiment. This five-category sentiment score allows for a fine-grained analysis of OSN user sentiments, capturing subtle variations in public sentiment.

To focus on users who freely express themselves on OSNs, we carefully selected user samples, omitting official accounts and news outlets. The experiment is conducted with 3000 users from New Jersey and 3000 users from New York, distributed in 15 Counties. Social media behavior is complex and subject to complex influence factors, especially for internet celebrities, the government, and the news. The study of crowd sentiment prediction, on the other hand, focuses on the intuitive and reactive response of the general netizens. The user should not be socially influential but rather influenced by OSN. We should exclude bots, news outlets, internet celebrities, public figures, and government voices. Therefore, we strategically select the users to represent the general netizens based on the following criteria.

1. The user's followees are accessible, indicating active interaction and engagement within the community.

2. The user's tweet history should contain mentions of "sandy" or "hurricane" to prevent irrelevance.

3. The threshold for total tweet counts is set to range from 10 to 1000, filtering out extremely inactive and active users.

4. The user's followers count and friends count were restricted to between 100 and 2000, to avoid unusually high or low social influences.

5. The presence of a geographic label was required to perform spatial segmentation for the user crowd. To perform geographical studies, we carefully select users from 15 different counties across New York and New Jersey.

6. The user must have comments before, during, and a week after the emergency to form a comprehensive view and the evolution of user sentiment.

There are a total of 124,876 users satisfying these criteria with a total of 5,038,920 tweets. For large enough users, we could assume the distribution of their behaviors. We group the users based on their geographical label. A region of the directly affected area was defined based on the trajectory and the diameter of the wind cycle of Hurricane Sandy. The users were segmented to be either directly affected or not affected. Moreover, to predict the evolution of user sentiment, we made predictions during and after the landfall.

A.5 Performance Analysis for Different LLMs

Due to freedom of speech, social media content may include swear words and politically incorrect expressions. Some popular LLMs, including the GPT series from OpenAI and the Llama series from Meta, are tuned to omit processing and generation of such content (Achiam et al., 2023). Instead of conducting the analysis, these LLMs would say that they can't proceed with such content. When we tried to use GPT and Llama to perform an analysis of social media comments, the tone of voice and attitude for aggressive and politically incorrect phrases could not be analyzed, and the social media comment generation failed due to censorship aimed at aligning with human values (Pletenev, 2024). When we try to perform feature extraction and subjective role-play to generate the post-landfall social media comments, 1547 out of 3000 social media users from New Jersey and 1633 out of 3000 social media users from New York are detected to have aggressive expressions. For the same 6,000 users, Llama avoided generating social media comments

Table 5: The extended ablation study performed with aligned model Llama 3,1. Bold denotes the best (highest) results

			Sentiment				Sentiment Polarity			
			T1		T2		T1		T2	
			Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
Llama 3.1	NJ	MPR-RP	0.150	0.150	0.127	0.106	0.185	0.150	0.254	0.174
		MPR-FE	0.192	0.115	0.308	0.159	0.243	0.165	0.422	0.247
		MPR-OB	**0.184**	**0.113**	**0.312**	0.161	0.240	0.168	0.424	0.151
		MPR	0.173	0.108	0.308	**0.163**	**0.255**	**0.173**	**0.425**	**0.247**
	NY	MPR-RP	0.137	0.143	0.197	0.148	0.138	0.113	0.267	0.204
		MPR-FE	**0.178**	0.098	0.250	0.141	0.223	0.148	0.337	0.216
		MPR-OB	0.172	0.101	**0.254**	**0.154**	0.226	0.160	**0.342**	0.217
		MPR	0.158	**0.103**	0.243	0.152	**0.227**	**0.163**	0.340	**0.218**

for 468 users from New Jersey and 854 New Jersey users. We treat these results as false for all categories, significantly decreasing the accuracy rate and the F1 score as shown in Table 5, the extended ablation study with Llama 3,1.

We used the bert-base-multilingual-uncased-sentiment model to perform sentiment analysis for the tweets detected to be politically incorrect or toxic. Over 82% are assigned the score -2 (strongly negative), and 94% are detected to be negative. It aligns with our assumption that the aggressive comments are mostly strongly negative. On the other hand, most politically incorrect tweets with positive sentiments use some swear words or politically incorrect terms as slang or exclamation.