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

Anonymous ACL submission

Abstract

The sentiment of users on social media reveals their underlying social trends, crises, and needs. Researchers have analyzed users' past messages to trace the evolution of sentiments and reconstruct sentiment dynamics. However, predicting the imminent sentiments of an ongoing event is rarely studied. In this paper, we address the problem of **sentiment forecasting** on social media to predict the user's future sentiment in response to the development of the event. We extract sentiment-related features to enhance the modeling skill and propose the multi-perspective role-playing (MPR) method to simulate the process of human response. Our preliminary results show significant improvement in sentiment forecasting on both microscopic and macroscopic levels. Code for reproduction is available at `https://anonymous.4open.science/r/Sentiment_Forecasting`.

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 sentiments can assist disaster relief organizations in adjusting their key objectives to focus on areas with heightened collective distress (Zhang et al., 2020). Nevertheless, a more useful sentiment research would be **sentiment forecasting** which foresees the evolution of public sentiment w.r.t. the development of an ongoing event. Sentiment forecasting would allow early detection of trends, proactive decision-making, and resource allocation.

Retrospective sentiment analysis methods typically assign a discrete or linear sentiment score to sentences, paragraphs, or entire documents (Mousavi et al., 2022). Researchers harnessed 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 during COVID-19. However, these methods only 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, such development has complex semantic information which is too intricate to be formulated as an input for the above neural network or probabilistic linguistic methods.

The advent of large language models (LLMs) sheds light on comprehensive and prospective sentiment reasoning in real-world scenarios (Chang et al., 2023). LLMs understand complex external event contexts with embedded common-sense knowledge. Zhang et al. (2023) fined-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. Such advantages provide an opportunity to pivot from traditional retrospective studies of sentiment analysis to prospective studies of sentiment forecasting.

However, there are still gaps in accurately forecasting the sentiment of different users on social media. To begin with, it requires comprehensive user features to model the user behavior (Iosifidis and Ntoutsi, 2017). These labels are difficult to obtain on social media due to anonymity and privacy

regulations. Moreover, even with sufficient labels, the evolution of human sentiments is too intricate and subtle to model (Kuppens and Verduyn, 2017; Wang et al., 2022). At different times, people in various situations would have diverse sentiment responses to the same events (Hu et al., 2024). Yet it is challenging to model such subtle reactions.

In this paper, we focus on the problem of **Sentiment Forecasting** to predict people's future sentiments on social media under real-world events. To comprehensively incorporate the context, we leverage LLMs to understand the complex semantic information. To enrich the features for user-specific sentiment modeling, we extract the features from user-generated content, specifically the tone of voice and attitude toward the event. To address the complexity and variety of sentiment evolution, we develop a multiperspective role-playing method to forecast the users' social media comments. Specifically, a general LLM would role-play the diverse users to express themselves on social media. A Fine-tuned "psychologist" LLM with expertise would analyze the response and feedback to the general LLM for reflections to ensure consistency. The proposed method addresses the complexity of user-specific sentiment forecasting, enabling more precise and nuanced predictions of both individual and collective sentiments.

Our contributions lie in the following aspects:

- We focus on **Sentiment Forecasting** on social media, aiming to predict the future sentiment response of social media users towards ongoing real-world events. We formulate it as a reasoning problem to incorporate external context.

- To enrich user-specific features, we created an LLM-based feature extraction method. The retrieved features would enhance the ability to simulate the sentiment response for different social media users.

- We present a multiperspective role-playing method to predict future sentiment responses. A general LLM role-plays common social media users with extracted features to generate responses. A fine-tuned LLM role-plays objective psychologist to ensure behavioral consistency.

2 Related Work

2.1 The Evolution of Sentiments

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) fit 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.

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. Therefore, though inspiring, such methods could only be used to predict some specific fabricated character's reaction to some events with limited generalizability in real-life applications.

3 Methodology

In this section, we first provide a general overview of the context-aware sentiment forecasting framework. We then present the feature extraction method, the subjective role-play method, and the objective reflective role-play method with fine-tuned large language models.

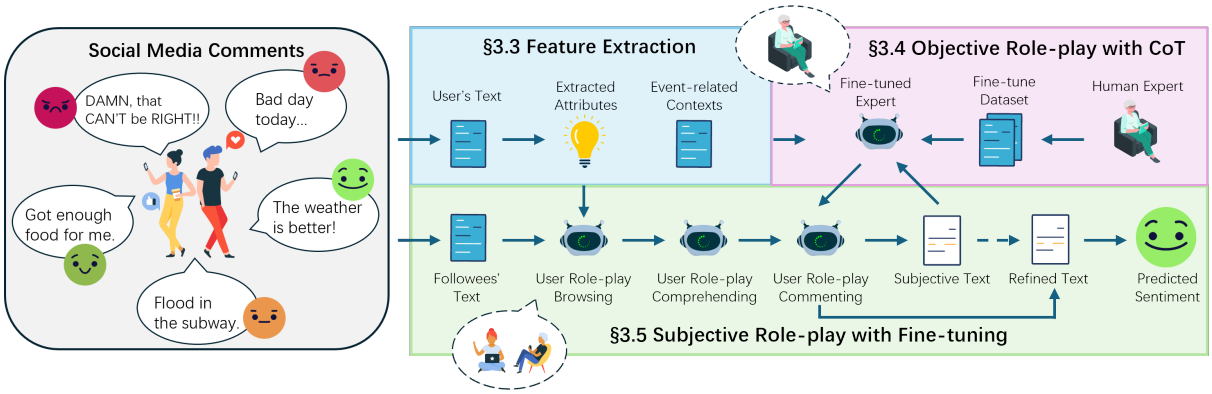

Figure 1: The overview of the proposed sentiment forecasting framework.

3.1 Problem Definition

A retrospective sentiment analysis task could be expressed as follows:

$$\sigma_t = F(\cdot_t), \tag{1}$$

where the sentiment label σ at time t is extracted from input $\cdot_t$ which is normally the user-generated content on social media at time t.

However, sentiment forecasting uses the information at time t to predict a user's sentiment in the near future $t+1$:

$$\sigma_{t+1} = F(\cdot_t), \tag{2}$$

where $t+1$ means the sentiment in the near future. It is not necessarily one time step ahead.

3.2 Overall Framework

To effectively integrate this semantic information, we develop an LLM-based framework with Chain-of-thought (CoT) reasoning. We formulate a reasoning problem, where the goal is to generate the user's future social media comments. The sentiment label is retrieved from this comment with state-of-the-art sentiment analysis.

As shown in Figure 1, the framework is composed of three components: feature extraction, subjective role-play with CoT, and objective role-play with fine-tuning. Feature extraction aims to identify user features derived from social media comments. This process is designed to execute NLP tasks that capture the tone of voice and infer the user's attitude towards a given event. With the extracted features, an LLM is employed to simulate the behavior of the user. It would act as a social media user to comprehend the specific contexts and skim through the followees' social media comments. Subsequently, it is prompted to generate a new comment. To ensure consistency in the user's tone of voice and attitude flow, a fine-tuned "psychology" LLM reviews this comment to filter potential behavioral inconsistency in the tone of voice. Further details are provided in the sections below.

3.3 Feature Extraction

The tone of voice reflects a user's unique expression on social media. Studies demonstrate that people tend to maintain a consistent tone of voice according to their social image on social media (Cingi et al., 2023; Rettberg et al., 2017). Given the LLMs's abilities in NLP and text pattern recognition, we designed a feature extraction prompt to extract the tone of voice. We analyze the tone of voice τ^u given the social media comments of a specific user u:

$$\tau^u = \text{LLM}(\mathcal{C}^u, p_\tau), \tag{3}$$

where p_τ denotes the designed prompt for analyzing the user's tone of voice, and $\mathcal{C}^u$ represents the user's previous social media comments. The output is structured as three descriptive adjectives that encapsulate the tone of voice. An example is shown in A.1.

The attitude towards an event can also be inferred from social media comments. The user's attitude may shift moderately as the event evolves. Therefore, the event context $\mathcal{E}$ serves as an important influence factor in determining the user's attitude. With an impression of the event and an understanding of the tone of voice, the user's attitude is defined as:

$$\alpha^u = \text{LLM}(\tau^u, \mathcal{C}^u, \mathcal{E}, p_\alpha), \tag{4}$$

Symbol	Description of Notations
$\mathcal{C}$	User's social media comments
$\mathcal{U}$	User attributes
$\mathcal{F}$	Followees' social media output
$\mathcal{E}$	Context of Ongoing Event
p_n	Prompt for task n
τ	Tone of voice
α	Attitude towards the event
ψ	Predicted social media comment
θ	The analysis from expert
σ	Sentiment label

Table 1: Notations

where α^u represents the user-specific attitude. p_α represents the designed prompt to analyze the user's attitude toward the event.

3.4 Subjective Role-play with CoT

The subjective role-play strategy employs the CoT approach to prompt LLMs to role-play social media users based on the extracted features. The role-played user is instructed to maintain the tone of voice and attitude towards the event. Moreover, the user's previous social media comments $\mathcal{C}^u$ are provided as the few-shot learning sample. The subjective role LLM^s is defined as:

$$\text{LLM}_t^s = \text{RP}(\tau_t^u, \alpha_t^u, \mathcal{U}_t^u, \mathcal{C}_t^u, \mathcal{E}_t, p_r), \quad (5)$$

where LLM^s indicates the LLM designed to role-play a subjective social media user. $\mathcal{U}$ is the attainable social media user attribute, e.g. self-reported location. The subscript $_t$ indicates that the features at time t. The features and user-generated content until time t are harnessed to role-play the users.

With such a role-played social media user, we could simulate the scene of "browsing through social media". A sample of the followees' social media comments is extracted based on empirical social media influence factors, such as relevance to the topic, frequency of interaction, number of followers, etc. The reactive process is broken down into three stages: browsing, comprehending, and responding. It is formalized as follows:

$$\psi_{t+1}^u = \text{LLM}^s(\mathcal{F}_t^u, \mathcal{C}_t^u, \mathcal{E}_t, p_s), \quad (6)$$

where ψ_{t+1}^u is the generated future user comment regarding the ongoing event w.r.t. the context and user features.

3.5 Objective Role-play with Fine-tuning

To mitigate over-subjectiveness and prevent unusual stochastic behaviors, we fine-tune an LLM to criticize the produced comment generated by the subjective role-playing process. To ensure consistency, we enlisted several Ph.D. students in behavioral psychology to analyze the alignment between the generated social media comments and the few-shot learning examples with a consistent tone of voice and a reasonable and coherent attitude. We construct more than 25,000 sets of social network comment review Q&A pairs with GPT 4o with expert analysis as few-shot learning examples. These Q&A pairs are utilized for low-rank adaptation 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 Llama 3 8b Instruct model, functioning as an objective behavioral psychologist. Details of the fine-tuning process are illustrated in Appendix A.2. The composed prompts for the fine-tuned model are crafted to match the format of the questions in the Q&A pairs, ensuring optimal performance of the model.

We define the fine-tuned model to be LLM^f, role-playing a behavioral psychologist. The consistency analysis is defined as follows:

$$\theta_{t+1}^u = \text{LLM}^f(\tau_{t+1}^u, \alpha_t^u, \mathcal{C}_t^u, \psi_{t+1}^u, p_o), \quad (7)$$

where θ^u is the review from the fine-tuned expert, and p_o is the prompt for consistency analysis.

If the output is consistent, the generated comment from the subjective role is treated as the final output. If not, the analysis from the expert is then fed back to the subjective role to help regenerate a more consistent social media comment.

$$\psi_{t+1}^u = \text{LLM}^s(\theta_{t+1}^u, \mathcal{U}_t^u, \mathcal{C}_{t+1}^u, \psi_{t+1}^u, p_g), \quad (8)$$

where ψ_r^u is the regenerated social media comment for a role-played social media user whose previously generated comment is determined to be inconsistent. p_g indicates the prompt for regeneration.

The generated comment with a consistent tone of voice and coherent attitude is the result of the formulated reasoning problem. Then, the sentiment of the generated comments could be analyzed with state-of-the-art sentiment analysis methods. The notations are depicted in table 1.

Table 2: Macroscopic performance comparison between two states during Hurricane Sandy's landfall, evaluated using JSD. T1 represents the forecast made immediately after the hurricane's landfall, while T2 represents the forecast one week later. Sentiment evolution reflects the change from T1 to T2. Bold indicates the best (lowest) results, and underlining denotes the second-best results.

Metrics	Distribution of Sentiment				Distribution of Sentiment Polarity				Evolution of Sentiment	
Location	New Jersey		New York		New Jersey		New York		New Jersey	New York
Time	T1	T2	T1	T2	T1	T2	T1	T2		
Voter	0.2822	0.3234	0.2245	0.2581	0.2543	0.2423	0.2454	0.2341	0.6544	0.6425
DeGroot	0.2376	0.1869	0.2511	0.2243	0.2082	0.1744	0.1310	0.1213	0.6528	0.6067
SLANT+	0.3630	0.30355	0.2912	0.2874	0.2516	0.2681	0.2905	0.2734	0.6862	0.6401
NN	0.1904	0.2161	0.1733	0.1804	0.1517	0.1379	0.1270	0.1355	0.6066	0.6031
SINN	0.1673	0.1359	0.1504	0.1377	0.1201	0.1312	0.1022	0.1216	0.5450	0.5427
MPR_G	0.0243	**0.0105**	0.0456	0.0396	0.0211	**0.0064**	0.0290	0.0239	0.3315	0.3543
MPR_M	**0.0148**	0.0192	**0.0220**	**0.0313**	**0.0114**	0.0100	**0.0116**	**0.0204**	**0.3057**	**0.2781**

4 Experiment

In this section, we evaluate the effectiveness and robustness of the proposed method. We performed extensive experiments with different models and compared the results with several state-of-the-art baselines. The source code of our implementation is available[1].

4.1 Experiment Settings

4.1.1 Dataset

Hurricane Sandy, the largest hurricane of 2012 and one of the most destructive and widely recognized disasters in U.S. history (Kryvasheyeu et al., 2015), garnered significant attention on social media upon its formation on Oct. 22, 2012. This dataset comprises a comprehensive collection of Twitter messages, capturing digital traces on social media from Oct, 15 to Nov. 12, spanning the period before the hurricane's formation and extending beyond its landfall in the continental United States. In addition to the textual content, the dataset provides metadata, including user identifiers, followee 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. This dataset allows for in-depth analysis not only of the text itself but also of the temporal-spatial attributes and user interactions, 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[2], which achieves an accuracy as high as 87% on multiple datasets (Sahoo et al., 2023; Contreras Hernández et al., 2023). Details of data preprocessing are shown in Appendix A.3.

4.1.2 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.

- **Voter** (Muslim et al., 2024): The voter model is one of the classic models where user sentiment is adapted randomly from their followees.
- **DeGroot** (Degroot, 1974; Wu et al., 2023): The DeGroot model assumes users update their sentiment iteratively to the weighted average of the followee's opinion.
- **SLANT+** (Kulkarni et al., 2017): SLANT+ is a non-linear generative model applied recurrent neural network (RNN) with the point process model to learn the non-linear evolution of the user's sentiment.
- **NN** (De et al., 2016): A pure neural network (NN) method to learn the evolution of users' sentiments based on previous patterns.
- **SINN** (Okawa and Iwata, 2022): The Sociologically-informed Neural Network

[1] https://anonymous.4open.science/r/Sentiment_Forecasting/

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

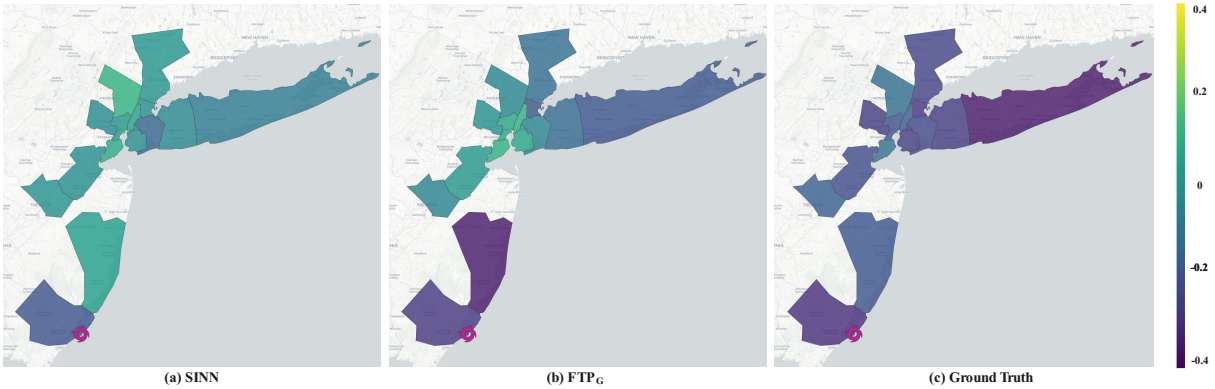

Figure 2: The collective sentiment level for selected counties from New York and New Jersey. The red circle indicates the location of the Landfall of Hurricane Sandy. Figures (a) and (b) are the reconstructed sentiment distribution maps for SINN and our proposed MPR$_G$, respectively. Figure (c) is the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

(SINN) method harnesses sociological models to guide the model to approach users' sentiment evolution.

4.1.3 Metrics

We measure the alignment of the predicted sentiment with the real-world data on both the individual level and the general crowd level. For individuals, we measure the performance of sentiment forecasting with accuracy and Macro F1. To get quantitative evaluations on the fidelity of the generated data, we adopt the Jensen-Shannon divergence (JSD) to measure the distribution similarity of the above pattern between the predicted sentiment for the users and the sentiment of the actual user texts. It is formalized as follows:

$$\mathrm{JSD}(\mathbf{p}||\mathbf{q}) = \frac{1}{2}\mathrm{KL}(\mathbf{p}||\frac{\mathbf{p}+\mathbf{q}}{2}) + \frac{1}{2}\mathrm{KL}(\mathbf{q}||\frac{\mathbf{p}+\mathbf{q}}{2}), \tag{9}$$

where H represents the entropy, $\mathbf{p}$ and $\mathbf{q}$ are two distributions and $\mathrm{KL}(\cdot||\cdot)$ 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 use the distribution of sentiment and the distribution of sentiment polarity $p = \mathrm{sgn}(s)$ 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. We use the evolution of sentiment from *immediately after landfall (T1)* to *a weeks after landfall (T2)* to measure the predicted sentiment shift.

4.2 Experiment Results

We present the macroscopic results to show the crowd-level performance followed by the user-specific performance.

4.2.1 Macroscopic Performance

Table 2 presents the performance of our models in two different states. Our multi-perspective role-play method, compatible with most LLMs, was tested using Gemma2 9b (Team et al., 2024) and Mistral NeMo 12b (Jiang et al., 2023), denoted as MPR$_G$ and MPR$_M$, respectively. MPR represents the generated comment following analysis by the fine-tuned psychologist LLM. For the Macroscopic JSD metric, our proposed method 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. The model-based methods tend to converge to a certain point 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. Our MPR method, however, considers the context information through reasoning. This is particularly effective when events

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) results and underline denotes the second-best results

	New Jersey				New York			
	T1		T2		T1		T2	
	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
Voter	0.199	0.133	0.169	0.128	0.194	0.137	0.156	0.125
DeGroot	0.183	0.143	0.310	0.217	0.187	0.150	0.288	0.196
SLANT+	0.213	0.168	0.227	0.187	0.190	0.131	0.202	0.137
NN	0.285	0.146	0.302	0.155	0.253	0.187	0.239	0.125
SINN	0.353	0.179	0.364	0.167	0.385	0.168	0.327	0.152
MPR_G	0.413	0.302	**0.453**	**0.331**	0.396	0.292	0.418	**0.329**
MPR_M	**0.445**	**0.312**	0.438	0.309	**0.482**	**0.310**	**0.429**	0.301

shift drastically, as was the case with Hurricane Sandy's landfall, which significantly altered the living environment for social media users.

Moreover, we analyzed users from different geographical regions to assess group-specific performance. As shown in 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.

4.2.2 Microscopic Performance

In addition to the crowd-level analysis, we also heavily tested the performance of our method at an individual level. As shown in Table 3, our proposed methods outperformed the baselines for both Gemma2 and Mistral NeMo. With Gemma2, our proposed method achieves an average of 6.23% improvement in accuracy and 14.7% improvement in Macro F1 compared to the best baseline SINN. Mistral NeMo demonstrates a slightly better overall performance, with an increase of 9.13% in accuracy and 14.15% in Macro F1. With the proposed method, we have a 45% chance on average to correctly forecast the users' sentiments based solely on information available from news 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 immediate environment. This would enable the MPR method to generate social media comments that more closely reflect the user's actual situation.

4.3 Ablation Studies

To demonstrate the effectiveness of our method, we conducted extensive ablation studies using different LLMs. As shown in Figure 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 and the role-play is conducted solely based on the extracted features of the users.

Our proposed method consistently outperforms the compared methods short of important modules. The MRP-RP method, 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. However, there was one instance where MRP-FE outperformed the other methods using Llama 3.1. This is due to the overall compromised performance of the Llama model which is addressed in the following subsection. The SRP method, 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 method can achieve an accuracy rate as high as 63.9%, demonstrating its capability to comprehend, reason, and forecast sentiment for social media users.

Table 4: The ablation study was performed by removing components across three models. 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
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 NeMo	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**
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**

4.4 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 for 468 users from New Jersey and 854 New Jersey users. For these results, we treat them as false for all categories, significantly decreasing the accuracy rate and the F1 score.

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% is assigned with 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.

5 Conclusion

In this paper, we propose the problem of sentiment forecasting in social media, addressing the effect of the event context on the evolution of user sentiments. It is formalized as a reasoning problem to predict the next comments for social media users. The proposed method harnessed LLMs to role-play the social media users. Experiments show that our proposed method outperforms the state-of-the-art methods on both macroscopic and microscopic levels.

Limitations

First, some popular LLMs, including GPT and LLama series are trained to avoid generating social media comments w.r.t. inappropriate or strongly negative tone of voice. It indicates that our methods are limited to LLMs that are not trained with a sense of political correctness. The performance also varies from the choice of LLMs. Moreover, the proposed method is limited to the text which only represents a fraction of the information on social media. We see significant potential when incorporating multimodality to perform sentiment forecasting in a more comprehensive way.

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*.

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.

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.

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.

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.

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*.

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.

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.

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.

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 Prompt Demonstration

Our proposed prompting methods are demonstrated in Figure 3. We took an OSN 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 prompt should be designed to omit the effect of retrospective knowledge by constructing a hypothetically identical environment as if it were in a parallel universe.

A.2 Detailed Procedure for Fine-Tuning LLaMA3 8B with LoRA

The fine-tuning process was carried out using the LLaMA3 8B Instruct model. The model architecture consists of several components: the embedding layer, decoder layer, normalization layer, and output layer. Specifically, the embedding layer includes `embed_tokens` with dimensions (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 LLaMA3 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 is shown as 4

A.3 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. Prompting the LLMs with references 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 2024. This ensures that knowledge about Hurricane Sandy is vaguely correlated.

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 prompt might produce different sentiment labels for the same sentence, introducing instability into the proposed methods. After careful consideration, we selected the state-of-the-art supervised learning model bert-base-multilingual-uncased-sentiment.

A.3.1 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 col-

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 prompts for feature extraction, subjective role-play, and objective role-play.

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.

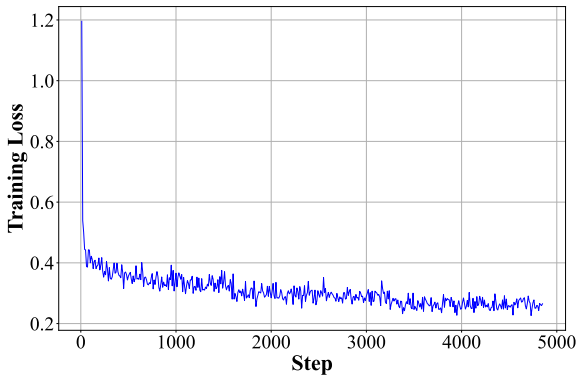

Figure 5: Fine tune training loss for Llama 3 8b Instruct.

lection 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.3.2 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 subjective to complex influence factors, especially for internet celebrities, government, and 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 was 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 that the distribution of their behaviors We group the users based on their geographical label. A region of 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.