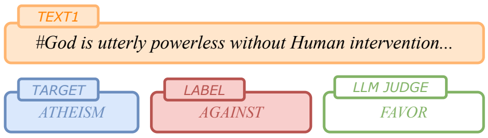
Understanding and Tackling Label Errors in Individual-Level Natural Language Understanding

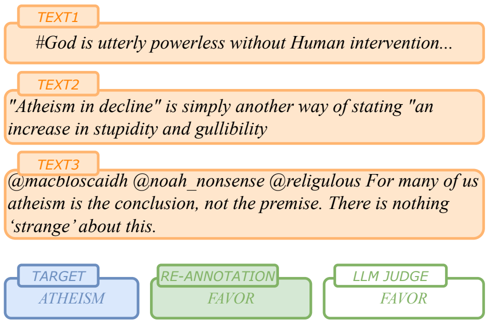
Anonymous ACL submission

Abstract

Natural language understanding (NLU) is a task that enables machines to understand human language. Some tasks are closely related to individual subjective perspectives, thus termed individual-level NLU. Previously, these tasks are often simplified to text-level NLU tasks, ignoring individual factors such as demographic information or worldview over a period of time. This can lead to a large number of label errors and also make inference difficult and unexplainable. To address the above limitations, we propose a new NLU annotation guideline based on individual-level factors. We find that the error rate in the dataset samples exceeded 20%, with the highest reaching 48.7%. We further use large language models to conduct experiments on the re-annotation datasets and find that the large language models perform well on these datasets after re-annotation. We also verify the effectiveness of individual factors through ablation studies. Our re-annotation dataset can be found at https://anonymous.4open.science/r/Individual-NLU-A0DE.



(a) The original dataset contains text, target and label. LLM judge on the text is different from the label.



(b) Expanded dataset, add other posts of the same individual (user) for the same target, and re-annotate the user's stance. The LLM Judge is consistent with the re-annotation label.

Figure 1: A typical example of potential label error in stance detection.

1 Introduction

Natural language understanding (NLU) is the task that uses semantic and syntactic analysis to enable computers to understand human-language inputs. Common natural language understanding tasks include fake news detection (Shu et al., 2017), sentiment analysis (Wankhade et al., 2022), stance detection (AlDayel and Magdy, 2021), toxicity/hate speech detection (Pavlopoulos et al., 2020), and sarcasm detection (Joshi et al., 2017). Existing NLU datasets are predominantly text-based, relying solely on short text information without accounting for social factors. While text-level NLU simplifies many tasks, its limitations begin to be recognized, such as poor inference performance (Hovy and Yang, 2021; Bhattacharya et al., 2025). So, some researchers have propose frameworks integrating social factors into NLU (Hovy and Yang, 2021). Additionally, various studies have incorporated different social factors such as user information and background knowledge into specific tasks to improve NLU accuracy (Yang and Eisenstein, 2017; Aldayel and Magdy, 2019).

However, current research has not explored in depth which NLU tasks will have huge deficiencies when using only textual information (without any other factors). In this paper, we define a type of NLU tasks as individual-level NLU tasks, where the labels reflect the identity or perspective of the individual (typically the web user who posts the text) rather than the content of the text itself. We argue that inference only with short texts is flawed in such tasks. Tasks that fall under individual-level NLU include sentiment analysis, sarcasm detection, and stance detection etc. A key characteristic of these tasks is that their labels are inherently tied to the publishers rather than the readers, or at least the publisher's identity and perspective play an important role in the label. An NLU task that does not incorporate an individual's perspective is not considered individual-level NLU. Such tasks are usually annotated based on social consensus or objective facts. For example, in tasks such as natural language inference, the labels usually represent a broadly accepted interpretation rather than an individual user's perspective (Bowman et al., 2015). In fake news detection (Shu et al., 2017) and authorship detection (Huang et al., 2024), the label remains unchanged regardless of whether one or multiple individuals share or endorse it.

This deficiency is reflected in the creation of the datasets. Current research often implicitly assumes that the labels in original datasets are accurate. However, individual-level NLU datasets are often created using text-level guidelines, and annotators' interpretations may differ from those of the original publishers. Such misalignment can lead to a significant number of labeling errors. Prior works have attempted to mitigate this issue by leveraging the individual factors in the datasets. For instance, datasets like Amazon reviews (Zhang et al., 2015) and IMDB (Maas et al., 2011) assign labels directly based on user scores, reducing the likelihood of annotation inconsistencies. However, many individual-level NLU datasets, such as the Twitter stance detection dataset (Mohammad et al., 2016) and the Twitter sentiment analysis dataset (Rosenthal et al., 2019), depend mostly on manual annotation, since social media posts do not come with explicit "scores" and must be annotated manually or inferred through hashtags instead. A recent study demonstrates that large language models (LLMs) perform well when human annotators do but fail in cases where human annotators struggle to reach consensus (Li and Conrad, 2024). This suggests that inconsistencies among annotators stem from the inherent ambiguity of the text rather than from annotator negligence.

From a sociolinguistic perspective, the attitude of an individual should be tied to the original publisher's intent at the time of posting (Kockelman, 2004), rather than being subject to the variability of annotator interpretations. Annotation inconsistencies often arise due to insufficient information and poor data quality. This phenomenon is referred to as systematic label errors (Cabrera et al., 2014). Prior work has reported non-trivial label error even under the original text-level stance detection; the error rate for the atheism category in the SemEval-2016 dataset was as high as 22.7% (Garg and Caragea, 2024). Beyond such mislabels, we argue that a single post is often informationally insufficient for user-level inference, motivating the use of multi-post evidence. We later validate this necessity via a simple ablation on the amount of user context.

To address this research gap, we propose guidelines for three NLU subtasks: stance detection, topic-based sentiment analysis, and hate speech detection. These guidelines aim to identify and mitigate systematic labeling errors that may exist in text-level NLU datasets. Specifically, we incorporate additional posts from the same user within a similar timeframe to assess the accuracy of dataset labels. We mitigate these errors by incorporating additional posts from the same user within a bounded time window, using cross-post consistency as evidence for the individual-level label. Our analysis reveals a substantial number of labeling errors. To further evaluate these errors, we employ three mainstream large language models (LLMs) to evaluate the datasets. Our findings indicate that LLMs achieve exceptionally high accuracy on the re-annotated datasets using only simple prompts, demonstrating the necessity of introducing individual-level NLU and individual factors.

We summarize our contributions as follows:

- We propose a novel guideline to reduce labeling errors in individual-level NLU.
- We show that widely used individual-level benchmarks can substantially underestimate model performance due to systematic label errors, and that evaluation on our corrected labels yields markedly higher and more consistent scores across models.
- We show that using multi-post user context is crucial for user-level inference, supported by controlled ablations on the amount of available context.

2 Related Work

In this section, we will introduce existing the definition for label errors. Then we will introduce pre-trained and large language models, and explain their potential in detecting label errors in the individual-level NLU task.

2.1 Label Errors

The inconsistency between the labels and groundtruths in the training dataset is often called "noisy labels" (Song et al., 2022). If the labels are inconsistent with the groundtruths in the test dataset, it is called label errors. Label errors are common in test datasets and may affect the evaluation of the model. There is an average of 3.3% label error in ten commonly used datasets (Northcutt et al., 2021).

Large number of label errors (over 20%) are probably not due to the negligence of the annotators but the defects in the annotation guidelines themselves. Label errors at such scale also imply pervasive noisy labels in the training data, biasing both training and test distributions away from the true data-generating process and thereby undermining classical label-error detection methods that assume a relatively clean test set or stable noise patterns. For example, the 23.7% label error rate in the TADRED dataset is because of inappropriate guidelines (Stoica et al., 2021). Annotation guidelines serve as the instruction manual for annotators, drafted by product owners. The process can be simply summarized as follows: (1) Annotators are recruited and given data samples and the description of guidelines; (2) Annotators provide the labels based on their knowledge and experience, by strictly complying with the guidelines (Klie et al., 2024).

2.2 Pre-trained Language Models

Before the emergence of large language models, studies have shown that pre-trained language models are better than support vector machines or other deep learning models (Ghosh et al., 2019). Many works demonstrate that using external knowledge can effectively enhance the performance of individual-level NLU tasks such as stance detection tasks (He et al., 2022; Hanawa et al., 2019; Li et al., 2021). Since large language models were pre-trained with a large corpus, many researchers began to explore their performance in individual-level tasks such as stance detection (Zhang et al., 2022; Cruickshank and Xian Ng, 2023; Lan et al., 2024; Li and Conrad, 2024; Gatto et al., 2023), sentiment analysis (Zhang et al., 2023; Korkmaz et al., 2023). However, these works focus on how to guide LLMs to achieve better performance, and no work has considered the potential impact of label errors in individual-level NLU tasks. If the dataset is systematically and consistently mislabeled, the evaluation of LLMs can become both misleading and unreliable.

3 Methodology

In this section, we illustrate the process of mitigating label errors in individual-level NLU tasks. We begin by highlighting the unique characteristics of individual-level tasks. Next, we present our methods, using representative tasks such as stance detection and topic-based sentiment analysis.

3.1 Tasks and Dataset Selection

According to the definition of individual-level NLU, it's difficult for annotators to directly infer a publisher's perspectives, but they can approximate them using indirect contextual information about the user. Relying solely on a single piece of text often results in inaccurate annotations. This highlights the critical need for a more comprehensive understanding of an individual's background in NLU tasks, including physiological attributes (e.g., gender, age) and social factors (e.g., interests, occupation, and community affiliations). However, collecting such sensitive information from social media presents significant challenges, particularly regarding privacy concerns. Therefore, it is essential to simplify the problem by focusing on specific individual-level NLU tasks while minimizing privacy risks.

Therefore, we focus on three representative tasks: topic-based sentiment analysis, stance detection, and hate speech detection. One advantage of these tasks and datasets is that they have clearly defined topics or targets, making it easier to collect relevant posts from the users. Other tasks, such as sarcasm detection lack a specific target and often require a deep understanding of an individual's speaking style, interests, and other contextual factors, making data collection significantly more challenging. Additionally, previous studies have shown that users' stances and sentiment toward specific perspectives tend to remain stable over short periods (Borge-Holthoefer et al., 2015; Aldayel and Magdy, 2019). For example, in topic-based sentiment analysis, if the topic is Arsenal, a dedicated Arsenal fan is expected to maintain a positive sentiment toward the team over time. For our study, we select three datasets: the SemEval-2016 Task

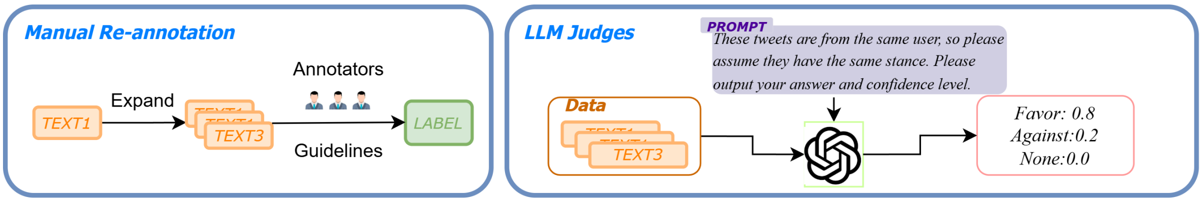


Figure 2: The process of manual re-annotation and LLM judges. In the manual re-annotation, after finding other posts related to the topic/target by individuals (network users), three annotators follow the guidelines to annotate individual-level labels. In LLM judges, the input is divided into two parts: data and prompts.

Dataset	Topics/Targets	Tweets
Semeval Stance	5	1,129
Semeval Sentiment	60	4,346
HateXplain	12	20,148

Table 1: Statistics of three datasets.

4 topic-based sentiment analysis dataset (Nakov et al., 2019), the SemEval-2016 Task 6 stance detection dataset (Mohammad et al., 2016), and the HateXplain dataset (Mathew et al., 2021). The basic statistics of three datasets are presented in Table 1. Stance detection has three classes: $Favor$, $Against$, and $None$. In topic-based sentiment analysis, the creators discard the $Neutral$ and only keep the $Positive$ and $Negative$ classes. HateXplain also has three classes, namely $Hatespeech$, $Offensive$, and $Normal$.

3.2 Data Expansion

To expand the dataset, we collect user posts related to the specified topic or target. We make the following assumption: given a set of posts $X = x_1, x_2, ..., x_n$ authored by a user about a topic or target t over a certain period, these posts should exhibit the same perspective. We further validate this assumption from a clustering perspective. Previous research has clustered posts based on textual features at the text level (Samih and Darwish, 2021), where posts with similar textual characteristics are positioned closer together and are more likely to share the same class label. At the individual level, drawing from prior studies (Borge-Holthoefer et al., 2015; Aldayel and Magdy, 2019), we extend this idea by assuming that users and their posts should be each other's nearest neighbors. In other words, if a dataset contains only one post x_1 from a given user, and we add k additional posts $x_1, ...x_k$, forming a cluster of nearest neighbors that share the same label. Previous studies have shown that detecting label errors requires as few as two nearest neighbor samples (2-NN) (Zhu et al., 2022). so we set $k = 2$ for the dataset creation (we also conduct experiments to demonstrate the impact of k in section 6.2). However, certain edge cases must be considered—such as when a user has only one post related to t, or when the original post itself isn't directly related to t. We provide specific guidelines for handling the cases in Section 3.3.

We start from the existing dataset, find the users corresponding to these posts, and then use the Twitter API to crawl other tweets from the same user within a certain period of time based on the corresponding keywords. This period is usually no more than two years. For example, for the stance dataset, we crawl the user's tweets from January 2015 to December 2016. The keywords (also called search queries) corresponding to different targets are given in the appendix. If more than three tweets are collected from a user, we filter the tweets: We first keep the tweets that explicitly contained the target (e.g., the target was Legalization of Abortion and the tweet explicitly contained abortion). If there are not enough tweets (less than three tweets), we manually collect the user's tweets in the following order: (1) tweets posted by the user related to the target or topic (regardless of time, the closer to the original tweet, the better); (2) tweets retweeted by the user related to the target or topic (regardless of time, the closer to the original tweet, the better); (3) tweets posted by the user closest to the original tweet.

Because some tweets have been deleted or restricted, we can only obtain the users corresponding to a part of the tweets. This is also the case in previous studies (Aldayel and Magdy, 2019). In the stance detection dataset, we selected four targets: Atheism (AT), Climate Change is a Real Concern (CC), Feminist Movement (FM), and Legalization of Abortion (LA) for data expansion. This is because worldview-anchored targets (e.g., atheism, feminism, abortion) than for candidate-centric targets, where attitudes can shift rapidly in response to campaign events or news (Sagiv and Schwartz, 2022). Therefore, we focus our re-annotation on the former and exclude politician targets (Hillary Clinton / Donald Trump) from this study. In topic-based sentiment analysis, we select two tweets for each topic, giving priority to one with the label $Positive$ and one with the label $Negative$. However, if all tweets of a certain class of a certain topic are inaccessible, we select two tweets of the same class. The data statistics are shown in Table 2. In HateXplain, each tweet has three annotators, and we select tweets with inconsistent annotators in the test set (e.g., two annotators annotated "Hatespeech" and one annotator annotated "Offensive"). There are 350 such tweets in total, but only 76 of them are still accessible. Our data expansion assumes that a user's stance toward a target is approximately stable within a bounded time window.

3.3 Manual Re-annotation Guidelines

After collecting user tweets, three annotators (all of them are authors) independently annotated them. Considering that we use LLMs for data evaluation and that the annotators may not understand some background knowledge, we allow the annotators to use search engines to assist in the annotation work but prohibit the use of LLMs.

In the annotation, we first followed the guidelines for three datasets. According to the characteristics of expanded data, we propose a new guideline for stance detection and topic-based sentiment analysis: for individuals whose sentiment/stance is difficult to determine, we use the following rules to annotate: (1) If none of the three posts can determine the individual's stance/sentiment on the target, it is annotated as $None$ ($Neutral$). (2) If one tweet clearly states that the stance is $Favor$ ($Positive$) or $Against$ ($Negetive$), and the remaining two tweets have unclear stances or are irrelevant to the target, the stance is still annotated as $Favor$ ($Positive$) or $Against$ ($Negetive$). (3) If more than half of the tweets are $Favor$ ($Positive$)/$Against$ ($Negetive$), please identify the user's stance/sentiment as $Favor$ ($Positive$)/$Against$ ($Negetive$).

The situation for the HateXplain dataset is more complicated. Hate speech detection is a complex task that combines individual factors and social consensus. First, determining whether a speech is offensive is mainly based on the perspective of readers/social consensus, while determining whether a speech is hate speech requires considering both individual factors and social consensus. Second, unlike stance detection and sentiment analysis tasks, if a post is judged as hate speech, it does not mean that other posts of the user are also hate speech; however, the individual level obtained from other posts can play an auxiliary role in judging whether the current post is hate speech. In fact, there are differences in the definition of hate speech and even the annotation guidelines of different datasets (Jahan and Oussalah, 2023). In HateXplain, $Hatespeech$, $Offensive$, and $Normal$ also rely mainly on the subjective judgment of the annotator, and there is no standard. Compared with the other two tasks, the fuzzy label boundaries of hate speech detection exacerbate the uncertainty of labels. We used the United Nations definition of hate speech (Strossen, 2021) in re-annotation and developed the standards for distinguishing three classes of labels. These standards are also input into LLMs as system prompts. The specific prompts can be found in Appendix B.

The Krippendorff's α values for stance detection, topic-based sentiment analysis, and hate speech detection datasets were 0.68, 0.71, and 0.52, respectively. Clearly, the ambiguity of the hate speech detection boundaries leads to greater inconsistencies. However, our goal is to re-annotate the labels of the original dataset; we cannot simply discard the samples when the three annotators are inconsistent. Therefore, if an inconsistency is found, the annotators will re-search the Twitter user's information and discuss it until they reach a consensus.

4 Experiments

In this section, we introduce the large language models used to evaluate Individual-level NLU performance and then our evaluation metrics.

4.1 LLM Evaluations

We use three representative large language models: GPT-4o (Achiam et al., 2023), Llama3-70B (Dubey et al., 2024), and PHI-4 (Abdin et al., 2024) to evaluate the performance of the datasets after expansion and correction.

To demonstrate that multiple posts are more effective than one post, we also conduct two ablation experiments. The first ablation experiment compares the performance when using the original tweet and two newly collected tweets and the per-

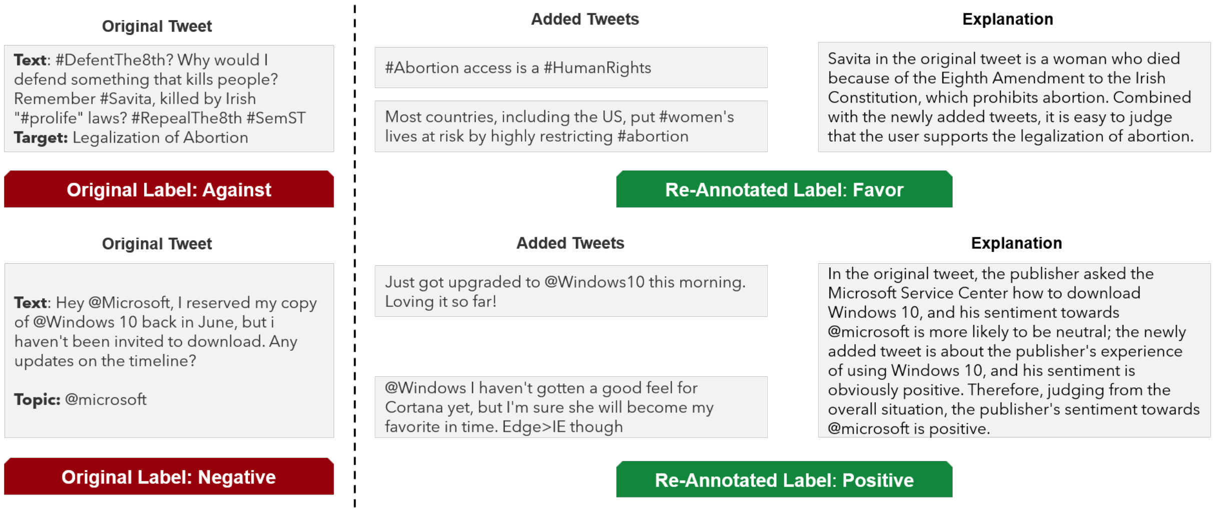


Figure 3: Some examples of correcting label errors. Using multiple posts from the same publisher can more accurately determine the user's sentiment or stance, and can effectively explain why this label is given.

formance when using the original tweet. The second ablation experiment verifies the performance of LLM when using different numbers of tweets. In the second ablation experiment, not all users can collect more than three tweets, so we only use users with more than or equal to five tweets collected in the stance detection dataset, and then randomly select one to five tweets from these users to input into LLM to evaluate the performance. In the one-tweet experiment, the input tweet can be different from the tweets in the original dataset.

4.2 Evaluation Metrics

Similar to previous work, we calculate the label error rate R_e according to (1).

$$R_e = S_e/S_t \tag{1}$$

S_e is the number of error samples, and S_t is the total number of samples. Previous work (Mohammad et al., 2016; Nakov et al., 2019) uses the average F1 value of positive and negative samples to evaluate model performance. However, we find that in some targets, the number of positive or negative samples that can still be accessed is very small, and directly using the average F1 value will cause a large bias. Thus we use the $Accuracy$ for evaluation. However, in the appendix, we also give the average F1 value of each model.

$$Accuracy = S_c/S_t \tag{2}$$

S_c is the number of samples predicted correctly. Since a user may have multiple tweets in the original dataset, each one may be annotated with a different label, we calculate R_e and $Accuracy$ based on the number of tweets rather than the number of users. This allows us to better compare our method with the original dataset.

Dataset	Target	Posts	Error	Error Rate
Stance	AT	133	30	29.3%
	CC	114	22	19.3%
	FM	114	50	43.9%
	LA	130	54	41.5%
	ALL	491	156	31.7%
Sentiment	All Topic	120	28	23.3%
HateXplain	All Topic	76	37	48.7%

Table 2: Error rate statistics of different datasets

5 Assessing Label Errors

According to our guidelines, we evaluated the labeling errors of the three datasets. In the stance detection dataset, there were 156 tweets with incorrect labels. The error rate was as high as 31.7%. Among them, the error rates of Atheism, Feminism, and Legalization of Abortion were as high as 29.3%, 43.9% and 41.5% respectively. In the topic-based sentiment analysis dataset, the error rate is 23.3%. In the hate speech detection dataset, the error rate is 48.7%. This may be related to the unclear standards when the dataset was constructed and the fuzzy boundaries between the three categories of labels. Table 2 shows the tweets and error rates in different targets or topics in the three datasets.

We then perform a qualitative analysis of the

	Stance Detection		Sentiment Analysis		Hate Speech Detection	
Model	Total	Total (OL)	Total	Total (OL)	Total	Total (OL)
GPT-4o	92.46	64.77	89.17	78.33	65.79	46.05
LLama3-70B	87.58	64.56	92.50	78.33	60.53	38.16
PHI-4	79.02	61.51	92.50	76.67	53.95	35.52

Table 3: The performance of different models on the dataset after label correction. OL means original labels, which is the label of the original datasets without correction.

errors in these labels. In the stance detection dataset, the target of Legalization of Abortion, a hashtag $\#repealthe8th$ repeatedly appears, which often means that the user is Irish and opposes the Eighth Amendment to the Irish Constitution. The Eighth Amendment to the Irish Constitution is a law against the Legalization of Abortion. Opposing the law means that the user's stance on the Legalization of Abortion is $Favor$. However, in the original dataset, a large number of tweets are annotated as $Against$. This is most likely because the annotators are not Irish and do not understand Irish culture and politics. In hate speech detection, mistakes often arise when annotations ignore individual-level context and pragmatic factors (e.g., speaker intent, in-group vs. out-group usage, quotation/irony, and the conversational setting). For instance, posts containing racial slurs may be labeled differently depending on whether the term is used to target a protected group or appears in reclaimed, self-referential, or quoted contexts; failing to account for such context can blur the boundary between "hate speech" and "offensive language," illustrating the complexity of reliable annotation. More examples are given in Figure 3. We also give more examples in the appendix figure.

6 Evaluation on Expanded Datasets

6.1 Quantitative Analysis

We first evaluate the performance of the three models on the new datasets. As shown in Table 3, GPT-4o and Llama-3 show good performance on stance detection and sentiment analysis tasks. PHI-4 performs slightly worse on the stance detection task; However, if we use the original labels (uncorrected dataset labels, OL) for evaluation, the accuracy of the three models will drop significantly. The three models perform slightly worse in the hate speech detection task, but are still better than the accuracy using the original labels. This shows that label errors in the original dataset will seriously affect the evaluation of model performance,

	SD		SA		HSD	
Model	MT	ST	MT	ST	MT	ST
GPT-4o	92.48	69.25	89.17	74.17	65.79	63.16
LLama3	87.58	74.13	92.50	72.50	60.53	59.21
PHI-4	79.02	60.29	92.50	71.67	53.95	50.00

Table 4: Comparison of results using multiple tweets and a single tweet. MT: Multiple Tweets. ST: Single Tweet. SD: Stance Detection. SA: Sentiment Analysis. HSD: Hate Speech Detection

6.2 Ablation Studies

We conduct two ablation experiments to evaluate the validity of multiple tweets from the same user. Table 4 shows the results of the first ablation experiment. When using only the original tweet, the accuracy of all LLMs drops. This proves the necessity of using individual factors. Different posts from the same individual can complement each other and enhance the accuracy of prediction.

Then we input LLM with one to five tweets from the same user in the stance detection task. We collected 281 users with more than five tweets, so we evaluated the effectiveness of multiple tweets on these 281 users. Since PHI-4 performed poorly before, we used LLama3-70B and GPT-4o for experiments. Figure 5 shows that three tweets can achieve good accuracy. Although the performance can continue to improve by increasing the number of tweets, the improvement is significantly reduced. Therefore, using three tweets is a choice that takes both performance and efficiency into consideration.

6.3 Case Study

We also conduct case studies of the results given by LLMs. We focus on two types of samples: The first type is samples where the new label is different from the original label. The second type is samples where the new label is the same as the original label, but the prediction results are different when using multiple tweets and a single tweet. We find that

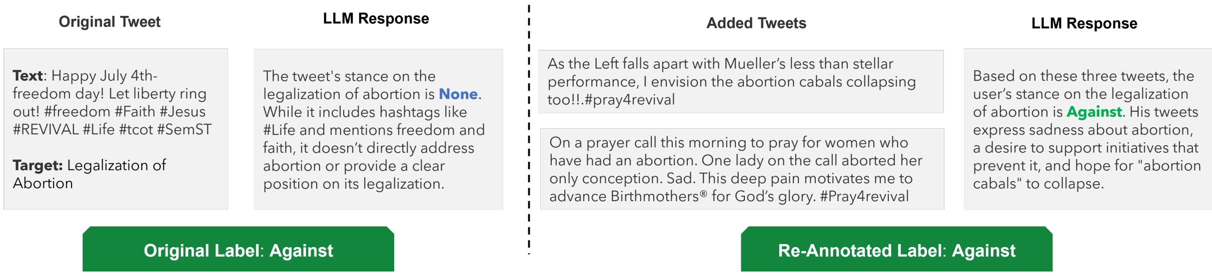


Figure 4: Multi-posts example. In this case, although the user's stance is "Against" in both the new and original datasets, it is difficult or even impossible to infer the user's stance from the text in the original dataset. After adding other tweets from the user, LLM gives an accurate prediction.

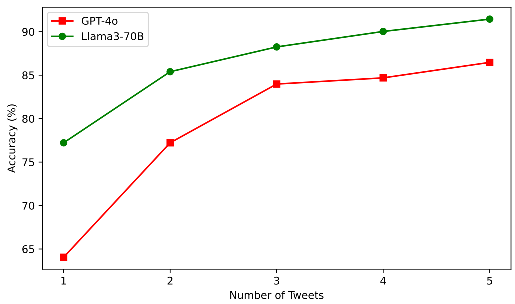


Figure 5: Performance on the Semeval stance detection dataset using different numbers of tweets as LLM input.

LLMs' explanations were basically consistent with the annotators' cognition. Figure 4 shows a typical example. The sample is annotated "Against" in both the new and original datasets, but even humans find it difficult to judge the user's stance on Legalization of Abortion through the original tweets. All three annotators also believe that the original tweet did not mention abortion at all, nor did it contain any clues supporting or opposing abortion. When only one tweet is used for stance detection, LLMs give a prediction result of "None", which is consistent with the annotator's cognition. After using the newly added two tweets, a total of three tweets for prediction, LLMs give the result of "Against". The three annotators also give the label "Against" based on the newly added tweets. This proves that expanding the dataset and increasing the information of the same user in the dataset is crucial for individual-level NLU. And further shows that annotation of individual tweets is prone to labeling errors. More samples are given in Appendix D.

7 Discussion and Conclusion

Our research demonstrates the limitations of reducing individual-level NLU tasks to text-level tasks. The information lost in the reduction process not only leads to poor model performance but also causes annotators to misunderstand semantic information, resulting in a large number of label errors. Therefore, we call on dataset creators to fully consider social factors and reasonably choose guidelines to reduce systematic label errors when creating individual-level datasets in the future.

Past studies have shown that online users' perspectives of a topic or target do not change over a period of time (Aldayel and Magdy, 2019). We draw inspiration from these conclusions and propose an individual-level annotation guideline. We collect posts related to topics/targets from online users over a period of time, use the consistency of the posts for cross-validation, and finally judge the perspectives of the online users. Case studies show that our method avoids the ambiguous semantics of a single post, allowing for more accurate annotation, and the labels we give are more explainable. At the same time, our method only collects posts to avoid collecting a large amount of invalid user information.

We used the re-annotated dataset to conduct zero-shot experiments on different LLMs. Comparing the labels with the original datasets, we found that incorrect labels seriously affect the evaluation model's performance on the individual-level NLU task; the current LLM performs exceptionally well in stance detection and topic-based semantic analysis. The hate speech detection performance is relatively poor, which may be related to the fuzzy hate speech boundaries. Through ablation experiments and case studies, we demonstrated the effectiveness of multiple posts compared to a single post and also showed that LLMs have human thinking patterns when facing single and multiple tweets.

8 Limitations

Our study also has some limitations. First, in individual-level NLU, the user's perspectives can be determined not only through the tweets posted by the user but also by using other information of the user. For example, the user's retweets, likes, follows, and profile. Although these data are rich, they are highly heterogeneous compared to the tweets posted by the user. For example, some user profiles may contain information to determine the user's stance, while some users may not even have profiles. This may be because different users have different habits when using social media. Insufficient information has a great impact on hate speech detection. It can also be seen from the article that the accuracy of hate speech detection is low. Effectively utilizing and modeling this information is one of our future directions.

Secondly, our current information retrieval methods are only applicable to tasks that involve determining topics, such as stance detection and topic-based sentiment analysis. The characteristic of this type of task is that we can use keywords to retrieve user posts. Some other individual-level NLU tasks, such as sarcasm detection, do not have similar characteristics and cannot find corresponding user tweets by keywords. This means that when facing this type of NLU task, we need new information retrieval methods and models. This is also the direction we need to explore.

Finally, our annotations are relatively small. Individual-level annotations require full consideration of each post, which greatly increases the annotation cost. The large number of tweets deleted and restricted further limits the number of tweets we can annotate. In particular, in the HateXplain dataset, only about 20% of tweets are still online. To verify the robustness of our method, we will increase the number of annotation samples and build a larger dataset in the future.

9 Ethics Statement

Our work on the datasets is conducted with a strong commitment to ethical principles. We prioritize privacy by collecting only publicly available tweets and strictly adhering to relevant guidelines for annotation and dataset sharing. In our research, we comply with the X Developer Agreement and Policy, ensuring that all content is used solely for academic research purposes. Tweets can only identify online users, not real individuals. Furthermore, we respect diverse religious beliefs and political perspectives.

Additionally, our research does not diminish the contributions of previous dataset creators; rather, we deeply appreciate their efforts. The datasets they developed serve as the foundation of our work.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

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*.

Abeer Aldayel and Walid Magdy. 2019. Your stance is exposed! analysing possible factors for stance detection on social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–20.

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

Prasanta Bhattacharya, Hong Zhang, Yiming Cao, Wei Gao, Brandon Siyuan Loh, Joseph JP Simons, and Liang Ze Wong. 2025. Rethinking stance detection: A theoretically-informed research agenda for user-level inference using language models. *arXiv preprint arXiv:2502.02074*.

Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. Content and network dynamics behind egyptian political polarization on twitter. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 700–711.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Guillermo F Cabrera, Christopher J Miller, and Jeff Schneider. 2014. Systematic labeling bias: Debiasing where everyone is wrong. In *2014 22nd International Conference on Pattern Recognition*, pages 4417–4422. IEEE.

Iain J Cruickshank and Lynnette Hui Xian Ng. 2023. Use of large language models for stance classification. *arXiv e-prints*, pages arXiv–2309.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Krishna Garg and Cornelia Caragea. 2024. Stanceformer: Target-aware transformer for stance detection. *arXiv preprint arXiv:2410.07083*.

Joseph Gatto, Omar Sharif, and Sarah Masud Preum. 2023. Chain-of-thought embeddings for stance detection on social media. *arXiv preprint arXiv:2310.19750*.

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance detection in web and social media: a comparative study. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pages 75–87. Springer.

Kazuaki Hanawa, Akira Sasaki, Naoaki Okazaki, and Kentaro Inui. 2019. Stance detection attending external knowledge from wikipedia. *Journal of Information Processing*, 27:499–506.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from wikipedia to enhance stance detection. *arXiv preprint arXiv:2204.03839*.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 588–602.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? *arXiv preprint arXiv:2403.08213*.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Paul Kockelman. 2004. Stance and subjectivity. *Journal of Linguistic Anthropology*, 14(2):127–150.

Adem Korkmaz, Cemal Aktürk, and Tarik Talan. 2023. Analyzing the user's sentiments of chatgpt using twitter data. *Iraqi Journal For Computer Science and Mathematics*, 4(2):202–214.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 891–903.

Mao Li and Frederick Conrad. 2024. Advancing annotation of stance in social media posts: A comparative analysis of large language models and crowd sourcing. *arXiv preprint arXiv:2406.07483*.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2021. Improving stance detection with multi-dataset learning and knowledge distillation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6332–6345.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2019. Semeval-2016 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.01973*.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2019. Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*.

Lilach Sagiv and Shalom H. Schwartz. 2022. Personal values across cultures. *Annual Review of Psychology*.

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

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13843–13850.

Nadine Strossen. 2021. United nations free speech standards as the global benchmark for online platforms' hate speech policies. *Mich. St. Int'l L. Rev.*, 29:307.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Yi Yang and Jacob Eisenstein. 2017. Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics*, 5:295–307.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zhaowei Zhu, Zihao Dong, and Yang Liu. 2022. Detecting corrupted labels without training a model to predict. In *International conference on machine learning*, pages 27412–27427. PMLR.

A Keywords in searching

In topic-based sentiment analysis and hate speech detection, we directly use topics as the keyword for search. In stance detection, the keywords are:

- Atheism: $Atheism$, God, $Pray$
- Climate Change is a Real Concern: $Climate$, $Globalwarming$
- Feminist Movement: $Women$, $Feminism$, $Feminist$
- Legalization of Abortion: $Abortion$, $Women$, $Legal$

B Prompts Design

We use a very simple prompt:

- For stance detection:

  Read the question, provide your answer, and your confidence in this answer. Please make sure that the confidence level of your answers adds up to 1. Only output confidence levels. Do not output any other things. Please decide the following users' stance on the $Target$: Is it FAVOR, AGAINST, or NONE? These tweets are from the same user, so please assume they have the same stance. $Tweet1$ $Tweet2$ $Tweet3$

- For topic-based sentiment analysis:

  Read the question, provide your answer, and your confidence in this answer. Please make sure that the confidence level of your answers adds up to 1. Only output confidence levels. Do not output any other things. Please decide the following users' sentiment on the $Topic$: Is it POSITIVE or NEGATIVE? These tweets are from the same user, so please assume they have the same sentiment. $Tweet1$ $Tweet2$ $Tweet3$

- For hate speech detection:

  The definition of hate speech and offensive speech is as follows: hate speech refers to offensive discourse targeting a group or an individual based on inherent characteristics (such as race, religion or gender) and that may threaten social peace. Note that hate speech can only be directed at individuals or groups of individuals. It does not include communication about States and their offices, symbols or public officials, nor about religious leaders or tenets of faith. Please consider both the individual factors of the publisher and social consensus before judging it as hate speech. For example, if a black person uses 'Nigger', it is usually not judged as hate speech, but if a white person uses 'Nigger', it is considered hate speech.

  Offensive speech is defined as speech that makes others feel uncomfortable, insulted, or degraded, but does not necessarily have the consensus intention of systematic discrimination or inciting hatred. When judging offensive speech, it is necessary to consider social consensus, that is, whether the tweet is considered appropriate by the majority of people or a certain group in society. For example, some black people do not agree with the use of the word 'Nigger' among black people, then even if the publisher is black, please Judge the speech as offensive speech.

Other speeches are normal speech.

Please decide whether this tweet is hate speech, offensive speech, or normal speech. Only output hatespeech, offensive or normal, don't output any other things: $Tweet1$.

You can use these two tweets to help you make your judgment, they are other tweets posted by the same user: $Tweet2$ $Tweet3$

C Average F1 value

In previous work, the average F1 value of positive and negative samples was often used to evaluate model performance of stance detection and sentiment analysis. The formula is as follows:

$$F_{avg} = \frac{F_P + F_N}{2} \tag{3}$$

F_P is the F1 value of the positive sample, and F_N is the F1 value of the negative sample. In stance detection, positive samples are samples with the label $Favor$, and negative samples are samples with the label $Against$.

	Stance Detection		Sentiment Analysis	
Model	Total	Total (OL)	Total	Total (OL)
GPT-4o	94.36	68.61	88.44	77.08
LLama3-70B	90.29	68.94	92.06	76.67
PHI-4	78.87	65.96	92.06	74.88

Table 5: The average F1 value of different models on the dataset after label correction. OL means original labels, which is the label of the original datasets without correction.

D More Error Samples and LLM Responses

Original Tweet: These days, the cool kids are atheists. #freethinker #SemST
Target: Atheism
Original label: Against

Added Tweet 1: Just a reminder that supernatural entities, e.g. Odin or God, had nothing to do with creating today's #WinterSolstice. **Added Tweet 2**: Right, and it seems highly likely that both the afterlife and God were completely made up by people and don't actually exist.
Re-annotated label: Favor
Explanation: In the two newly added tweets, the publisher denied that God or Odin created the "winter solstice" and believed that afterlife and God were both human imaginations and not real existences. Combined with the original tweets, the publisher obviously supports Atheism.
LLM response: The user's stance on Atheism is Favor based on their tweets denying the existence of God and the afterlife, as well as their endorsement of atheism as a perspective.

Original Tweet: avg house in US consumes 10,656 kWh per year 2006, Gore devoured nearly 221,000 kWh more than 20 X the nat average. #onpoli #SemST
Target: Climate Change is a real concern
Original label: Favor

Added Tweet 1: How many people in room at climate change rah rah have significant investments/salaries from Green Industry taking taxpayer $$? #onpoli
Added Tweet 2: @sunlorrie it's okay Lorrie, they are going to fix world climate next, same way they fixed hydro. #onpoli
Re-annotated label: Against
Explanation: The original tweet only describes the electricity consumption in a certain area. It is difficult to judge whether the tweet has a Favor stance. However, in the two added posts, the publisher is very dissatisfied with the use of climate change to make profits. Overall, the publisher's stance should be Against.
LLM response: The user's stance on Climate change is a real concern appears to be Against. Their tweets criticize prominent climate advocates (e.g., Al Gore's energy consumption), question the financial interests of those promoting climate action, and express skepticism toward climate policies.

Original Tweet: Feminism is a hate group!! - anti fem Twitter user that does nothing but harass and spread hate. #SemST
Target: Feminist Movement
Original label: Against

Added Tweet 1: If men have more rights than women, that's not equality. Keeping up? @Indoctr1n3 @Grumpy_P_Sloth @LeexxxW
Added Tweet 2: It's amazing how many people scream at feminists that we are equal and then tweet about some form of inequality seconds later @SjwNation
Re-annotated label: Favor
Explanation: The publisher expressed the idea of gender equality in the newly added tweet 1; in the newly added tweet 2, the publisher expressed sarcasm towards anti-feminists. Combined with the original tweet, the original tweet also expressed sarcasm towards anti- feminists, so the user should be supporting the feminist movement.
LLM response: The user's stance on the Feminist Movement appears to be Favor. Their tweets defend feminism against critics, highlight gender inequality, and push back against those who claim feminism is unnecessary.

Original Tweet: Amazon Prime Day: What all the fuss is about: Amazon's Prime Day promised massive deals rivaling Black Friday,... http://usat.ly/1HO0P1i
Topic: Amazon Prime Day
Original label: Negative

Added Tweet 1: Amazon's "Prime Day" sparks summer sales fury: Amazon's Prime Day launches a summer sales fury. http://usat.ly/1HE2xCn
Added Tweet 2: Amazon Prime Day deals that beat Black Friday: Prime Day's best deals http://usat.ly/29Ej6Js
Re-annotated label: Positive
Explanation: All three tweets describe the advantages of Amazon Prime Day. Obviously, the sentiment of the publisher should be positive.
LLM response: The user's sentiment on Amazon Prime Day appears to be Positive. Their tweets highlight the event's major deals and compare it favorably to Black Friday without any negative commentary.

Original Tweet: there's too many wypipo asking what this is and if they can get in.... this is a niggers only event
Topic: African
Original label: hatespeech

Added Tweet 1: just because someone has notoriety doesn't mean they're not a complete dumbass. look at ben carson. we as a black community cannot grow if we still view such arbitrary things like race mixing as detrimental. so once again, your statements were ignorant and of ill taste
Added Tweet 2: y'all really don't get how much seeing representation means to minorities. like i definitely acknowledge my privilege as a lighter black girl but seeing dark skins getting important roles outside the "angry black girl" EXCITES me
Re-annotated label: offensive
Explanation: From the two new tweets, we can judge that the publisher is black; some black people think that using "N-word" is inappropriate, so the label is offensive.
LLM response: It is not hate speech (given the speaker's identity and intent), but it is offensive speech, due to its racially exclusionary nature and use of a highly sensitive term in a public setting.