Why do Large Language Models Judge Differently than Humans? An Examination of Sentiment Analysis of Movie Reviews

Nils Messerschmidt
SnT, University of Luxembourg
nils.messerschmidt@uni.lu

Amir Sartipi
SnT, University of Luxembourg
amir.sartipi@uni.lu

Antragama Ewa Abbas
SnT, University of Luxembourg
antragama.abbas@uni.lu

Orestis Papageorgiou
SnT, University of Luxembourg
orestis.papageorgiou@uni.lu

Igor Tchappi
SnT, University of Luxembourg
igor.tchappi@uni.lu

Gilbert Fridgen
SnT, University of Luxembourg
gilbert.fridgen@uni.lu

Abstract

This research investigates the root causes of divergence between Large Language Model (LLM)-based and human sentiment judgments. Using an inductive approach, we qualitatively analyzed a movie review dataset and identify two main causes: (i) contextual statements, where sentiment depends on situational factors (e.g., describing a film as "childish" may be positive for younger audiences but negative for adults); and (ii) linguistic statements, where sentiment shifts due to complex constructions such as sarcasm or double negation. Our study thus highlights the importance of both context (where, when, and for whom a statement is made) and linguistic form (how it is phrased) in sentiment interpretation. We contribute to the literature by identifying justificatory mechanisms behind differences in sentiment judgments between humans and LLMs. This may initiate a broader discourse on whether machine-generated sentiment can serve as a valid proxy for human interpretation. Even more, human-in-the-middle approaches may still outperform solely LLM-based sentiment interpretations.

Keywords: Generative Artificial Intelligence, Large Language Models, Sentiment Analysis, Stanford Sentiment Treebank.

1. Introduction

The sentiment analysis market will grow from USD 5.1 billion in 2024 to USD 11.4 billion by 2030, reflecting an average annual growth rate of 14.3% (Research and Markets, 2025). Organizations apply sentiment analysis to extract customer opinions from various sources, including social media, blogs, and online forums. Given its relevance to economic activity, sentiment analysis has long attracted attention from Information Systems (IS) scholars (e.g., Abbasi et al., 2008), with renewed interest following the emergence of LLMs (e.g., Mukta & Islam, 2025). Although still limited, IS researchers have begun to explore the use of LLMs for sentiment analysis, resulting in two main literature streams. The first *proposes* an LLM-based approach for sentiment analysis (e.g., Ipa et al., 2024; Mukta & Islam, 2025; Xing, 2025), while the second *compares* the performance of established P-trained Language Models (PLMs) such as RoBERTa (e.g., Gautam et al., 2025; Li et al., 2024) in the realm of Natural Language processing (NLP).

Surprisingly, what remains lacking is (a) an evaluative comparison between LLM-based sentiment analysis and human judgment, and (b) a theoretical evaluation of why their judgments differ. This knowledge gap is problematic because the main purpose of sentiment analysis is to approximate human satisfaction. Without assessing the alignment between LLM- and human-based judgment, it implies that the IS community may *implicitly assume* that machine-generated sentiment is a valid proxy for human interpretation. However, this assumption may not hold. Previous works show that even mature sentiment analysis tools struggle with limitations (e.g., Wankhade et al., 2022). For instance, such tools struggle to handle ambivalent reviews that contain both positive and negative statements (Abulaish et al., 2020). What remains unclear is whether LLMs experience the same struggles, posing epistemic risks if their outputs are assumed to be accurate. In such cases, organizations may optimize decisions on sentiment analysis that misrepresents how customers actually feel.

Against this backdrop, this research investigates the root causes of divergence between LLM-based and

URI: https://hdl.handle.net/10125/111524
978-0-9981331-9-5

human sentiment judgments. To achieve our objective, we employed an inductive approach, allowing the data to guide our analysis rather than starting with predefined hypotheses. This approach is appropriate, as we found no prior studies that explicitly compare LLM outputs with human judgments in sentiment analysis. Specifically, we qualitatively examined the well-known Stanford Sentiment Treebank (SST) movie review dataset through the content coding technique. We contribute to IS literature at the intersection of LLMs and sentiment analysis by offering insights to inform future theorizing on why LLMs judge sentiment differently from humans. In short, reliable LLM-based sentiment analysis requires equal attention to both *where*, *when* and *for whom* a statement is made (context) and to *how* it is phrased (linguistics).

2. Research Background

Sentiment analysis has evolved from rule- and lexicon-based methods (Kour et al., 2021), limited by their static nature, to machine learning approaches that relied on manual feature engineering (e.g., TF–IDF, n-grams, POS tags). Deep learning further advanced the field through neural network architectures such as CNNs and LSTMs, supported by distributed word representations like Word2Vec and GloVe (Kim, 2014). The major breakthrough came with the Transformer architecture and PLMs like BERT and RoBERTa (e.g., Zhuang et al., 2021), which achieved state-of-the-art results with minimal task-specific fine-tuning.

Limitations of previous approaches include the fact that, although PLMs were powerful, their performance without fine-tuning was limited, and they struggled with reasoning over implicit sentiment, which required task-specific adaptation. These shortcomings have driven research toward more advanced models. LLMs like GPT-4 leverage large-scale pretraining with substantially more parameters and introduce zero-shot and few-shot learning. This enables sentiment analysis directly from instructions or a handful of examples without fine-tuning, thereby surpassing prior approaches (OpenAI, 2024). These capabilities demonstrate the potential of LLMs for a more accurate understanding and application in sentiment analysis.

To demonstrate that LLMs outperform prior approaches, researchers typically conduct two evaluations: (1) error analysis and (2) benchmarking against existing tools. Error analyses suggest that LLMs often stumble on core linguistic challenges, for instance, they tend to underestimate emotional depth and consistently misinterpret sarcasm or irony (Bojić et al., 2025). Subtle structural changes, such as negation, double negation, or entity substitution, can drastically shift their outputs (Yasunaga et al., 2022). These difficulties are compounded by their limited ability to generalize relational patterns: a model trained on "A is B" may still fail to infer the inverse (Berglund et al., 2024).

Another stream of literature has focused on benchmarking LLMs against earlier methods or across large task suites (Gautam et al., 2025; Li et al., 2024). For instance, evaluations on thirteen sentiment-related tasks found that while LLMs excel at simple classification, they continue to struggle with nuanced phrasings that require deeper linguistic understanding (Zhang et al., 2024). Similarly, Bavaresco et al. (2025) introduced a benchmark spanning twenty NLP tasks and showed that, although LLMs occasionally align with human judgments, their reliability varies widely across tasks and properties.

To sum up, most existing works provide broad quantitative comparisons. While useful as early signals, these studies rarely compare LLM outputs directly to human judgments. As a result, we risk claiming LLM superiority based on invalid proxies (i.e., comparisons to earlier approaches that introduce their own limitations). A deeper analysis is thus needed to uncover the root causes of divergence. To our knowledge, only one study has taken a qualitative approach to assess LLM-based sentiment analysis outcomes (Ochieng et al., 2024). While the analysis followed a predefined framework, we adopted a more inductive approach that allowed patterns to emerge from the data. This opened the opportunity to uncover alternative explanations that may have been overlooked using a fixed lens.

3. Methodology

3.1. Data Collection

We qualitatively investigated the divergence of human and LLM-based sentiment judgment of movie reviews in the SST dataset introduced by Socher et al. (2013), one of the most widely used benchmark corpora in sentiment analysis research. The objectives of the analysis were twofold: a) to identify why LLMs and human annotators diverge in their sentiment judgments using inductive reasoning, and b) to act as a "human-in-the-middle" to make a judgment about the sentiment. For both objectives, the process was repeated until an agreement between both researchers was achieved. For the former objective, the researchers chose a qualitative content analysis approach (Mayring, 2014) while using elements of grounded theory to initially discover divergence categories (i.e., open

coding) and later cluster them according to common themes (i.e., axial coding). This approach is suitable because it allows for in-depth examination of individual statements, enabling the identification of patterns that go beyond purely linguistic features.

The dataset is built based on movie reviews collected from *rottentomatoes.com* (originally collected by Pang and Lee (2005) and consists of 215,154 sentences, with the sentiment of each sentence being evaluated by three independent Amazon Mechanical Turk annotators (with approximately 7% of the sentences being double-checked by up to 6 annotators). For each sentence, the annotators assigned a score between 1 and 25, where 1 represents a completely negative sentiment, and 25 represents a completely positive sentiment. For each sentence, the sentiment scores provided by the annotators were averaged, and the resulting scores were scaled to take values in [0,1] using min–max scaling.

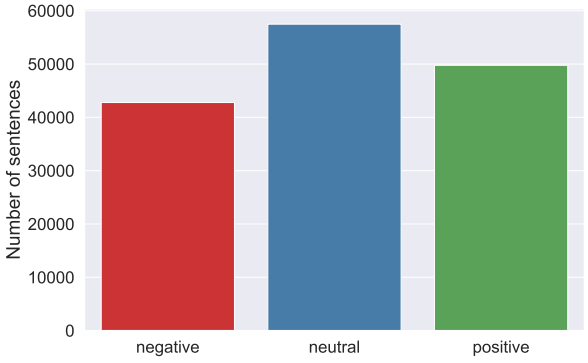
For our analysis, and to ensure that each sentence contained adequate contextual information, we removed sentences that contained fewer than 20 characters. This pre-processing step resulted in a dataset comprising 150,030 sentences. Finally, we categorized the normalized sentiment scores into three clusters. Sentences with scores in the interval [0, 0.4] were labeled as having a negative sentiment, (0.4, 0.6] were labeled as neutral, and (0.6, 1] were labeled as positive.

We then used OpenAI's GPT-4o model to evaluate the sentiment of the filtered sentences. We chose GPT-4o because it is among the most widely adopted LLMs. Each sentence was submitted to the GPT-4o API using the following prompt: *Your role is to analyze the sentiment of the entire text provided and strictly classify it into exactly one of the following three categories: Positive, Negative, or Neutral. You must not deviate from these categories. Provide a justification for your classification, basing it exclusively on the content of the text provided. Do not infer sentiment from context outside the given text or use assumptions. Now, analyze the following text:* {*text*}

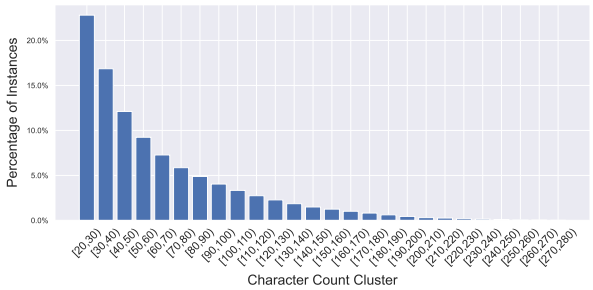
We crafted the prompt to explicitly define the sentiment-analysis task, prevent any deviation, and require the model to justify each classification, as prior research indicates that these measures improve its accuracy (Bu et al., 2024; Sun et al., 2024; White et al., 2023). To further enhance its accuracy, we assigned the role of "sentiment analysis expert" to the model and set $temperature = 0$, towards more deterministic outputs supporting reproducibility.

3.2. Data Analysis

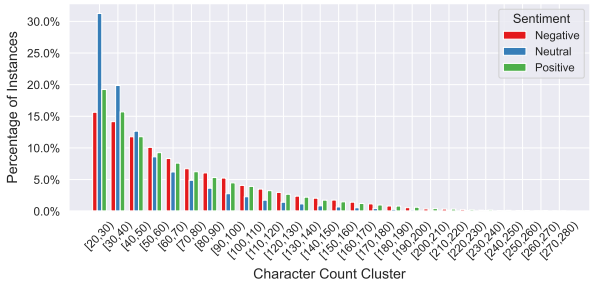
Overall, the resulting dataset is reasonably balanced, with 38.31% of the sentences being neutral and 33.17% positive (Figure 1a). Sentence lengths tend towards a smaller number of characters, with 39.6% containing fewer than 40 characters, and 53.6% fall between 40 and 130 characters (Figure 1b). Among the neutral sentences, 51.12%, have fewer than 40 characters. In contrast, longer sentences (over 40 characters) are predominantly negative, with 70.2% of all negative sentences belonging in this cluster (Figure 1c).



(a) Sentiment distribution.



(b) Distribution of sentence lengths (in characters).



(c) Sentiment distribution against sentence length.

Figure 1: Dataset composition.

We observe that alignment between human and GPT-4o annotations was highest for negative sentences with 82.6%, declining to 68% for neutral and 63.6%

for positive sentences. We attribute this difference to sentence length, as most negative sentences in our dataset are longer (Figure 1b), providing the model with richer context to infer the sentiment. Conversely, many of the neutral and positive sentences contain fewer than 40 characters, which likely hinders GPT-4o's ability to identify their sentiment. Figure 2 provides a summary of the alignment between human and LLM sentiment annotations in our dataset.

Subsequently, we focused on the statements with opposite sentiment judgments between human annotators and the LLM, yielding a total of 2,735 statements. Of those, 230 (8.4%) capture cases in which the LLM judged the statement positively, whereas the human annotators judged it negatively (hereafter referred to as *Type 1 statement*). The other 2505 (91.6%) cases are the opposite, in which the human annotators judged the statements positively, while the LLM judged it negatively (hereafter referred to as *Type 2 statement*).

In an effort to focus only on the most relevant statements and also achieve a balance between Type 1 and Type 2 statements, we narrowed down the subset even further. While the positive judgments ranged from an average human annotator score of 0.6 to 1, only 227 statements were considered strong positive (i.e., having an average score of 0.8 or above). This subset was consequently deemed appropriate as it is closely tied to the 230 Type 1 entries (ranging from 0 to 0.4).

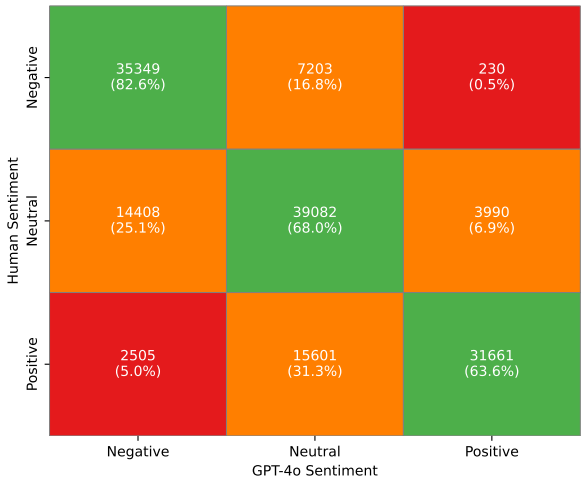


Figure 2: Human and LLM sentiment confusion matrix.

4. Results

The content analysis of the final subset containing 457 statements yields two broader categories of divergence. These statements are contextual (i.e., sentiment is inferred from knowledge outside the given review) and linguistic (i.e., sentiment is inferred from within the given review). Both are, in turn, assigned to several inductively derived sub-categories, which are presented in the following. Note that the term "reviewer" refers to the author of the movie review, not to be confused with the "annotator" who later assesses the sentiment from the given review. In the following, each sub-category is explained, and exemplary codes are provided. Tables 1 and 2 show a quantitative overview of each sub-category.

4.1. Contextual statements

Movie reviewers using contextual statements incorporate situational factors outside of the given movie into their review, such as prior knowledge or individual preferences. A total of 175 codes portraying contextual statements were identified. Depending on the type of situational factor, this can include conditional, combining, referencing, and temporal statements.

Table 1: Numerical Overview of Contextual Statements

Sub-Category	Total	Type 1	Type 2
Conditional Statement	49	31	18
Combining Statement	67	39	28
Referencing Statement	38	34	4
Temporal Statement	21	14	7

Conditional statement. This sub-category captures statements where the perception is subject to specific conditions being met. If viewers fulfill the condition, they will consequently perceive the movie positively; otherwise, they will not. This includes demographic and sociopolitical attributes of viewers, and further includes the state of mind a viewer needs to be in.

Three main demographic attributes are captured in the codes. One such attribute is age: The movie reviewers mention that, while the movie may not be a fit for every audience, young viewers would enjoy it: "*A young audience which will probably be perfectly happy with the sloppy slapstick comedy*" (human: negative [0.37]; LLM: positive). The second demographic attribute frequently referenced by reviewers is ethnicity. One review suggests that, without being part of a specific ethnic group, one might not understand the movie: "*American Chai encourages rueful laughter at stereotypes only an Indian-American would recognize*" (human: negative [0.37]; LLM: positive). The last common demographic attribute refers to the nationality or place of residence. One reviewer suggests that the perception of the movie, or its relevance, might differ depending on the viewer's nationality: "*A fascinating,*

bombshell documentary that should shame Americans, regardless of whether or not ultimate blame finally lies with Kissinger. Should be required viewing for civics classes and would-be public servants alike." (human: positive [0.91]; LLM: negative).

Sociopolitical attributes capture personal beliefs and societal actions that define whether someone is part of a societal subgroup, such as supporting a social movement. One commonly named theme is the context of empowerment: "*Shrewd feminist fairy tale*" (human: negative [0.39]; LLM: positive). Other themes are more politically motivated, ranging from issues related to trust in governments to conspiracy theories: "*Delivers a powerful commentary on how governments lie*" (human: positive [0.8]; LLM: negative). Lastly, conditions can also relate to religious beliefs: "*Christians sensitive to a reductionist view of their Lord as a luv-spreading Dr. Feelgood or omnipotent slacker will feel vastly more affronted than secularists, who might even praise God for delivering such an instant camp classic.*" (human: positive [0.8]; LLM: negative).

Lastly, a common theme of reviewers is to use conditional statements referring to the emotional state of mind one needs to be in to enjoy a given movie: "*This ready-made midnight movie probably won't stand the cold light of day but under the right conditions it's goofy - if not entirely wholesome - fun*" (human: negative [0.39]; LLM: positive). Frequently, annotators mention that viewers need to be resilient to explicit content, such as profanity or violence, highlighting that it is not suited for every audience: "*If you can get past the fantastical aspects and harsh realities of 'the isle' you'll get a sock-you-in-the-eye flick that is a visual tour-de-force [...]*" (human: negative [0.4]; LLM: positive).

Combining statement. This sub-category captures statements that include knowledge that the reviewer has gained from external sources. This knowledge is then combined with the experience gained from the movie under review, leaving those who do not have that subgroup knowledge unable to make a judgment about the quality of the movie.

One common external source used for combination is other works of those involved in the production of the given movie. On the one hand, this concerns producers: "*What might have been readily dismissed as the tiresome rant of an aging filmmaker still thumbing his nose at convention takes a surprising, subtle turn at the midway point.*" (human: negative [0.23]; LLM: positive). Similarly, actors are commonly referenced: "*When cowering and begging at the feet a scruffy Giannini, Madonna gives her best performance since Abel Ferrara had her beaten to a pulp in his Dangerous Game.*" (human: positive [0.88]; LLM: negative).

A further type commonly used is using additional meta information about the given movie, such as the production cost: "*The writer-director of this little $1.8 million charmer, which may not be cutting-edge indie filmmaking*" (human: negative [0.37]; LLM: positive).

Lastly, reviewers may also include genre-specific knowledge gained from watching related movies. Depending on the preferred genre, viewers might judge the movie differently. Especially false anticipation might lead to a negative perception, despite being a good movie overall: "*By no means a slam-dunk and sure to ultimately disappoint the action fans who will be moved to the edge of their seats by the dynamic first act, it still comes off as a touching, transcendent love story.*" (human: negative [0.27]; LLM: positive).

Referencing statement. Similar to the combining statement, reviewers using referencing statements include external knowledge into the given review. However, in contrast to combining statements, this sub-category captures statements that defer the reader away from the initially reviewed movie to another external entity. One common theme of deferring readers is to another movie by the same actors or producers: "*Go rent 'Shakes The Clown', a much funnier film with a similar theme and an equally great Robin Williams performance*" (human: negative [0.27]; LLM: positive).

Other commonly used themes refer the reader to a previous version of the same movie: "*If you want a movie time trip, the 1960 version is a far smoother ride*" (human: negative [0.31]; LLM: positive). Some reviewers also reference external media, usually in which the same storyline has been presented: "*The video game is a lot more fun than the film*" (human: negative [0.36]; LLM: positive).

Lastly, reviewers suggested doing something else altogether, in an effort to save valuable time that otherwise would be lost: "*Every sequel you skip will be two hours gained. Consider this review life-affirming*" (human: negative [0.19]; LLM: positive).

Temporal statement. This sub-category captures statements that include some temporal element. Some have explicitly mentioned that the actual duration of the movie could have been shortened: "*A ravishing consciousness-raiser if a bit draggy at times*" (human: negative [0.41]; LLM: positive). Similarly, several reviews pointed out that the perceived duration of the movie felt different than its actual one: "*Love that feels significantly longer than its relatively scant 97 minutes*" (human: negative [0.41]; LLM: positive).

Reviewers have noted that a movie might keep one occupied enough to forget the surroundings, thus reflecting back on the external environment: *"The only possible complaint you could have about Spirited Away*

is that there is no rest period, no timeout" (human: positive [0.87]; LLM: negative).

4.2. Linguistical statements

Linguistic statements refer to the usage of stylistic elements that impact the sentiment of the movie review. A total of six linguistic sub-categories, namely sarcastic, coping, double-negated, contradicting, comparing, and other, were identified, representing 282 codes in total.

Table 2: Numerical Overview of Linguistic Statements

Sub-Category	Total	Type 1	Type 2
Sarcasm	30	15	15
Coping	25	5	20
Double-Negation	6	4	2
Contradiction	72	38	34
Comparison	46	6	40
Other	103	41	62

Sarcastic statement. This sub-category covers sarcastic statements or idioms. Frequently, the LLM appears not to be able to pick up on the sarcastic meaning of the statement: "*This one aims for the toilet and scores a direct hit*" (human: negative [0.09]; LLM: positive). Yet, from time to time, human annotators also appear to misjudge sarcasm: "*The only pain you'll feel as the credits roll is your stomach grumbling for some tasty grub.*" (human: negative [0.39]; LLM: positive). In some cases, the LLM appears to hallucinate by incorrectly assuming that a common phrase is used, which would typically convey either a positive or negative sentiment: "*Going to win any academy awards*" (human: positive [0.81]; LLM: negative).

Coping statement. Several statements refer to individual coping mechanisms of reviewers. Given that coping is an inherently personal process, the statements may be judged ambiguously. This is particularly the case for movies discussing disturbing topics or displaying darker scenes: "*Creates some effective moments of discomfort for character and viewer alike.*" (human: negative [0.31]; LLM: positive). In several cases, coping extends far beyond the duration of the movie itself: "*Ends with scenes so terrifying I'm still stunned. And I've decided to leave a light on every night from now on*" (human: positive [0.81]; LLM: negative).

Double-negated statement. While only a few statements containing double-negations have been identified, the usage appears to result in different judgments: "*Never seems hopelessly juvenile*" (human: negative [0.41]; LLM: positive).

Contradicting statement. Several reviewers used rhetorical devices such as oxymora to combine two words or phrases that have contradictory positive and negative connotations. When assessing the statement, one needs to make a decision on whether the newly combined phrase carries a more strongly positive or negative sentiment. Two such examples include "*Is an undeniably worthy and devastating experience*" (human: negative [0.41]; LLM: positive) and "*You'll gasp appalled and laugh outraged*" (human: positive [0.8]; LLM: negative).

Comparing statement. Several reviewers used comparisons to describe the movie. A common figurative device used was similes: "*Like life on an island*" (human: negative [0.37]; LLM: positive). Another commonly used technique is the usage of metaphors: *Splashed with bloody beauty as vivid*" (human: positive [0.8]; LLM: negative).

Other linguistic statement. A total of 103 statements with clear usage of linguistic elements, rather than contextual ones, could not be grouped any further. These are commonly of rather short nature, leaving room for speculation when judging the sentiment such as *"Laughing at his own joke*" (human: negative [0.39]; LLM: positive) and *"Is an earnest study in dispair"* (human: positive [0.81]; LLM: negative). It further captures primarily descriptive statements, e.g., referring to certain scenes or the storytelling: *"That led to their notorious rise to infamy"* (human: positive: [0.8]; LLM: negative). They are therefore considered as "other".

5. Discussion

5.1. Divergence interpretation

We find that contextual statements, specifically conditional, combination, reference, and temporal ones, as well as linguistic statements, consisting of various stylistic elements, lead to divergent sentiment judgments between human annotators and LLMs. Linguistic theory helps us to better position our findings. In particular, we find divergence occurring across different levels of analysis: The meso- and micro-level. We consider established theories on both levels in the following.

On the meso-level, the Rhetorical Structure Theory (RST) provides a systematic way to describe the structure of natural text (Mann & Thompson, 1987). It has readily been applied in sentiment analysis contexts, for instance, to improve its performance on document level (Märkle-Huß et al., 2017). The first step in using the RST is to divide a given statement into arbitrary units of interest. These units are then connected through an extensive list of relations. We find that some relations are linked to our findings, including "conditions" (conditional statements), "background"

and ”elaboration” (combining statements), and ”contrasts” (contradicting statements). We find that our coding scheme relates to the RST in the sense that contextual statements are statements that typically consist of several rhetoric units, whereas linguistic statements merely consist of one or two.

Consider the following contextual statement, which we can divide into two units of interest: *”This one* ***aims for the toilet*** *(unit 1) and* ***scores a direct hit*** *(unit 2)”*. We would assume that both of these units form a ”sequence”, with unit 2 reinforcing unit 1 by ”scoring a direct hit”. Assuming only the rhetorical structure lens, the statement therefore displays a positive sentiment. Only when considering the wider context of the movie does it receive its negative sentiment, which the human annotators correctly identified.

Next, we consider a linguistic statement, which only consists of one unit of interest: ”*Going to win any academy awards*”. According to the RST, this statement makes use of an ”evaluation” relation; however, it lacks the necessary preceding unit. By presenting humans and LLMs with statements that deviate from the usual sentence structure, judgments diverge.

While RST allows us to identify whether statements satisfy rhetoric structures, it does not allow us to make normative statements on the divergent judgments. Therefore, we draw on Roseman’s appraisal-emotion framework to understand sentiment at the micro-level. The framework offers a theoretical basis to understand how variations in contextual interpretation trigger different emotional appraisals and, in turn, divergent sentiment assessments. The framework identifies cognitive factors that shape emotional responses (i.e., whether people interpret an event as positive or negative) (Roseman, 1996). These factors include, for instance, situational state (whether an event supports or hinders personal goals), motivational state (whether the motive is appetitive-seeking reward, or aversive- avoiding harm), agency (who caused the event), unexpectedness (whether the event was anticipated), probability (certainty of outcomes), and control (perceived ability to influence outcomes).

When considering a conditional statement such as “*Only young audiences will enjoy this sloppy slapstick comedy*”, we can analyze it through the appraisal factors of situational state, motivational state, and control. Assuming the annotator is an adult, this reflects motive inconsistency, where the benefit (enjoyment) is not intended for them but for younger viewers. As a result, the adult cannot satisfy their appetitive motive (seeking pleasure). Watching a movie is also a low-control situation. The viewer cannot change their age or the film. According to Roseman, these conditions increase the likelihood of a sadness appraisal. This explains why humans, especially adult annotators, may tend to assign negative sentiment, while the LLM focuses on the positive verb *enjoy* and gives a positive judgment.

In the case of referencing statement such as *“Go rent ’Shakes the Clown’, a much funnier film with a similar theme and an equally great Robin Williams performance”*, three Roseman appraisal factors may explain why human readers judge the sentiment as negative. The *situational state* turns unpleasant, as the current film is framed as failing to meet the appetitive goal of entertainment. The agency to blame lies with an external party (i.e., likely the filmmakers), who are held responsible for the shortfall. Control potential is high, as the suggestion to “rent” another title offers the viewer a clear way to resolve the failure. According to Roseman’s matrix, the combination of motive inconsistency, external agency, and high control typically elicits anger. Humans, therefore, assign a negative valence, while LLMs focus on words like *funnier* and *great*, judging them as positive.

To sum up, RST and Roseman’s model help us analyze divergence at meso- and micro-level. While RST allows us to make claims about divergence in contextual statements due to unexpected structural configuration of the reviews (e.g., missing units of sentences), Roseman’s framework helps explain the divergence in contextual sentiment judgments between humans and LLMs. In particular, the theory shows that emotion arises from a configuration of appraisals (e.g., situational, motivational, and control factors rather than a single cue). Humans intuitively assemble these patterns and form implicit emotional judgments. In contrast, LLMs appear to focus primarily on the situational layer (i.e., whether a goal is achieved) and rely heavily on surface-level language cues.

5.2. Theoretical contributions

We contribute to IS literature at the intersection of LLMs and sentiment analysis by offering insights into why LLMs and humans may interpret sentiment differently, laying the groundwork for future theorizing. Specifically, we propose treating contextual and linguistic statements as constructs for theorizing. This brings two theoretical implications.

First, our findings address an underexplored area in the literature: the role of context in shaping differences between LLM and human sentiment judgments (i.e., where, when, and for whom a statement is made). This finding may then shift the sentiment analysis debate away from deciding whether humans or models are “right” or “wrong.” Instead, they highlight that

both may be correct, depending on context and point of view. In doing so, we reveal potential boundary conditions that can inform and deepen the theorizing process in sentiment analysis. This finding is in contrast with mainstream literature, which primarily focuses on linguistic structure (e.g., Barnes et al., 2019; Bojić et al., 2025). Instead, our findings align with Ochieng et al. (2024), who also emphasize the importance of contextual and linguistic factors. However, we go further by unpacking distinct subcategories within each factor. These subcategories can serve as formative indicators in future behavioral-quantitative studies aiming to operationalize and measure these constructs.

Second, we contribute to this mainstream literature by offering additional empirical evidence that language structure remains a critical issue in LLM-based sentiment analysis. This finding is thus consistent with previous research that finds LLM-based sentiment analysis, indeed, still struggles with sarcasm (e.g., Simanjuntak et al., 2024). From a practical perspective, these insights lay the groundwork for developing targeted solutions, such as fine-tuning approaches, to improve LLM-based sentiment analysis.

Table 3: Assessment of LLM Sentiment Analysis

Statement Type	**Total**	**Type 1**	**Type 2**
Contextual Statement	92/175 (0.526)	82/118 (0.695)	10/57 (0.175)
Linguistic Statement	111/282 (0.394)	79/109 (0.725)	32/173 (0.185)
Total	203/457 (0.444)	161/227 (0.709)	42/230 (0.183)

Aside from the two theoretical implications, our findings may open a discussion around the potential of a human-in-the-loop approach as a promising pathway for sentiment analysis, in line with recent research (Gu et al., 2025). Table 3 shows the percentage to which the two researchers who qualitatively reviewed the statements align with the LLM in the sentiment judgment. Despite not claiming representativeness, it allows us to infer some insights on the nature of divergence and its position in the wider debate of LLM-enabled sentiment analysis.

We find that for a striking majority of Type 1 statements, the LLM judgment is closer to our assessment (0.709), whereas for Type 2 statements, it is the opposite (0.183). Given that we considered a subset of roughly the same number of Type 1 and Type 2 statements, both the human and the LLM judgments are aligned with ours in roughly half of the cases, respectively (human: 0.556; LLM: 0.444). Therefore, we conclude that the divergence between humans and LLMs enables us to identify relevant borderline statements, in which both LLMs and human annotators are able to spot relevant nuances that predict positive sentiment, albeit inconsistently.

5.3. Limitations and further research

Given the widespread use of the SST dataset, GPT-4o was likely exposed to reviewer sentiment outcomes during training, introducing potential overfitting risks, even though we instructed the model to rely solely on the provided sentences. This limitation is justifiable, as SST remains one of the most widely used benchmark datasets in sentiment analysis. Importantly, we still observe clear differences between LLM and human sentiment judgments, suggesting that the potential overfitting risk did not undermine our analysis. This is likely due to the inherent variability in LLM outputs. Still, future research should address this concern by developing new datasets to more rigorously assess human–LLM discrepancies. Our findings should therefore be interpreted within the context of a widely used benchmark dataset, and can serve as a foundation for future comparisons using entirely new datasets.

In this study, we focused on assessing the ability of LLMs to judge sentiment based on a dataset with usually one sentence each by implicitly assuming that LLMs are valid proxies for sentiment analysis. While linguistic theories on the meso- and micro-level of analysis allowed us to explain the identified divergence, the scope of our research did not allow us to challenge the reasoning of LLMs on the macro-level. Future research is suggested to consider this level (e.g., by using critical discourse analysis), to challenge the broader role of LLM's reasoning in sentiment analysis.

On another note, while the LLM outperforms human annotators across Type 1 and 2 in assessing contextual statements, such as conditional (0.653) and combining (0.544), it performs worse in linguistic statements, such as coping (0.24) and comparing (0.239). On a broader level, our findings suggest that human-in-the-loop approaches, whereby both LLMs and humans assist in determining the sentiment, promise the best results. Our study thus provides opportunities for future research to move beyond qualitative coding, e.g., by training a classifier, to increase the quality of sentiment analysis by building on individual strengths of both LLM- and human-based judgments.

6. Conclusion

This research investigates the root causes of divergence between LLM-based and human sentiment

judgments of a movie review dataset. We find two underlying root causes: contextual statements and situational statements.

Contextual statements arise from differing interpretations of situational factors and their relative meaning. Contextual statements have four sub-categories. First, conditional statements trigger sentiment outcomes based on specific conditions (e.g., age, ideology, or mood). Interpretation shifts depending on whether the viewer aligns with or falls outside that condition. Second, combining statements links on-screen impressions with external facts (e.g., budget or franchise history), so interpretation relies on external knowledge. Third, referencing statements redirect the annotator's focus to another work, making the sentiment dependent on implicit comparison rather than the current film. Fourth, temporal statements anchor sentiment in time-related cues, such as duration. We show that meso-level linguistic theories, such as RST, can help us interpret the divergence.

Linguistic statements emerge due to the complexity of language constructions. We identify five sub-categories. Sarcastic statements reverse the literal meaning or use figurative language. Coping statements express personal feelings (e.g., fear), making sentiment dependent on the annotator's own experience. Double negation uses two negative markers in the same clause. Contradictory statements combine opposing positive and negative cues within the same phrase. Comparative statements express sentiment through comparison with other works. We show that micro-level linguistic theories, such as Roseman's framework, can help us interpret the divergence.

Our study thus highlights the importance of both context (where, when, and for whom a statement is made) and linguistic form (how it is phrased) in shaping sentiment interpretation. This may prompt a broader discourse challenging the implicit assumption of LLM-based sentiment analysis as inherently superior. Moreover, human-in-the-loop approaches may outperform sentiment interpretations based solely on LLMs.

Acknowledgements: This research was funded in part by the Luxembourg National Research Fund (FNR) and PayPal, PEARL grant reference 13342933/Gilbert Fridgen. For the purpose of open access, and in fulfillment of the obligations arising from the grant agreement, the author has applied a Creative Commons Attribution 4.0 International (CC BY 4.0) license to any Author Accepted Manuscript version arising from this submission.

References

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst.*, *26*(3).

Abulaish, M., Kamal, A., & Zaki, M. J. (2020). A survey of figurative language and its computational detection in online social networks. *ACM Transactions on the Web (TWEB)*, *14*(1), 1–52.

Barnes, J., Øvrelid, L., & Velldal, E. (2019). Sentiment analysis is not solved! assessing and probing sentiment classification. In T. Linzen, G. Chrupała, Y. Belinkov, & D. Hupkes (Eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 12–23).

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., & Testoni, A. (2025). LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks.

Berglund, L., Tong, M., Kaufmann, M., & Balesni, M. (2024). The reversal curse: LLMs trained on "A is B" fail to learn "B is A".

Bojić, L., Zagovora, O., Zelenkauskaite, A., Vuković, V., Čabarkapa, M., Veseljević Jerković, S., & Jovančević, A. (2025). Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific Reports*, *15*(1), 1–16.

Bu, K., Liu, Y., & Ju, X. (2024). Efficient utilization of pre-trained models: A review of sentiment analysis via prompt learning. *Knowledge-Based Systems*, *283*, 1–18.

Gautam, H., Gaur, A., & Yadav, D. K. (2025). A survey on the impact of pre-trained language models in sentiment classification task. *International Journal of Data Science and Analytics*, 1–39.

Gu, F., Li, Z., Colon, C. R., Evans, B., Mondal, I., & Boyd-Graber, J. L. (2025). Large language models are effective human annotation assistants, but not good independent annotators.

Ipa, A. S., Roy, P. N., Rony, M. A. T., Raza, A., Fitriyani, N. L., Gu, Y., & Syafrudin, M. (2024). BdsentiLLM: A novel LLM approach

to sentiment analysis of product reviews. *IEEE Access*, *12*, 189330–189343.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1746–1751).

Kour, K., Kour, J., & Singh, P. (2021). Lexicon-Based Sentiment Analysis. In G. S. Hura, A. K. Singh, & L. Siong Hoe (Eds.), *Advances in Communication and Computational Technology* (pp. 1421–1430). Springer Nature Singapore.

Li, Z., Chan, F. Y., Gao, C., & Ye, Q. (2024). *Will sentiment extraction based on ChatGPT yield better predictive outcomes? Evidence from an online travel agency*.

Mann, W. C., & Thompson, S. A. (1987). Rhetorical structure theory: Description and construction of text structures. In *Natural language generation: New results in artificial intelligence, psychology and linguistics* (pp. 85–95). Springer.

Märkle-Huß, J., Feuerriegel, S., & Prendinger, H. (2017). Improving sentiment analysis with document-level semantic relationships from rhetoric discourse structures. *Proceedings of the 50th Hawaii International Conference on System Sciences*, 1142–1151.

Mayring, P. (2014). Qualitative content analysis: Theoretical foundation, basic procedures and software solution.

Mukta, M. S. H., & Islam, N. (2025). *LLM-driven sentiment analysis for predicting customer insights in enterprise systems: A computational design science approach*.

Ochieng, M., Gumma, V., Sitaram, S., Wang, J., Chaudhary, V., Ronen, K., Bali, K., & O'Neill, J. (2024). Beyond metrics: Evaluating LLMs effectiveness in culturally nuanced, low-resource real-world scenarios.

OpenAI. (2024). GPT-4 Technical Report.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 115–124.

Research and Markets. (2025). Sentiment analytics - global strategic business report. https://www.researchandmarkets . com / report / sentiment - analysis?

Roseman, I. J. (1996). Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognition & Emotion*, *10*(3), 241–278.

Simanjuntak, V. O. F., Velni, F., & Wiputra, R. (2024). Prompt design for monetary policy: An LLM-based sentiment analysis for macroeconomic decision-making. *Proceedings of the Australasian Conference on Information Systems (ACIS 2024)*.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642.

Sun, X., Zhang, K., Liu, Q., Bao, M., & Chen, Y. (2024). Harnessing domain insights: A prompt knowledge tuning method for aspect-based sentiment analysis. *Knowledge-Based Systems*, *298*, 1–17.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *Proceedings of the 30th Conference on Pattern Languages of Programs*.

Xing, F. (2025). Designing heterogeneous LLM agents for financial sentiment analysis. *ACM Trans. Manage. Inf. Syst.*, *16*(1).

Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2022). QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering.

Zhang, W., Deng, Y., Liu, B., Pan, S., & Bing, L. (2024). Sentiment analysis in the era of large language models: A reality check. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3881–3906).

Zhuang, L., Wayne, L., Ya, S., & Jun, Z. (2021). A robustly optimized BERT pre-training approach with post-training. In S. Li, M. Sun, Y. Liu, H. Wu, K. Liu, W. Che, S. He, & G. Rao (Eds.), *Proceedings of the 20th chinese national conference on computational linguistics* (pp. 1218–1227).