Systemic Functional Linguistics for Explainable and Sustainable Sentiment Analysis with GPT-5

Anonymous ACL submission

Abstract

Despite major advances, sentiment analysis (SA) still mostly relies on rigid polarity categories that oversimplify the complexity of human meaning-making. In addition to being theoretically reductionist, existing approaches are also opaque, increasingly costly, and environmentally unsustainable. This paper proposes to integrate linguistic theory—specifically Systemic Functional Linguistics (SFL) and the Appraisal Framework—into an LLM SA classification workflow. Using a manually annotated corpus of 8,601 Italian tweets—the *xLiMe Twitter Corpus*, we compare manual annotations and XLM-RoBERTa sentiment predictions with standard GPT-5 and SFL-enhanced GPT-5 classifications. Results show that against the manual gold standard, GPT-SFL achieves the best results, with an accuracy of 0.803 and a macro-F1 of 0.389, slightly higher than GPT-std (0.784 / 0.385). SFL integration also introduces fine-grained categories such as *neutral with humour*, *slight sadness*, and *cautionary*, which other systems collapsed into neutrality. Moreover, the method adds interpretive depth by providing explicit linguistic rationales for each assigned label. The findings demonstrate that embedding linguistic theory makes it possible to obtain human-aligned labels and explanations without costly fine-tuning offering a theory-grounded path towards more transparent, accurate, replicable, and sustainable SA.

1 Introduction

Sentiment Analysis (SA) has become one of the most widely applied techniques in natural language processing (NLP), with uses in consumer analytics, media studies, political communication, and cultural research. Research has consistently sought to refine this technique, either by softening rigid classification schemes (Munikar et al., 2019) (e.g., very positive, positive, neutral, negative, and very negative) or extending beyond polarity to (basic) emotion prediction (Kawade and Oza, 2017; Nandwani and Verma, 2021). For example, early dictionary-based methods, while foundational, were largely context-agnostic and struggled with complex forms of meaning such as irony and sarcasm (Hassan and Mahmood, 2017; Munikar et al., 2019; Meena et al., 2021, 2022, 2023). Transformer-based approaches improved accuracy by incorporating contextual embeddings, and more recently, large language models (LLMs) have demonstrated strong potential in more accurate sentiment classification. However, these computational advances never fundamentally challenge the epistemological assumption underpinning SA: that human emotions can be discretized into two or three universal categories (Viola, 2023a). As a consequence, even these models remain bound by traditional sentiment categories—typically positive, negative, and neutral—that oversimplify human affect.

Such simplifications are especially problematic when SA is applied beyond product reviews—the task for which it was originally designed (Puschmann and Powell, 2018; Viola, 2023b). When analyzing political or cultural discourse, the complexity of affective meaning cannot be adequately captured by rigid polarity categories. For example, neutral categories often serve as a "catch-all" for cases of irony, hedging, cautionary stance, or mixed affect (Koppel and Schler, 2006; Valdivia et al., 2018; Vassallo et al., 2024). At the same time, human annotators also struggle: linguistic categories are difficult to define unequivocally (Langacker, 1983; Talmy, 2000; Croft and Cruse, 2004; Dancygier and Sweetser, 2012; Gärdenfors, 2014; Paradis, 2015), and inter-annotator agreement in SA rarely exceeds 65% (Bobicev and Sokolova, 2017).

In addition, contemporary SA practices increasingly rely on computationally expensive LLMs, which raises concerns about financial costs and environmental sustainability. Training and running

large-scale transformer models has been shown to require substantial energy resources and to generate significant carbon emissions (Strubell et al., 2019; Bender et al., 2021; Patterson et al., 2021; Anthony et al., 2020; Bakhtiarifard et al., 2025). At the same time, the opacity of LLM predictions makes it difficult to interpret, replicate, or explain why a model arrives at a particular classification (Danilevsky et al., 2020), highlighting a growing tension between performance, transparency, and sustainability in NLP workflows.

This study proposes to integrate the Systemic Functional Linguistics (SFL) theory, in particular Systemic Functional Grammar (SFG) (Halliday et al., 1985) and the Appraisal Framework (Martin and White, 2005), to help overcome these limitations. The central hypothesis is that embedding SFL principles into LLM SA prompts leads to more nuanced, explainable, replicable, and interpretable sentiment classifications, while also offering a lighter-weight and more sustainable alternative to fine-tuned models. Unlike standard SA, which reduces evaluation to rigid polarity categories, and unlike emotion analysis (EA), which maps text onto universalist taxonomies of basic emotions (e.g., joy, anger, sadness), the study argues that this approach will allow us to detect linguistically grounded realizations of attitude, graduation, and engagement. The study asks: does embedding SFG and the Appraisal Framework into an LLM SA workflow enable more granular, transparent, and environmentally conscious classifications that also better reflect the complexity of meaning-making in natural language?

To answer this question, we compare manual sentiment annotations and transformer-based predictions (XLM-RoBERTa) (Barbieri et al., 2022) against both GPT-5[1] predictions without linguistic theory (here referred to as 'standard GPT-5') and GPT-5 enhanced with SFL principles (here referred to as 'SFL-enhanced GPT-5'). The dataset for the analysis is the publicly available *xLiMe Italian Twitter Corpus*[2] (Rei et al., 2016), which includes 8,601 Italian tweets. The complete annotated dataset, including SFL-enhanced labels and explanations, is released separately and described in detail in (Anonymous, 2026).

[1]All GPT-based annotations were generated using ChatGPT-5 accessed via the OpenAI API. We refer to this system as GPT-5 throughout the paper for simplicity.

[2]MIT License

2 Literature review

With its applications extending far beyond consumer review analysis, SA has evolved methodologically, rapidly progressing from lexicon-based and traditional machine learning approaches to neural architectures, transformers, and, most recently, large language models. Early models such as Naïve Bayes and SVM offered interpretability and computational efficiency but were inadequate for sarcasm and irony detection as well as contextual nuance (Sharma et al., 2025; Alahmadi et al., 2025). Deep learning approaches, including CNNs and LSTMs, later introduced hierarchical and sequential modelling that significantly improved performance on short, noisy text (Sharma et al., 2025; Albladi et al., 2025).

But it was the introduction of transformer architectures such as BERT and RoBERTa that truly marked a paradigm shift, enabling contextual embeddings through self-attention and setting new benchmarks for sentiment classification (Alahmadi et al., 2025; Sharma et al., 2025). These models have demonstrated state-of-the-art performance, especially in handling multilingualism, sarcasm, and mixed sentiments (Albladi et al., 2025) and laid the foundation for modern LLMs, including GPT variants and ChatGPT, which extend capabilities to few-shot and zero-shot tasks.

Recent studies have confirmed the superiority of LLMs over traditional SA approaches, particularly in handling context-rich and longer texts. For instance, Kanth et al. (2025) demonstrate that GPT-4 consistently outperforms classical classifiers across various SA tasks, while Shah et al. (2025) show that GPT-4 and Llama-2 surpass both rule-based systems (e.g., SAFE, ReUS) and fine-tuned transformer models (e.g., RE-BERT), especially when using few-shot prompting. These findings underscore the powerful generalisation and adaptability of LLMs, even under zero-shot conditions.

Yet, this impressive performance has not addressed a longstanding theoretical gap in the field. While psychological frameworks have long informed emotion recognition models—most notably in works like Ekman's basic emotions theory (Ekman, 1992), Cowen and Keltner's mapping of emotional experience (Cowen et al., 2019), and the GoEmotions taxonomy (Demszky et al., 2020)—linguistic theory has remained conspicuously absent from computational SA. This absence is especially notable given the availability of robust

linguistic models, such as SFL, which offer principled ways of analysing interpersonal meaning, evaluative stance, and the fine-grained resources speakers use to position themselves in discourse.

At the same time, the widespread adoption of LLMs has surfaced persistent concerns about their scalability, fairness, and interpretability. As Kanth et al. and Shah et al. both note, high performance often comes at the cost of increased computational demands, reduced transparency, and ethical challenges—particularly in low-resourced languages. These trade-offs call for approaches that preserve the strengths of LLMs while enhancing their interpretability and reducing resource intensity.

This paper responds to these challenges by exploring, to the author's knowledge, the first integration of a full-fledged linguistic theory—SFL—into an LLM-based SA pipeline via prompting. Rather than relying on costly fine-tuning or opaque black-box predictions, this method leverages theory-driven instruction to guide classification in a transparent, replicable, and linguistically meaningful way. In doing so, it addresses key limitations in current SA practices: it enhances explainability, supports low-cost deployment, and bridges the gap between computational efficiency and theoretical rigour.

3 Systemic Functional Linguistics

In this study, the Systemic Functional Linguistics (SFL) (Halliday et al., 2004) theory is employed as the foundation for enhancing SA classification with GPT-5. The study's research hypothesis is that embedding SFL principles into LLM workflows allows sentiment classifications to move beyond rigid polarity labels, offering categories that are both linguistically grounded and interpretable.

SFL conceptualizes grammar as a resource for making meaning; the theory posits that speakers (or writers) create meaning by making linguistic choices (e.g., by using optional forms). SFL proposes that language simultaneously realizes three metafunctions:

- Ideational meaning, which represents human experience;
- Interpersonal meaning, which enacts social relations;
- Textual meaning, which organizes discourse in relation to context.

Treating grammar as a meaning-making resource, SFG in particular insists on the interrelation of form and meaning helping analyse systematic relations between choices and grammatical forms and more abstract strata such as situational and cultural context. The present study focuses particularly on the interpersonal metafunction, which is closely tied to evaluative meaning, that is how speakers participate in speech events (appraisal). To deepen this dimension, Martin and White's Appraisal Framework (2005) is integrated, providing analytical tools for how language evaluates, constructs social stances, and manages writer/speaker–reader/listener relationships. The Appraisal Framework comprises three interrelated systems:

- Attitude, which captures emotions and viewpoints;
- Engagement, which accounts for how attitudes are sourced and negotiated;
- Graduation, which tracks the intensity and scope of evaluations.

The RH is that integrating the Appraisal - SFG (Martin and White 2005) Framework into GPT-5 SA workflow will reveal the complex ways in which speakers express and construct interpersonal relationships, attitudes and stances towards events and entities, enriching the SA classification, for example introducing more nuanced stances. Moreover, it will introduce a layer of explainability: the classifications will be grounded in explicit linguistic theory, so that model outputs can be traced back to identifiable features of language use, such as evaluative lexis, making the annotations transparent and replicable.

In this sense, the integration of SFL into the SA classification task links model predictions to transparent linguistic principles, clarifies how interpersonal meaning is constructed in text, and reveals how speakers and writers position themselves towards events and entities. This dual role—enhancing classification performance while providing interpretability—marks the innovative contribution of the present study. At the same time, the approach is more sustainable and cost-efficient, as it leverages prompting rather than resource-intensive fine-tuning, thereby reducing computational overhead and environmental impact.

4 Methodology and dataset

The research hypothesis is tested on the Italian task of the *xLiMe Twitter Corpus* (Rei et al., 2016) which contains 8,601 tweets (162,269 tokens) manually annotated for Part of Speech, Named Entities, and Sentiment. The analysis thus compares the manually annotated labels (i.e., positive, negative, neutral) and the transformer-based machine learning annotations (i.e., positive, negative, neutral) using XLM-RoBERTa (Barbieri et al., 2022)[3] against both GPT-5 predictions without SFL, and GPT-5 predictions with SFL[4]. In this way, two baselines are employed for the comparative evaluation. First, the manual annotations from the *xLiMe Twitter Corpus* are used as the golden standard human benchmark. However, although these annotations represent the closest available approximation to "ground truth," it is important to note that the inter-annotator agreement for this dataset is reported as only slight (Cohen/Fleiss $\varkappa = 0.02$). This low level of agreement underscores the inherent difficulty of sentiment annotation and highlights the limitations of rigid polarity schemes, particularly for ambiguous or culturally nuanced data (Viola, 2023b). Including these labels as a baseline is nevertheless crucial, as it foregrounds the interpretative challenges faced even by human annotators.

Second, we adopt XLM-T (Barbieri et al., 2022) as the computational baseline. XLM-T is a domain-adapted multilingual transformer model built on XLM-R —a multilingual variant of RoBERTa (Conneau et al., 2020)—and further pre-trained on millions of tweets across more than thirty languages, including Italian. Empirical evaluations demonstrate that such Twitter-specific multilingual models consistently outperform general-domain multilingual models (such as XLM-R or mBERT). Its inclusion ensures methodological robustness by providing a strong computational, transformer-based point of comparison for GPT-5.

Taken together, these baselines allow us to evaluate the performance of GPT-5 with and without SFL integration against both human annotation practices and a state-of-the-art transformer trained on Twitter data, thereby situating our results in relation to both interpretative and computational standards. While we do not modify GPT-5's parameters, we operationalise SFL through prompting, guiding the model to classify sentiment categories and provide linguistic explanations based on systemic-functional principles.[5]

5 Analysis

We first compare the two annotation benchmarks: manual annotations from the *xLiMe Twitter Corpus* and XLM-T annotations. The results are displayed in Table 1.

Label	Human annotations	XLM-T annotations
Positive	554	2021
Negative	523	3439
Neutral	7519	3141

Table 1: Sentiment annotations of human annotators and XLM-T annotations.

The results demonstrate a marked discrepancy. Human annotators overwhelmingly assigned the neutral label (7,519 instances), with relatively few texts judged as positive (554) or negative (523). This may reflect the challenge of achieving consistent sentiment judgements across human coders. By contrast, the model produced a more dispersed distribution, skewed towards negative labels (3,439), while assigning substantially fewer neutral (3,141) and more positive (2,021) labels than the human coders. This divergence underscores the inherent difficulty of sentiment annotation, both human and automated, particularly in contexts where affective cues are subtle or ambiguous. At the same time, it raises important questions about XLM-T role as a computational "gold standard."

In line with the methodological design, we now proceed to evaluate GPT without SFL against the two established baselines —human annotations and XLM-T —situating GPT's default performance relative to both interpretative and computational stan-

[3]Although specifically fine-tuned on Italian Twitter data, the study does not use FEEL-IT (Bianchi et al., 2021). This is because FEEL-IT employs a binary label scheme (positive/negative) that is not directly comparable with the three-way manual annotation scheme (positive/negative/neutral) used by the *xLiMe Twitter Corpus* annotators.

[4]The full prompt is provided in 10

[5]The experiments with GPT-5 and GPT-5-SFL were carried out via the OpenAI API. Since GPT-5 is a proprietary model, the exact number of parameters and training infrastructure are not publicly disclosed. By contrast, the XLM-RoBERTa baseline we used contains approximately 270M parameters. Our annotation workflow required only inference calls, with a total of $\approx 3.4M$ tokens processed (8,601 tweets $\times$ 2 annotation settings $\times \approx$ 200 tokens average). This corresponds to a modest computational budget compared to model training or fine-tuning, and no dedicated GPUs were used on our side beyond a standard workstation for orchestration and data analysis.

dards. We then extend the comparison to include the SFL-enhanced version, thereby assessing the added value of systemic-functional cues for sentiment classification. The results are reported in Table 2.

Label	Human annotations	XLM-T annotations	GPT (std)	GPT with SFL
Positive	554	2021	647	551
Negative	523	3439	504	426
Neutral	7519	3141	7450	7598
Neutral with humour	–	–	–	253
Slight sadness	–	–	–	121
Cautionary	–	–	–	21

Table 2: Distribution of sentiment annotations across human annotators, XLM-T, GPT without SFL, and GPT with SFL.

When extending the comparison to include GPT-std further discrepancies emerge. XLM-T distributed labels more broadly than human annotators; GPT without SFL, by contrast, mirrors the human distribution more closely, assigning 7,450 instances to neutral and only 647 and 504 to positive and negative, respectively. While this behaviour aligns more closely with human annotation practices than with the computational baseline, it may also highlight the model's reluctance to over-polarise subtle evaluative meanings. XLM-T's inclination to over-assign negative sentiment may reflect a tendency to sharpen polarity distinctions which may also exaggerate negativity in contexts where human annotators perceive neutrality. Thus, the contrast illustrates that the computational standard itself is not a stable reference point but rather encodes a particular distributional logic that diverges from human interpretative practices. In this sense, evaluating GPT against both human and XLM-T baselines reveals the challenges of achieving cross-system consistency as well as the contested nature of what constitutes "ground truth" in SA.

By contrast, the introduction of the SFL-enhanced GPT output enriches the sentiment classification landscape by demonstrating how systemic-functional categories capture evaluative nuances beyond rigid polarity schemes. In addition to the three core labels, the model identifies categories such as "cautionary" (21), "neutral with humour" (253), and "slight sadness" (121). While the overall polarity balance remains broadly consistent with human annotation tendencies, the emergence of these additional categories illustrates how appraisal distinctions surface layers of meaning that would otherwise be collapsed into neutrality. This suggests that disagreements partly reflect the richer appraisal distinctions offered by SFL, rather than annotation noise.

This refinement has methodological significance: by introducing intermediate and nuanced categories, the SFL framework highlights interpretative subtleties—such as irony, caution, or mild affect—that a coarse three-way scheme cannot accommodate. Importantly, this provides a plausible explanation for the very low inter-annotator agreement reported in the corpus ($\varkappa$ = 0.02). Rather than mere inconsistency between annotators, many disagreements may reflect the inadequacy of traditional polarity categories to capture the full range of evaluative meanings embedded in the texts. This extension suggests that incorporating linguistic theories of meaning does not radically alter the overall polarity balance, but it does foreground interpretative subtleties—such as cautionary stances—that are otherwise flattened into neutrality.

5.1 Agreement, divergence, and performance

We now compute Cohen's $\varkappa$ between GPT-5 standard, SFL-enhanced GPT-5 (both base-3 and extended labels), manual annotations, and XLM-T (Table 3). While agreement with manual gold remains low across all models ($\kappa \approx 0.07$), raw agreement is much higher: both GPT models match manual labels in around 78–80% of cases. The apparent gap between raw agreement and $\varkappa$ reflects the fact that $\varkappa$ corrects for chance, which is inflated when one category (here, neutral) dominates the distribution.

SFL-enhanced GPT-5 diverges from manual annotations in two systematic ways. First, it often reclassifies manual positives and negatives as neutral, reflecting a stricter treatment of evaluative cues when prompted with explicit linguistic principles. Second, it introduces intermediate categories such as *neutral with humour* or *slight sadness*, redistributing a portion of manual neutrals. These distinctions suggest that what annotators labelled as "neutral" often conceals weak evaluative meaning.

Comparisons with XLM-T show even lower $\varkappa$ (0.040–0.048) and raw agreement of only 38–39%. Interestingly, $\varkappa$ is slightly higher when SFL extended categories are retained (0.048), even though this increases visible disagreement. This effect arises from $\varkappa$'s sensitivity to prevalence: distributing predictions across more categories lowers expected chance agreement, sometimes raising $\varkappa$ de-

spite lower overlap.

These results underscore two points. First, $\varkappa$ evidences the inadequacy of three-way polarity schemes to capture evaluative nuance. Second, theory-informed prompting surfaces distinctions that explain apparent disagreement—humour, mild affect, or cautionary stance—that would otherwise remain hidden under a single neutral label, as shown in Table 4. For example, $\approx 77\%$ of the category "neutral with humour", $\approx 86\%$ of "slight sadness", and $\approx 76\%$ of "cautionary" were manual neutrals. Thus, the apparent disagreement between humans and models may not reflect unreliability but rather that the manual neutral category was absorbing a lot of subtle evaluative content (irony, humour, mild affect, hedging), as reported in the literature. SFL-enhanced GPT-5 recognizes nuances that the 3-way manual scheme flattens.

Table 3: Cohen's κ between manual, XLM-T, GPT-5 standard, and GPT-5 SFL.

Comparison	Scheme	κ
Manual vs. GPT-SFL	Extended	0.077
Manual vs. GPT-SFL	Base 3	0.075
Manual vs. GPT-std	Base 3	0.076
XLM-T vs. GPT-std	Base 3	0.040
XLM-T vs. GPT-SFL	Extended	0.048
XLM-T vs. GPT-SFL	Base 3	0.043

Table 4: New GPT-SFL categories across manual labels.

SFL cat.	Neg	Neu	Pos
Cautionary	3	16	2
Neutral humour	13	196	44
Slight sadness	12	104	5

SFL-enhanced GPT also diverges from GPT-std in 716 cases. The main point of disagreement lies in SFL-enhanced GPT-5 redistribution of instances that GPT-std classified as "neutral" into "neutral with humour" (148), "slight sadness" (111), or even into "negative" and "positive" categories (Table 5). But SFL-enhanced GPT-5 also redistributes some polarity shifts: e.g., 141 cases GPT-std labelled "negative" became "neutral" under SFL-enhanced GPT-5, showing that the SFL framework introduces interpretive nuance.

Finally, the disagreement rate with XLM is much higher (5,378 cases vs. 716 for GPT-std). As reported in Table 6, 2,975 cases labelled "negative" by XLM were "neutral" in SFL-enhanced GPT-5, 1,545 cases labelled "positive" by XLM were "neutral" in SFL-enhanced GPT-5. This would again suggest that XLM is overly polarizing, while SFL-enhanced GPT-5 identifies milder evaluative stances.

Table 5: Disagreements between GPT-SFL and GPT-std.

SFL cat.	Neg (std)	Neu (std)	Pos (std)
Cautionary	1	20	0
Negative	0	103	5
Neutral	141	0	72
Neutral humour	4	148	101
Slight sadness	4	111	6

Table 6: Disagreements between GPT-SFL and XLM.

SFL cat.	Neg (XLM)	Neu (XLM)	Pos (XLM)
Cautionary	14	4	3
Negative	0	162	83
Neutral	2975	0	1545
Neutral humour	57	76	120
Positive	111	107	0
Slight sadness	65	31	25

In addition to $\varkappa$, we report overall performance of both GPT-5 systems against the two benchmarks (Table7). For GPT-SFL we provide both an extended evaluation (retaining nuanced categories) and a collapsed-to-base-3 evaluation (mapping nuanced categories to neutral) for comparability. Bootstrap confidence intervals (95%) are reported for macro-F1. A complete overview of per class metrics is provided in 10.

Table 7: Overall performance (Accuracy, Macro-F1) of GPT-std and GPT-SFL against manual gold and XLM-T.

	vs Manual		**vs XLM-T**	
System	**Acc**	**M-F1**	**Acc**	**M-F1**
GPT-std	0.784	0.385	0.383	0.289
GPT-SFL→base3	**0.803**	**0.389**	**0.388**	0.280
GPT-SFL (ext.)	0.767	0.191	–	–

Against the manual gold standard, GPT-SFL (collapsed to three categories) achieves the best results, with an accuracy of 0.803 and a macro-F1 of 0.389, slightly higher than GPT-std (0.784 / 0.385). These gains, though modest, indicate that prompting GPT with SFL-informed categories not only preserves but slightly improves performance compared to a standard polarity-based prompt. When the extended SFL categories are retained, performance drops to 0.767 accuracy and 0.191 macro-F1. This apparent decline is expected, since manual annotations collapse such distinctions into neutral, as already discussed. Rather than signalling model

failure, this mismatch highlights a key limitation of conventional evaluation: many tweets labelled neutral by human annotators in fact encode weak evaluative stances that the SFL framework is able to identify.

Comparisons against XLM-T further underscore these interpretive differences. Agreement between GPT models and XLM-T is low ($accuracy \approx 0.38$, $macro-F1 \approx 0.28$). Both GPT variants therefore diverge strongly from XLM, with GPT-SFL showing slightly better alignment than GPT-std. This indicates once more that whereby XLM-T polarises aggressively, GPT-SFL defaults to neutral unless evaluative lexis is clearly marked.

5.2 Explainability

Beyond distributional comparisons, the SFL-enhanced GPT system contributes a key methodological advantage: explainability. Unlike both human annotations, which provide only categorical judgements, and transformer-based baselines such as XLM-T, which at the very best output a confidence score, the SFL-enhanced GPT model generates explicit rationales grounded in the Appraisal framework. Each sentiment label is accompanied by a short explanation that cites the linguistic cues driving the classification. These cues are organized across the three appraisal dimensions: Attitude (affective, ethical, or aesthetic evaluation), Engagement (stance-taking through irony, rhetorical questions, or intersubjective positioning), and Graduation (scaling of intensity through intensifiers or mitigators).

For example, tweets labelled "negative" are often explained through the presence of evaluative lexis such as *vergognarsi* ('to be ashamed'), while tweets categorized as cautionary highlight directive forms like *mi raccomando* ('make sure to…'). Neutral classifications, in contrast, tend to be justified by the absence of evaluative or intensified expressions. By making these interpretative criteria explicit, the SFL-enhanced GPT increases the transparency of computational SA while foregrounding the linguistic mechanisms that underlie evaluative meaning.

Explainability has two important implications. First, it offers a bridge between computational predictions and human interpretive practices, allowing researchers to understand why the system assigns a given label. Second, it provides a potential resolution to the low inter-annotator agreement observed in the dataset. Many apparent "disagreements" between annotators may in fact reflect interpretative differences that are invisible within rigid polarity categories but become intelligible when the classification system is based on a solid linguistic theory. In this sense, explainability is not merely an add-on but a methodological tool for diagnosing annotation inconsistencies and rethinking what constitutes reliable sentiment "ground truth." A sample of the explainability output[6] is provided in the Appendix.

6 Qualitative analysis

To further illustrate how SFL-enhanced GPT-5 enriches SA beyond coarse polarity assignments, we present a set of qualitative examples (1-3 below). Table 8 reports the annotations across the four methods, including SFL-enhanced GPT-5 explanation.

1. *io avevo l'ansia prima di comprare i biglietti immaginati ora ahah*
   I was anxious before buying the tickets, imagine me now ahah
2. *ora piango ciao ciao): ok no la smetto, ti amo anche io ... *fa il labbuccio* buonanotte *gli stampa un bacio sulle labbra**
   now I cry bye bye): ok I'll stop, I love you too … *pouts* goodnight *kisses on the lips*
3. *Mi raccomando, il prossimo 10 Agosto durante la notte di San Lorenzo, portatevi un caschetto, che potreste essere investiti dai meteoriti.*
   Make sure, next 10 August during the night of San Lorenzo, to wear a helmet, as you might be hit by meteorites.

In Example 1, humour cues ('ahah') are ignored by GPT-std and XLM, which opt for neutral and positive respectively, while SFL-enhanced GPT-5 interprets them as markers of stance mitigation, assigning neutral with humour, showing affective softening that was otherwise flattened. Example 2 reveals a different dynamic: although both XLM and GPT-std classify the text as positive likely due to affectionate language, SFL-enhanced GPT-5 highlights co-occurring sadness ('piango,):') and assigns slight sadness. This demonstrates the model's ability to integrate affective contradictions within the utterance, rather than privileging one

[6] To assess the reliability of the SFL-based annotations and explanations, two independent validators manually validated a stratified sample of 600 tweets (≈7% of the corpus), ensuring coverage of all categories including minority labels. We found that while the majority of labels were consistent, 5–10% of cases contained errors, either in the label assignment or in the linguistic explanation. The validation results are released alongside the published dataset (Anonymous, 2025).

Table 8: Summary of qualitative examples (cf. Examples 1–3).

Ex.	Manual	XLM	GPT-std	GPT-SFL	Explanation
1	Positive	Positive	Neutral	Neu. humour	Marker *ahah* signals irony/mitigation; stance softened.
2	Neutral	Positive	Positive	Sl. sadness	*Piango* ('I cry') + emoticon *):* convey mild sadness with affection.
3	Positive	Negative	Neutral	Cautionary	Directive *mi raccomando* ('make sure') + conditional cue caution.

polarity or flattening into neutral. Finally, Example 3 illustrates how SFL-enhanced GPT-5 recognizes cautionary stance. While manual, XLM, and GPT-std scatter across positive, negative, and neutral labels, SFL-enhanced GPT-5 identifies directive lexis ('mi raccomando') and conditional warning as markers of a cautionary evaluation. Here, the model detects pragmatic force rather than literal polarity, revealing the limits of rigid sentiment categories.

These cases demonstrate how SFL integration into GPT does not merely reassign polarity but captures evaluative nuance that is otherwise flattened by polarity-driven approaches thereby providing a more interpretable explanation for apparent disagreement across models and annotators.

7 Discussion

The results confirm that much of what annotators treated as "neutral" actually encodes either weak evaluative stances otherwise obscured or contradicting sentiments within the same text. Compared to GPT-std, SFL-enhanced GPT-5 diverges in only 716 cases, primarily by reallocating neutral predictions into nuanced categories and softening strong polar classifications. This indicates that the SFL framework introduces interpretive refinement rather than wholesale reclassification, maintaining overall polarity balance while adding layers of meaning. By contrast, disagreements with XLM are far more frequent (5,378 cases), reflecting XLM's systematic tendency to over-polarize texts that SFL-enhanced GPT-5 interprets as neutral or subtly evaluative. This suggests that polarity-driven architectures risk overstating sentiment in contexts where evaluative stance is more ambivalent.

On the whole, these findings have four main methodological implications. First, they demonstrate that, beyond the inherent impossibility of unequivocally define linguistic categories, the persistent problem of low inter-annotator agreement in SA may reflect the limitations of rigid polarity schemes. Second, they show that integrating SFL principles into LLM-based sentiment classification provides an explainability framework that accounts both for systematic divergences and the typical lack of transparency in understanding how the categories are assigned. Third, the results highlight the impact of linguistic theory in computational modelling: even at the level of prompt, SFL-enhanced GPT-5 enriches the analysis without destabilizing the overall polarity distribution, thereby offering a scalable, interpretable, theoretically-grounded method that bridges qualitative linguistic insights with large-scale automated classification. Finally, the study carries methodological implications for the sustainable use of LLMs in . By relying on prompting rather than resource-intensive fine-tuning, the SFL-enhanced workflow demonstrates that it is possible to obtain nuanced and human-aligned annotations without incurring the high computational and environmental costs typically associated with model retraining.

8 Conclusion

This study addressed a long-standing limitation in SA: the reliance on rigid polarity categories that oversimplify the complexity of evaluative meaning in natural language. It proposed to integrate Systemic Functional linguistic theory—specifically Systemic Functional Grammar and the Appraisal Framework—into an LLM SA workflow. Using a manually annotated corpus of Italian tweets, we compared manual annotations and XLM-RoBERTa predictions against GPT-5 prompted with and without SFL principles.

The results produced more nuanced and transparent classifications, added interpretive depth, and provided explicit linguistic rationales, thereby enhancing transparency and explainability grounded in established linguistic frameworks. This approach offers a path towards greener and more cost-efficient SA, showing that linguistic theory can also serve as a strategy for reducing the carbon footprint of annotation and classification tasks.

9 Limitations

While this study provides valuable insights, several limitations must be acknowledged. The study focused on a single corpus of Italian tweets. Future studies could apply SFL principles on other languages and genres to enhance generalisability across languages and genres, such as political communication, advertisement, etc. The integration of SFG categories was operationalised through prompting rather than fine-tuning, and further work is needed to explore hybrid architectures or training strategies that embed linguistic theory more deeply. Moreover, while explainability was improved, evaluation still relied on inherently limited annotation schemes, suggesting that future work should also revisit annotation guidelines in light of linguistic theory. Finally, the study only tested the effectiveness of SFL on GPT-5. Further research should run a comparison across open models.

Despite the study's limitations, this paper has shown that incorporating SFL into LLM-based SA offers a principled way to move beyond polarity-driven reductionism, bridging qualitative linguistic insights with computational scalability. In doing so, SA can become not only more accurate, but also more transparent, theoretically grounded, replicable, and environmentally less costly.

References

Khaled Alahmadi, Sultan Alharbi, Juan Chen, and Xianzhi Wang. 2025. Generalizing sentiment analysis: a review of progress, challenges, and emerging directions. *Social Network Analysis and Mining*, 15(1):45.

Aish Albladi, Minarul Islam, and Cheryl Seals. 2025. Sentiment Analysis of Twitter Data Using NLP Models: A Comprehensive Review. *IEEE Access*, 13:30444–30468.

Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. *arXiv preprint*. Version Number: 1.

Pedram Bakhtiarifard, Pınar Tözün, Christian Igel, and Raghavendra Selvan. 2025. Climate And Resource Awareness is Imperative to Achieving Sustainable AI (and Preventing a Global AI Arms Race). *arXiv preprint*. ArXiv:2502.20016 [cs].

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2021. FEEL-IT: Emotion and Sentiment Classification for the Italian Language. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 76–83, Online. Association for Computational Linguistics.

Victoria Bobicev and Marina Sokolova. 2017. Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective. In *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, pages 97–102. Incoma Ltd. Shoumen, Bulgaria.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alan S. Cowen, Hillary Anger Elfenbein, Petri Laukka, and Dacher Keltner. 2019. Mapping 24 emotions conveyed by brief human vocalization. *The American Psychologist*, 74(6):698–712.

William Croft and D. Alan Cruse. 2004. *Cognitive Linguistics*, 1 edition. Cambridge University Press.

Barbara Dancygier and Eve Sweetser, editors. 2012. *Viewpoint in Language: A Multimodal Perspective*. Cambridge University Press, Cambridge.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Peter Gärdenfors. 2014. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press, Cambridge, MA, USA.

M. A. K. Halliday, C. M. I. M. Matthiessen, M. Halliday, and C. Matthiessen. 1985. *An Introduction to Functional Grammar*. Taylor & Francis.

M. A. K. Halliday, Christian M. I. M. Matthiessen, Michael Halliday, and Christian Matthiessen. 2004. *An Introduction to Functional Grammar*, 3 edition. Routledge, London.

Abdalraouf Hassan and Ausif Mahmood. 2017. Deep Learning approach for sentiment analysis of short texts. *2017 3rd International Conference on Control, Automation and Robotics (ICCAR)*, pages 705–710. Conference Name: 2017 3rd International Conference on Control, Automation and Robotics (ICCAR) ISBN: 9781509060887 Place: Nagoya, Japan Publisher: IEEE.

Kamala Kanth J, Vidhya Saraswathi P, and Radha T. 2025. Advancing Sentiment Analysis with LLM: Comparative Study with Traditional ML on Social Media Reviews. In *2025 11th International Conference on Communication and Signal Processing (ICCSP)*, pages 1328–1333, Melmaruvathur, India. IEEE.

Dipak R. Kawade and Kavita S. Oza. 2017. Sentiment analysis: machine learning approach. *International Journal of Engineering and Technology*, 9(3):2183–2186.

Moshe Koppel and Jonathan Schler. 2006. THE IMPORTANCE OF NEUTRAL EXAMPLES FOR LEARNING SENTIMENT. *Computational Intelligence*, 22(2):100–109.

Ronald W. Langacker. 1983. *Foundations of Cognitive Grammar*. Indiana University Linguistics Club.

J. R. Martin and P. R. R. White. 2005. *The Language of Evaluation*. Palgrave Macmillan UK, London.

Gaurav Meena, Krishna Kumar Mohbey, and Ajay Indian. 2021. Categorizing Sentiment Polarities in Social Networks Data Using Convolutional Neural Network. *SN Computer Science*, 3(2):116.

Gaurav Meena, Krishna Kumar Mohbey, and Ajay Indian. 2022. Categorizing Sentiment Polarities in Social Networks Data Using Convolutional Neural Network. *SN Computer Science*, 3(2):116.

Gaurav Meena, Krishna Kumar Mohbey, Sunil Kumar, and K. Lokesh. 2023. A hybrid deep learning approach for detecting sentiment polarities and knowledge graph representation on monkeypox tweets. *Decision Analytics Journal*, 7:100243.

Manish Munikar, Sushil Shakya, and Aakash Shrestha. 2019. Fine-grained Sentiment Classification using BERT. *arXiv preprint*. ArXiv:1910.03474 [cs, stat].

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81.

Carita Paradis. 2015. Conceptual spaces at work in sensory cognition: Domains, dimensions and distances. In *Applications of Conceptual Spaces*, pages 33–55. Springer.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon Emissions and Large Neural Network Training. *arXiv preprint*. ArXiv:2104.10350 [cs].

Cornelius Puschmann and Alison Powell. 2018. Turning Words Into Consumer Preferences: How Sentiment Analysis Is Framed in Research and the News Media. *Social Media + Society*, 4(3):205630511879772.

Luis Rei, Dunja Mladenić, and Simon Krek. 2016. A multilingual social media linguistic corpus. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*, Ljubljana, Slovenia.

Faiz Ali Shah, Ahmed Sabir, Rajesh Sharma, and Dietmar Pfahl. 2025. How Effectively Do LLMs Extract Feature-Sentiment Pairs from App Reviews? In Anne Hess and Angelo Susi, editors, *Requirements Engineering: Foundation for Software Quality*, volume 15588, pages 123–138. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

Neeraj Anand Sharma, A. B. M. Shawkat Ali, and Muhammad Ashad Kabir. 2025. A review of sentiment analysis: tasks, applications, and deep learning techniques. *International Journal of Data Science and Analytics*, 19(3):351–388.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Leonard Talmy. 2000. *Toward a Cognitive Semantics*. MIT Press.

Ana Valdivia, M. Victoria Luzón, Erik Cambria, and Francisco Herrera. 2018. Consensus vote models for detecting and filtering neutrality in sentiment analysis. *Information Fusion*, 44:126–135.

Marco Vassallo, Giuliano Gabrieli, Valerio Basile, and Cristina Bosco. 2024. Neutral Score Detection in Lexicon-based Sentiment Analysis: The Quartile-based Approach. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it*

2024), pages 976–982, Pisa, Italy. CEUR Workshop Proceedings.

Lorella Viola. 2023a. *The Humanities in the Digital: Beyond Critical Digital Humanities*. Palgrave Macmillan, S.l. OCLC: 1341438097.

Lorella Viola. 2023b. On the use of sentiment analysis for linguistics research. Observations on sentiment polarity and the use of the progressive in Italian. *Frontiers in Artificial Intelligence*, 6:1101364.

10 Appendix

The full prompt used for integrating SFL principles into Chat GPT-5 was:

```
You are a sentiment analysis system enhanced with
Systemic Functional Grammar (SFG) and the
Appraisal Framework. Your task is to analyze
the text and classify its sentiment by
considering the three metafunctions of language
(Halliday & Matthiessen, 1985) and the Appraisal
subsystems (Martin & White, 2005):

Attitude - What emotions, judgments, or
appreciations are expressed?

Engagement - How are these attitudes sourced or
attributed? Is the speaker
aligning with or distancing from other voices?

Graduation - What is the intensity or scaling of
the evaluation?

Based on this analysis, assign a sentiment
category that is more nuanced
than a simple polarity classification, such as:

- Positive
- Negative
- Neutral
- Mixed (e.g., "negative with a touch of humour")
- Cautionary
- Slight annoyance / slight sadness
- Joyful / optimistic

(Expand categories as appropriate according to
the Appraisal analysis.)

In addition to the category, provide a brief
explanation linking the classification to
specific linguistic features (e.g., subject
pronouns, verb aspect, intensifiers, irony, or
evaluative lexis).
```

Note that the number of items differs slightly between evaluations reported in Table 9 and Table 10: $n = 8595$ corresponds to tweets with gold annotations available, whereas $n = 8601$ corresponds to the full set of tweets where both GPT and XLM-T provided predictions.

Table 11 presents a representative sample of the explainability outputs generated by SFL-enhanced GPT-5. Each instance includes both the assigned label and the accompanying explanation derived from systemic-functional cues (e.g., lexis of affect, directive force, rhetorical markers). This sample illustrates the typical reasoning patterns by which the model justifies its classifications across different categories.

Given space constraints, only a small subset of cases is shown here. The complete dataset is published and available at (Anonymous, 2025).

Table 9: Performance of GPT-std and GPT-SFL vs manual gold ($n = 8595$).

	GPT-std			**GPT-SFL→base3**		
Class	P	R	F1	P	R	F1
Neg	0.087	0.084	0.086	0.102	0.090	0.096
Neu	0.884	0.875	0.879	0.882	0.900	0.891
Pos	0.176	0.206	0.190	0.197	0.164	0.179
Acc		0.784			0.803	
Macro-F1		0.385			0.389	
W-F1		0.787			0.797	

Table 10: Agreement of GPT-std and GPT-SFL with XLM-T ($n = 8601$).

	GPT-std			**GPT-SFL→base3**		
Class	P	R	F1	P	R	F1
Neg	0.415	0.061	0.106	0.470	0.063	0.111
Neu	0.371	0.881	0.523	0.374	0.914	0.531
Pos	0.492	0.157	0.238	0.529	0.121	0.197
Acc		0.383			0.388	
Macro-F1		0.289			0.280	
W-F1		0.289			0.285	

Table 11: Sample of GPT-SFL explainability outputs (full dataset available as supplementary material).

GPT-SFL label	Explanation
Neutral	No strong polarity; default neutral stance.
Cautionary	Directive lexis (e.g., *mi raccomando*) signals cautionary stance.
Negative	Judgemental/critical lexis indicates negative stance. Use of rhetorical question suggests dialogic engagement.
Neutral	No strong polarity; default neutral stance. Presence of intensifier increases graduation force.
Slight sadness	Lexis of sadness (e.g., *piango*, *sogno*) signals affective stance.
Positive	Positive affect detected from lexical choices and polarity.
Negative	Negative affect detected from lexical choices and polarity.
Neutral	No strong polarity; default neutral stance. Use of rhetorical question suggests dialogic engagement.
Neutral	No strong polarity; default neutral stance.
Neutral	No strong polarity; default neutral stance.