Issue Generation from Consumer Complaints through LLMs

William Barbosa[1], **Marcos Gôlo**[1], **Ricardo Marcacini**[1], **Solange Rezende**[1]

[1]University of São Paulo (ICMC-USP), São Carlos, São Paulo, Brazil

williancarddd@usp.br, {ricardo.marcacini,solange}@icmc.usp.br

Abstract. *Issue generation transforms raw textual reports, such as consumer complaint narratives, into descriptions to support triage and service monitoring. Issue generation is inherently an open-world task since new problems continuously emerge. Even though Large Language Models (LLMs) make this generative formulation viable, the transformation of narratives into issues remains underexplored. In this paper, we introduce and empirically study issue generation via opinion mining, and propose five LLM-based strategies that explore different input compositions, including full narratives, keywords, and explicit opinion aspects. We evaluate these approaches on four datasets with real complaints from the Consumer Financial Protection Bureau (CFPB). Our results reveal a consistent trade-off between alignment with the human issues. While statistical analysis shows no single strategy strictly dominates, aspect-guided methods offer significant structural advantages by providing explicit opinion units that support enterprise auditing, monitoring, and interpretability.*

1. Introduction

Consumer-generated textual feedback, including reviews and complaint narratives, is an abundant and accessible source of information about problems experienced with products and services [Lima and Nogueira 2021]. In many domains, customers already describe failures, dissatisfaction, and unmet expectations through free-text reports submitted to organizations and regulatory agencies [Barros and De Bona 2021]. However, these narratives are often long, noisy, and subjective, making it difficult for developers, managers, institutions, and decision-makers to directly use them for monitoring and action [Jorge et al. 2024]. Issue generation addresses this problem by transforming raw textual reports into concise, traceable, and actionable descriptions. The generated issues can support triage, prioritization, and product improvement. Figure 1 shows the issue generation.

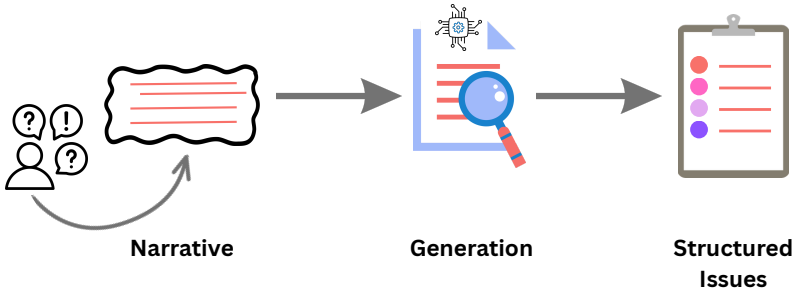


Figure 1. Issue generation: raw complaint narratives are transformed into issues through LLMs, supporting triage and monitoring at scale.

Opinion mining in consumer feedback represents a strategic opportunity for organizations seeking to systematically monitor problems, recurring failures, and sources of

dissatisfaction in their products and services [Lee et al. 2023]. In high-volume settings, this challenge intensifies, as manual analysis of free-text narratives becomes impractical and the need for automated mechanisms to transform fragmented texts into structured representations of reported problems becomes critical [De Caigny et al. 2020]. However, this potential remains underexplored in practice, as organizations may fail to operationalize complaint information, leaving rich customer sentiments and behavioral signals systematically underutilized [Bambauer-Sachse and Stuhldreier 2026, De Caigny et al. 2020]. Issue generation can be complemented by opinion mining, which aims not only to surface opinions but also to identify and organize the problems that emerge from user feedback.

In real-world settings, new problems continuously emerge with the launch of products and services, and the set of issues to be described evolves over time [de Lima et al. 2025, Win et al. 2022, De Caigny et al. 2020]. These conditions make issue generation an inherently open-world task, since the space of possible issues is dynamic and cannot be fixed in advance. A natural family of candidates for such open-ended text generation is automatic summarization, which condenses long texts into shorter sentences and ranges from extractive ranking to abstractive neural models [Watanangura et al. 2023]. More recently, the widespread availability of LLMs has further expanded the possibilities for generating structured text from free-text narratives [Allam et al. 2025], whether through summarization-style prompting or task-specific instructions. In this context, assessing which model or strategy performs best for issue generation becomes an important question. However, to the best of our knowledge, issue generation from consumer complaint narratives remains underexplored in the literature.

The absence of established methods for this task requires that candidate strategies be proposed and empirically evaluated, particularly regarding the role of opinion aspects, such as product and service characteristics, in guiding issue generation and improving the interpretability of the resulting outputs. This leads to the central question of this work: ***What LLM-based generation strategies perform better for issue generation from complaint narratives, and how do they empirically compare in terms of textual quality and interpretability when opinion aspects are explicitly incorporated?*** To answer this research question, we investigate five LLM approaches for issue generation from complaint narratives. In summary, our main contributions are:

1. We evaluate issue generation as a distinct, open-world natural language processing task within opinion mining;
2. We propose five generation strategies with different input compositions, ranging from purely narrative-based approaches to aspect- and keyword-guided methods;
3. We empirically evaluated our issue-generation strategies on four datasets comprising 133,368 real complaints from the Consumer Financial Protection Bureau.

2. Related Work

Summarization as a Related Abstraction Task. Issue generation is related to automatic summarization because both tasks transform longer texts into shorter representative outputs. Comparative surveys cover the techniques we adopt as baselines, spanning unsupervised extractive ranking such as TextRank and pre-trained abstractive models such as BART, along with their typical limitations, including the cohesion problems of extractive selection and the limited ability of lexical metrics such as ROUGE to capture semantic

equivalence [Watanangura et al. 2023]. More recently, LLMs prompted in an in-context learning setting have also been used to produce summaries directly, without task-specific training, which we include as an additional abstractive baseline. However, issue generation differs from summarization in its objective: rather than merely reducing length, it requires problem-oriented semantic abstraction. The generated output must identify a concrete failure, dissatisfaction, or source of complaint, remain grounded in the original narrative, and be structured enough to function as an operational issue. Thus, issue generation requires explicit attention to sentiment, opinion structure, and actionable problem formulation, which are not central optimization goals in conventional summarization.

LLMs for Opinion Analysis and Generation. LLMs have recently been explored for opinion analysis, taxonomy induction, and feedback structuring. TnT-LLM, for instance, uses LLMs to generate and refine taxonomies from unstructured text and then trains lightweight classifiers from LLM-generated pseudo-labels, but the method still maps texts to a closed taxonomy rather than generating issues from complaint narratives [Wan et al. 2024]. SCLF uses LLM feedback to identify causal keywords for sentiment classification and reduce shortcut learning, yet it remains focused on predefined sentiment labels rather than open-world problem generation [Song et al. 2025]. In a related direction, prompt-based LLMs have been used to structure citizen feedback into organized reports for disaster monitoring, but the resulting problems are not formally defined as actionable issues nor systematically evaluated as generated outputs [Cantini et al. 2025]. These studies demonstrate the potential of LLMs for processing opinionated text, but they do not establish issue generation as a distinct task. Even with the use of LLM in various tasks related to issue generation, their use for structured issue generation remains limited.

Research Gap. Summarization methods reduce text length, but are not oriented toward problem-focused abstractions. LLM methods, despite strong capabilities in opinion analysis and text structuring, have not characterized what constitutes an issue, how issues should be generated, or how their quality should be evaluated. Thus, we highlight an open gap: issue generation has not yet been treated as an independent natural language processing task. In this sense, in the next section, we introduce five issue-generation strategies to evaluate the issue generation from consumer complaint narratives.

3. Issue Generation through Large Language Models

This section describes how the issue generation task is operationalized. Each complaint narrative $x \in \mathcal{X}$ is mapped into a generated issue $\hat{y} \in \mathcal{Y}$. The goal is to generate a structured, traceable, and actionable issue description in an open textual output space, rather than selecting from a fixed set of categories. We propose five foundational approaches to LLM-based issue generation that differ in how the input prompt is constructed and in the type of evidence provided to the model.

3.1. Formalization of the Proposed Approaches

Let m denote one of the proposed generation approaches. For each method, an evidence function $E_m(x)$ extracts or organizes the information provided to the model together with the original complaint narrative. The method-specific prompt is constructed as:

$$q_m = \pi_m(x, E_m(x), \mathcal{I}_m), \tag{1}$$

where π_m is the prompt construction function, $E_m(x)$ is the evidence used by method m, and $\mathcal{I}_m$ is the set of task instructions. The generated issue is then obtained by:

$$\hat{y}_m = G_\theta(q_m), \tag{2}$$

where G_θ denotes the complete generation process, including conditional next-token prediction, decoding, and response parsing.

We consider three types of evidence. The first is lexical: a keyword vector $\mathbf{k}(x) = [k_1, \ldots, k_n]$, composed of the ten highest-ranked TF-IDF terms [Sparck Jones 1972]. The second is based on Aspect-Based Sentiment Analysis (ABSA) [Pontiki et al. 2016, Liu 2012], where each opinion unit is represented as $t_i = (a_i, o_i, s_i)$, with a_i an aspect term, o_i the associated opinion expression, and s_i the sentiment polarity. The full set of opinions extracted from x is $T(x) = \{t_i\}_{i=1}^{m_x}$, obtained via a dedicated zero-shot call to G_θ. From this, we define the negative subset $T^-(x) = \{t_i \in T(x) \mid s_i = \texttt{negative}\}$, which isolates dissatisfaction-bearing units.

3.2. Proposed Approaches

The five approaches differ only in the evidence provided to the model:

- **GenOnly**: no evidence beyond the narrative, i.e., $E_{\text{GenOnly}}(x) = \emptyset$.
- **GenKey**: lexical evidence, $E_{\text{GenKey}}(x) = \mathbf{k}(x)$.
- **GenAllAsp**: full opinion structure, $E_{\text{GenAllAsp}}(x) = T(x)$.
- **GenNAsp**: negative opinions only, $E_{\text{GenNAsp}}(x) = T^-(x)$.
- **GenLAsp**: full opinion structure, but the model is additionally required to return $A^*(x) \subseteq \{a_i \mid (a_i, o_i, s_i) \in T(x)\}$, the subset of aspects explicitly linked to the generated issue. Thus, the output is the pair $(\hat{y}_{\text{GenLAsp}}, A^*(x))$.

All methods share the same generation pipeline G_θ and task instructions $\mathcal{I}_m$ (varying only in content, not structure), ensuring that performance differences are attributable to input composition rather than architectural or instructional variations.

3.3. Model Selection and Inference Setup

To ensure that performance differences reflect properties of the input composition rather than artifacts of a single model, all five approaches are instantiated across three open-weight models with deliberately distinct profiles: `Qwen2.5-7B-Instruct` [Team 2025], `Llama-3.2-3B-Instruct` [Meta 2024], and `gemma-3-12b-it-int4-awq` [Gemma 2025]. These models originate from independent organizations with independent pretraining pipelines and alignment strategies, covering three points on the cost-performance spectrum: lightweight 3B inference, moderate 7B inference, and INT4-quantized inference from a 12B base. Smaller and quantized models additionally serve as a structural conformity stress-test, since zero-shot JSON generation degrades under reduced parametric capacity and weight compression, a pressure that is particularly relevant for aspect-dependent approaches where chained reasoning over extracted tuples may be asymmetrically sensitive to quantization-induced distribution shifts.

All approaches are implemented within the zero-shot paradigm [Brown et al. 2020], with English prompts that include an explicit task definition,

behavioral rules, and a JSON output specification. Inference is conducted using the vLLM Python library with high-concurrency batch processing. The main parameters are set as follows: temperature `0.0` for deterministic, reproducible outputs across all open-weight models; and a maximum output length of `2048` tokens to prevent truncation in approaches that produce richer, more structured outputs.

4. Experimental Evaluation

This section describes the experimental setup for the proposed approaches. We detail the datasets, evaluation protocol, metrics, and results. The source codes and prompts used in each approach are publicly available in the project repository[1].

4.1. Datasets

We use four datasets from the Consumer Complaint Database maintained by the Consumer Financial Protection Bureau (CFPB), a publicly available repository of complaints about financial products and services, which is regularly updated and released in open formats. The CFPB publishes only narratives submitted with consumer consent and after personally identifiable information has been removed, making it a large-scale platform with millions of narrative-issue pairs suitable for the task investigated in this work. The datasets span a six-year time window, from January 1, 2020, to January 1, 2026, covering four financial product categories: Debt collection, Loans (payday, title, personal, or advance), Prepaid card, and Debt or credit management, with approximately 109k, 10k, 8k, and 4k narrative-issue pairs, respectively, totaling 133,368 complaints.

The six-year time window maximizes linguistic and contextual heterogeneity across the corpus, capturing complaints produced under distinct economic conditions, including periods of tighter credit access, regulatory changes, pandemic crises, and shifts in consumer behavior, thereby reducing the risk of overfitting to a narrow historical interval and improving the external validity of the findings. The four product categories correspond to everyday financial situations that are broadly recognizable, such as debt collection, short-term lending, prepaid card, and debt management, facilitating qualitative error analysis without requiring deep familiarity with complex financial instruments.

Each complaint in the CFPB database is annotated with two hierarchical fields: *issue*, defined as the main problem identified by the consumer in the complaint, and *sub-issue*, the more specific sub-problem within that complaint. These fields serve as the reference labels for evaluation. The four subsets present notably distinct annotation profiles: Debt Collection contains 7 issues and 29 sub-issues, all fully annotated; Loans contains 26 issues and 23 sub-issues, though more than 90% of instances lack a sub-issue annotation; Prepaid Card contains 6 issues and 19 sub-issues, with approximately 10% of instances missing sub-issue annotation; and Debt or Credit Management contains 6 issues and no sub-issue annotations whatsoever. This heterogeneity in issue granularity reflects the varying operational definitions used across financial product categories and motivates the use of the concatenated issue and sub-issue fields in the evaluation protocol.

4.2. Evaluation Protocol

Each generated issue $\hat{y}$ is compared against a reference issue r, derived from the CFPB dataset by concatenating the *Issue* and *Sub-issue* fields assigned by domain experts dur-

[1] `https://github.com/williancarddd/issue-generation`

ing complaint intake. To compare the two texts, we use BERTScore, which indicates how closely a generated issue semantically matches the reference issue. BERTScore [Zhang et al. 2019] is adopted as the primary alignment metric for its ability to compute semantic similarity via contextual token embeddings, reporting f_1 scores under the TF-IDF-weighted variant to reduce the influence of generic tokens. However, we normalize the raw scores by the ratio between candidate and reference lengths, $\hat{s} = s \cdot \frac{\min(|c|,|r|)}{\max(|c|,|r|)}$, since BERTScore does not penalize length deviations relative to the reference. This experimental choice is consistent with prior work showing that length bias is a systematic concern in automatic evaluation of text generation and that decoupling quality signals from sequence length yields more reliable assessments [Zhang et al. 2025].

where s is the original BERTScore, $|c|$ is the number of characters in the candidate $\hat{y}$, and $|r|$ is the number of characters in the reference r. The factor is bounded in $(0, 1]$, penalizing candidates that deviate from the reference length in either direction proportionally to the magnitude of the deviation. We treat this adjustment as one component of the evaluation protocol rather than a standalone validated metric, acknowledging that its correlation with human judgment of granularity remains an open empirical question.

To complement BERTScore with criterion-based evaluation, we adopt G-Eval [Liu et al. 2023], an LLM-as-a-judge protocol in which the evaluator (GPT-4o) reasons over an explicit rubric and scores each generated issue $\hat{y}$ in $[0, 1]$ given the complaint narrative x and the reference issue r. The rubric operationalizes four properties: **representative coverage**, rewarding $\hat{y}$ that captures the same underlying problem as r; **traceability**, requiring $\hat{y}$ to be grounded in x and penalizing hallucinations; **appropriate granularity**, penalizing generic or overly short outputs; and **adaptability**, accepting paraphrases and synonyms that preserve semantic intent. We do not apply the length normalization to G-Eval, since that correction targets an artifact specific to BERTScore, whose embedding-based similarity is sensitive to length deviation from the reference, whereas G-Eval already controls for output adequacy through the **appropriate granularity** criterion; normalizing a criterion-based judgment by length would confound the rubric it is meant to apply, so G-Eval is reported unnormalized.

This makes G-Eval complementary to BERTScore, as it additionally conditions on the source x and explicitly rewards grounding while penalizing hallucination. Since G-Eval scores are normalized to $[0, 1]$, the default acceptance threshold of 0.5 coincides with the midpoint of the scale, the point of indifference at which the judge is ambivalent and a generation is neither clearly acceptable nor clearly rejectable. To avoid counting such borderline cases as acceptable, we adopt a stricter threshold $\tau = 0.7$, which requires the judge to lean unambiguously toward the positive side of the scale. The acceptance rate then reports the fraction of issues a downstream triage pipeline could adopt without human revision, and we treat τ as a reporting convention that complements the mean score rather than a human-validated quality guarantee.

4.3. Results and Discussion

Table 1 reports BERTScore f_1 and G-Eval (GE) for the five proposed strategies through three models and the baselines, across the four datasets. Among the baselines, **GenSumm** denotes the summarization baseline in which the same open-weight LLMs are prompted for a generic summary without any issue-specific objective, and it is reported per model

as GenSumm (Qwen), GenSumm (Gemma), and GenSumm (Llama). Each cell pairs the two evaluation perspectives, the lexical-semantic alignment to the reference (f_1) and G-Eval. Read together, they show that methods that score high on one metric often score low on the other, and only the proposed strategies remain competitive on both. The baselines show this most clearly, since none is strong on both metrics simultaneously.

Table 1. Comparison of issue generation strategies and baselines across four financial product categories. Results are BERTScore f_1 and G-Eval (GE). In each column, bold and underline mark the best and second-best results.

Model	Method	Debt Col.		Loans		Prepaid Car.		Debt Man.	
		f_1	GE	f_1	GE	f_1	GE	f_1	GE
Baselines	TextRank	0.302	0.373	0.178	0.330	0.334	0.343	0.171	0.339
	BART	<u>0.544</u>	0.314	0.368	0.247	**0.528**	0.268	0.310	0.265
	GenSum (Qwen)	0.242	0.599	0.151	0.537	0.298	<u>0.596</u>	0.151	0.505
	GenSum (Gemma)	0.213	**0.611**	0.113	**0.558**	0.230	**0.604**	0.124	<u>0.508</u>
	GenSum (Llama)	0.271	<u>0.606</u>	0.151	0.541	0.268	0.587	0.138	**0.514**
Qwen-7B	GenOnly	0.535	0.539	0.384	0.528	<u>0.504</u>	0.565	0.385	0.494
	GenKey	**0.547**	0.528	<u>0.403</u>	0.522	0.501	0.545	<u>0.400</u>	0.479
	GenAllAsp	0.501	0.521	0.348	0.501	0.496	0.556	0.340	0.478
	GenNAsp	0.501	0.527	0.345	0.499	0.494	0.558	0.343	0.479
	GenLAsp	0.521	0.472	**0.467**	0.477	0.455	0.497	**0.480**	0.451
Gemma-12B	GenOnly	0.501	0.543	0.364	<u>0.547</u>	0.479	0.557	0.385	0.495
	GenKey	0.514	0.538	0.383	0.535	0.476	0.549	0.397	0.482
	GenAllAsp	0.513	0.536	0.350	0.532	0.491	0.553	0.361	0.484
	GenNAsp	0.513	0.536	0.354	0.519	0.488	0.550	0.368	0.484
	GenLAsp	0.530	0.526	0.353	0.517	0.503	0.544	0.364	0.475
Llama-3B	GenOnly	0.376	0.551	0.218	0.514	0.416	0.540	0.209	0.474
	GenKey	0.420	0.547	0.256	0.509	0.461	0.544	0.255	0.472
	GenAllAsp	0.332	0.516	0.202	0.469	0.368	0.505	0.196	0.444
	GenNAsp	0.335	0.504	0.205	0.487	0.379	0.517	0.197	0.461
	GenLAsp	0.341	0.469	0.222	0.447	0.386	0.451	0.218	0.431

BART achieves a high f_1 (0.54) on Debt Collection but the lowest G-Eval among all methods (0.31), while GenSumm (Gemma) shows the opposite behavior, with the lowest f_1 (0.21) but the highest G-Eval (0.61). This effect is an artifact of output length. The reference issues are short (34–78 characters on average across datasets), and while BART produces the shortest text (78.1 characters, 14.4 words), GenSumm the longest (259.6 char, roughly $3.3\times$ longer), so the length-normalized f_1 rewards BART's brevity while the LLM judge rewards GenSumm's fluent, complete text. Only the direct strategies (GenOnly, GenKey) remain close to the reference scale, consistent with their stronger f_1; full per-method length statistics for the references and candidates are available in the repository. TextRank, which selects central sentences verbatim, is concise on neither axis and ranks lowest overall. The summarizers' behavior reveals why summarization requires explicit direction for this task, since a summary contains more than just the problem.

In a Loans case (id 6007828), GenSumm produced a four-sentence summary that correctly described a financed solar installation, the vendor's bankruptcy, and the customer's request to be released from the contract. The issue is covered, yet in a broader account, so a human operator would still have to extract it. Since GenSumm is the same LLM used by the proposal, prompting it without a task-specific objective shows that gen-

erative models already hold the capacity to capture the problem, but, left undirected, stop at a summary rather than an issue, precisely the gap that issue generation addresses.

The metrics also vary systematically across datasets, driven by each reference's annotation profile. Debt Collection and Prepaid Card, which include sub-issue annotations, yield the highest f_1 values, whereas Loans and Debt Management, which largely or entirely lack sub-issue annotations, yield consistently lower f_1 across all methods and models. The cause is length alignment: without a sub-issue, the reference reduces to the issue field alone, which is short and generic, so the length-normalized f_1 penalizes the more detailed generations even when they are correct. This is expected, since a very short reference is dominated by generic, non-actionable keywords, against which a specific, actionable issue is necessarily misaligned in length. G-Eval, which does not depend on reference length, is far more stable across the four datasets, confirming that the cross-dataset f_1 gap reflects the reference-human issue rather than generation quality.

Model capacity affects the two metrics differently, and not as expected. We hypothesized that larger models would better internalize the criteria of an issue. Although this happens in the G-Eval, it did not happen in the f_1. The 12B model does not outperform the 7B model. Judging issue quality thus depends less on raw size than on the in-context formulation of the task. Capacity helps elsewhere, though. Larger models return valid JSON more consistently in the aspect-extraction stage. This is why the aspect-based strategies, which depend on well-formed triplet extraction, perform best on the 12B model and worst on the 3B model. Model size, therefore, matters for the structured intermediate representation, but not for issue quality itself.

Each proposed method exhibits a characteristic behavior. The direct strategies, GenKey and GenOnly, achieve the highest alignment with the reference issue. Prompted in an in-context learning setting, the LLM captures the problem-oriented nuances that the institutional issue encodes and produces text that closely resembles the form of an issue. The capacity-dependent pattern across models reinforces this. Aspect-based approaches perform best on Gemma-12B, while lexical and direct approaches perform best on Qwen-7B, indicating that structured evidence requires additional capacity to be exploited.

Under GenLAsp, Gemma-12B reaches $f_1 = 0.81$ (id 8744300) by synthesizing the linked units *medical hardship application [n]: submitted* and *credit report [n]: collections* into a concise issue, whereas Llama-3B, on the same structures, repeatedly enumerates the aspect *Wakefield [n]: violating my HIPAA RIGHT* (id 4358651, $f_1 = 0.00$) or regresses to narrative text (GenAllAsp, id 10470227, $f_1 = 0.02$). Errors in the aspect-based strategies may be confounded with ABSA noise, but a post-hoc analysis of the stored `aspects_json` outputs shows JSON parsing failures below 0.2%. The one exception is Llama-3.2-3B on GenNAsp, where a polarity miscalibration left 10–68% of inputs empty. The remaining failure modes contrast each method's best and worst cases:

- **GenKey:** relies on surface signal and is susceptible to dataset-specific artifacts. In its worst case (id 5429763, $f_1 = 0.07$), TF-IDF keywords dominated by anonymization placeholders (`xxxx`, `xxxx pay`) collapsed the generation into the trivial "owe $820.00". Aspect-guided strategies are more robust here, since the ABSA layer discards empty placeholders when extracting (a_i, o_i, s_i).
- **GenAllAsp / GenNAsp:** competitive on f_1, but regress to literal narrative copy when the aspect set provides no clear hierarchy (*account [n]: This account is not*

my account), or enumerate entities without unifying them into one issue.

- **GenLAsp:** can loop on repetitive generation when the linked set is narrow; in a case with $f_1 = 0.00$, the high-polarity aspect *Wakefield [n]: violating my HIPAA RIGHT* induced the loop, in contrast with cases of multiple balanced aspects.
- **GenOnly:** with multi-event narratives, it may paraphrase the narrative entirely; in a case with $f_1 = 0.05$, it produced a long description rather than an issue.

In addition to the f_1 and G-Eval evaluation, we performed Friedman's statistical test with Nemenyi's post-test [Trawiński et al. 2012] to compare the methods on Qwen-7B, the most competitive open model overall, analyzing each metric separately to avoid mixing the two opposite biases discussed above. The Friedman test ranks the methods by performance on each dataset and detects whether their average ranks differ overall, while the Nemenyi post-test identifies which pairs of methods differ significantly. Figures 2a and 2b present the critical difference diagrams for f_1 and G-Eval. The diagrams present the average rankings of the methods, where methods connected by a line do not differ statistically at the 95% confidence level ($\alpha = 0.05$).

For f_1 (Figure 2a), GenKey and GenOnly obtained the best average rankings, with the proposed strategies grouped at the top and no statistically significant differences among themselves, leaving the choice between them open. The summarization baselines obtained the worst average rankings, with GenSumm presenting a statistically significant difference from GenKey ($p = 0.012$) and GenOnly ($p = 0.050$). For G-Eval (Figure 2b), the ranking changes. GenSumm obtained the best average ranking, followed by GenOnly and GenKey, while BART obtained the worst. GenSumm presents a statistically significant difference only from BART ($p = 0.001$), and BART presents a statistically significant difference from GenOnly ($p = 0.012$). As discussed, GenSumm's advantage stems from coverage and verbosity rather than from concise issues, whereas BART's brevity, which helped it on f_1, is penalized by a criterion-based judge.



(a) BERTScore f_1. **(b) G-Eval.**

Figure 2. Critical difference diagrams using Friedman + Nemenyi tests for Qwen-7B. Methods connected by a line do not differ significantly.

Together, the two analyses make the central point. The summarization baselines are each strong on a single metric and collapse on the other: BART is last on G-Eval, and GenSumm is last on f_1, whereas the proposed direct strategies are the only methods ranking near the top on both, making GenOnly and GenKey the most robust choices. The aspect-based strategies do not lead to either metric, but they add a property that the direct methods lack. By exposing explicit opinion units (a_i, o_i, s_i) and linking them to the generated issue, GenLAsp lets each issue be mapped back to the polarities and aspects of the source text, enabling an audit of the model's reasoning. This traceability is valuable in enterprise scenarios such as product monitoring, regulatory triage, and emerging problem detection, where justifying an issue with concrete signals is an operational requirement

that the metrics do not measure. The aspect-guided strategies are therefore best seen not as replacements for the direct methods, but as an auditable alternative that trades a modest amount of criterion-based quality for traceability.

5. Conclusion and Future Work

This work introduced and empirically studied issue generation from consumer complaint narratives as a distinct, open-world NLP task within opinion mining. We proposed five LLM-based strategies and compared them against extractive and abstractive baselines on 133,368 real complaints from the CFPB, evaluated with BERTScore and G-Eval against a reference taxonomy. The evaluation revealed a length-driven dissociation between the two metrics, since BART, favored by alignment-based f_1, and GenSumm, favored by criterion-based G-Eval, occupy opposite extremes of output length, showing that no automatic metric, in isolation, captures issue quality. When both metrics are combined into a single ranking, this dissociation collapses, as the proposed strategies rank above both abstractive baselines, extractive summarization (TextRank) is significantly the weakest, and the proposed strategies are statistically indistinguishable from one another.

This last point reframes method selection. Since no proposed strategy is significantly better than another in textual quality, the choice should rest on operational criteria that aggregate scores do not measure. Aspect-guided generation, and GenLAsp in particular, is advantageous in this regard, since by exposing explicit opinion units and linking them to the generated issue it makes each issue auditable against the source text, supporting monitoring, regulatory triage, and emerging problem detection. The absence of human judgment is the main limitation of this study, together with the assumption that every complaint contains an actionable issue, which holds for our curated complaint–issue pairs but not necessarily in production, where a relevance classifier at the start of the pipeline would be pertinent to filter non-actionable narratives before generation. Accordingly, future work will quantify traceability with a direct metric over the linked aspects, validate issue quality through human evaluation of the operational properties defined for the task, and study the robustness of this filtering stage. We further intend to investigate additional LLMs, integrate aspect-guided generation with retrieval-augmented generation to mitigate anonymization artifacts, and address the full formulation with $m_x > 1$, which requires annotation schemes capturing multiple issues.

6. ACKNOWLEDGMENTS

This work was supported in part by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES), Finance Code 001, by the Brazilian National Council for Scientific and Technological Development (CNPq) [Grants No. 134627/2025-3, 317002/2026-0, 309575/2021-4, 307184/2025-0, 303031/2023-9, and 101057/2024-5], by the Academic Master's and Doctoral Program for Innovation (MAI/DAI) with a scholarship granted in partnership with BIRDIE.AI and administered by the Research Support Foundation (FUNAPE) [Scholarship No. 44519], by the Research Support Foundation of the State of Goiás (FAPEG) [Grant No. 23104.003594/2025-12], and by the São Paulo Research Foundation (FAPESP) [Grant No. 2025/07160-0].

References

Allam, H., Makubvure, L., Gyamfi, B., Graham, K., and Akinwolere, K. (2025). Text classification how machine learning is revolutionizing text categorization. *Information*.

Bambauer-Sachse, S. and Stuhldreier, S. M. (2026). Effective complaint management on social networks: the impact of response strategy, type and account on consumer reactions. *Journal of Services Marketing*, pages 1–21.

Barros, J. M. and De Bona, G. (2021). A deep learning approach for aspect sentiment triplet extraction in portuguese. In *Brazilian Conference on Intelligent Systems*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cantini, R., Cosentino, C., Marozzo, F., Talia, D., and Trunfio, P. (2025). Harnessing prompt-based large language models for disaster monitoring and automated reporting from social media feedback. *Online Social Networks and Media*, 45:100295.

De Caigny, A., Coussement, K., De Bock, K. W., and Lessmann, S. (2020). Incorporating textual information in customer churn prediction models based on a convolutional neural network. *International Journal of Forecasting*, 36(4):1563–1578.

de Lima, V. M. A., Barbosa, J. R., and Marcacini, R. M. (2025). Issue detection and prioritization based on mobile application reviews. *Software Quality Journal*, 33(1):6.

Gemma, T. (2025). Gemma 3. *Google Deep Mind*.

Jorge, G. A. Z., Bezerra, D. A., Xavier, C. C., and Pardo, T. A. S. (2024). Multilingual extractive summarization: Investigating state-of-the-art methods for english and brazilian portuguese. In *Brazilian Conference on Intelligent Systems*. Springer.

Lee, Y., Kim, J., Kim, D., Kho, Y., Kim, Y., and Kang, P. (2023). Painsight: An extendable opinion mining framework for detecting pain points based on online customer reviews. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 215–227.

Lima, B. and Nogueira, T. (2021). Incorporating text specificity into a convolutional neural network for the classification of review perceived helpfulness. In *Brazilian Conference on Intelligent Systems*, pages 480–495. Springer.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Springer Nature.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Meta, T. (2024). The llama 3.1 herd of models. *arXiv preprint*.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *Int. workshop on semantic evaluation*.

Song, R., Li, Y., Tian, M., Wang, H., Giunchiglia, F., and Xu, H. (2025). Causal keyword driven reliable text classification with large language model feedback. *Information Processing & Management*, 62(2):103964.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Team, Q. (2025). Qwen2.5 technical report.

Trawiński, B., Smźtek, M., Telec, Z., and Lasota, T. (2012). Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. *International Journal of Applied Mathematics and Computer Science*, 22(4):867–881.

Wan, M., Safavi, T., Jauhar, S. K., Kim, Y., Counts, S., Neville, J., Suri, S., Shah, C., White, R. W., Yang, L., et al. (2024). Tnt-llm: Text mining at scale with large language models. In *Conference on Knowledge Discovery and Data Mining*.

Watanangura, P., Vanichrudee, S., Minteer, O., Sringamdee, T., Thanngam, N., and Siriborvornratanakul, T. (2023). A comparative survey of text summarization techniques. *SN Computer Science*, 5(1):47.

Win, M. N., Ravana, S. D., and Shuib, L. (2022). Sentiment attribution analysis with hierarchical classification and automatic aspect categorization on online user reviews. *Malaysian Journal of Computer Science*, 35(2):89–110.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, Y., Xu, W., Liu, Z., Nakagawa, T., and Freitag, M. (2025). Penalizing length: Uncovering systematic bias in quality estimation metrics. *arXiv e-prints*.