Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Detection Is Not Early Warning: Taxonomy-Aware Statistical Surveillance of Emerging Themes in U.S. Consumer-Finance Complaint Narratives, 2019–2023

Sujan sapkota

sapkotasujan53@gmail.com

Independent Researcher

Research Article

Keywords: consumer complaints, consumer finance, financial regulation, emergence detection, surveillance, taxonomy harmonization

Posted Date: September 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11029331/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Consumer complaint narratives are increasingly used for financial-market monitoring, but longitudinal inference is vulnerable to classification breaks, duplicate text, selected observation, and the tendency to relabel contemporaneous shocks as early warnings. We audit and harmonize 2019–2023 U.S. Consumer Financial Protection Bureau (CFPB) complaint narratives across debt collection, checking/savings, mortgage, and credit-card/prepaid-card products. The supplied four-product extract contains 377,353 narratives; a post-taxonomy-change recovery file contributes 14,256 records, yielding 391,609 harmonized records and 373,776 unique narrative strings. We implement an auditable surveillance rule over ten interpretable themes using trailing historical baselines, one-sided Poisson tail screening, Benjamini-Hochberg false-discovery-rate control, minimum support, abnormal growth, and persistence. Under the primary two-month persistence rule, COVID-19/pandemic and mortgage-forbearance language first qualify in March 2020 but are not operationally confirmed until April 2020. Across 81 threshold specifications, COVID-related language triggers under all specifications and mortgage forbearance under 60; other themes are substantially less stable. Public benchmarks place the strongest signals at or after broad recognition of the underlying shock. The evidence therefore supports contemporaneous complaint surveillance, not systematic advance warning, and shows that taxonomy harmonization and explicit confirmation timing are prerequisites for defensible longitudinal monitoring.

JEL classification: C55; D18; G21; G28

1. Introduction

Consumer complaints are a direct but selected record of problems that individuals choose to report about financial products and services. For regulators and financial-services analysts, the attraction of complaint narratives is obvious: structured taxonomies summarize a complaint, while free text may reveal changing language, mechanisms, and harms before those patterns are obvious in aggregate reporting. The difficulty is equally important. Complaint data are not a random sample of consumers, narrative publication is selective, duplicate text can distort prevalence, and administrative taxonomies can change in ways that create artificial structural breaks.

Prior research has already established that topic modeling and natural-language processing can extract useful structure from CFPB complaints. Bastani, Namavari, and Shaffer (2019) used latent Dirichlet allocation and examined temporal topic patterns. Sangaraju et al. (2022) compared classical topic models with BERT-based representations in published IEEE conference work. Gao et al. (2024) developed NLP-based systematic-anomaly detection in the Journal of Operational Risk. More recent work has examined complaint outcomes and large-language-model adoption. Consequently, simply applying an NLP model to CFPB narratives is not a defensible novelty claim.

This study asks a narrower and more falsifiable question: when an analyst formalizes emergence using only historical information, how stable are complaint-based theme alerts, and do those alerts actually justify 'early-warning' language? The answer matters because an attractive time-series spike can be

statistically real without being anticipatory. A system that detects a shock after consumers begin describing it may still be useful for surveillance, but it should not be represented as prediction.

The paper makes three contributions. First, it documents and repairs an August 2023 CFPB product-taxonomy discontinuity that otherwise creates a severe artifact in credit-card complaint trends. Second, it provides an auditable statistical emergence rule with multiple-testing control, support, growth, and persistence requirements. Third, it stress-tests the distinction between detection and early warning using 81 statistical specifications and independently dated public benchmarks. The results favor the more conservative interpretation: complaint narratives can register large consumer-finance shocks quickly, but the evidence here does not establish systematic advance warning.

2. Related Literature and Research Gap

Bastani et al. (2019) demonstrated that CFPB complaint narratives support topic extraction and longitudinal analysis using LDA. Sangaraju et al. (2022) further showed the relevance of contextual embedding-based topic representations, including financial-domain embeddings. Gao et al. (2024) moved closer to surveillance by detecting systematic anomalies in complaint narratives. A 2026 working paper by Wang, Zhu, and Chen (arXiv:2606.22664) uses complaint narratives and machine-learning features in a temporally split monetary-relief prediction setting, while Shin, Kim, and Shin (2026) analyze more than one million CFPB complaints to study large-language-model adoption and complaint outcomes. These studies make clear that the remaining gap is not 'NLP on CFPB data' but disciplined longitudinal measurement and claim calibration.

The remaining gap is therefore methodological rather than algorithmic: a surveillance design must separate changes in the observation system from changes in consumer narratives, define emergence before interpreting it, prevent future information from leaking backward, and distinguish contemporaneous detection from genuine lead time. The present study addresses that gap with a transparent baseline. It intentionally prioritizes auditability over model complexity.

Table 1
Literature and competitor comparison

Study	CFPB narratives	Temporal analysis	Formal emergence	Taxonomy-break control	Early-warning validation
Bastani et al. (2019)	Yes	Yes	No	Not central	No
Sangaraju et al. (2022, IEEE)	Yes	Limited/benchmarking	No	Not central	No
Gao et al. (2024)	Yes	Anomaly-oriented	Yes, anomaly focus	Not central	No independent lead-time test
Wang et al. (2026 preprint)	Yes	Temporal train/test	No	Not central	Outcome prediction, not emergence
Shin et al. (2026)	Yes	Yes	No	Not central	No
Present study	Yes	Yes	Yes	Explicit	Benchmarked; no strong early-warning claim

3. Data and Dataset Reconstruction

The analysis uses CFPB Consumer Complaint Database records supplied as two extracts plus the scripts that created them. The first extract retained non-empty narratives received between January 1, 2019 and December 31, 2023 for Debt collection, Checking or savings account, Credit card or prepaid card, and Mortgage. The second extract recovered records labeled Credit card or Prepaid card after the CFPB product taxonomy changed in August 2023. Complaint ID was used to prevent overlap when appending the recovery file.

Table 2
Dataset audit and harmonized sample

Audit item	Verified result
Original four-product extract	377,353 records x 16 fields
Date range	2019-01-01 to 2023-12-31
Narrative missingness	0 by extraction design
Exact duplicate excess in original extract	14,802 records
Post-change credit-card/prepaid recovery	14,256 records
Harmonized record-level dataset	391,609 records
Unique narrative strings used for primary text surveillance	373,776
Debt collection	151,241
Checking or savings account	75,331
Mortgage	61,751
Harmonized credit card / prepaid card	103,286
Submission channel in supplied extracts	Web only

3.1 Taxonomy discontinuity and harmonization

The legacy product label 'Credit card or prepaid card' disappears in late 2023 because the CFPB complaint form moved to separate Credit card and Prepaid card product labels. The recovery extract begins on August 25, 2023 and contains 12,905 Credit card records and 1,351 Prepaid card records. For longitudinal aggregation, the historical combined label and the two newer labels are mapped to a harmonized family, 'Credit card / prepaid card,' while the original product field is retained. The mapping is explicitly a longitudinal harmonization rule, not an assertion that every sub-product or issue definition remained unchanged.

This correction is substantively important. Without it, a model would treat an administrative relabeling as a sudden collapse in complaint volume, potentially generating spurious change points and misleading model evaluation. Administrative taxonomy is therefore part of the measurement process, not merely a preprocessing detail.

3.2 Duplicate narratives, short texts, and selection

The primary theme-intensity analysis uses one copy of each exact narrative string. This choice limits mechanical inflation from copied correspondence or repeated submissions while preserving the full record-level dataset for descriptive counts. The original extract contains 2,268 narratives shorter than 10 words, 6,023 shorter than 15 words, and 16,463 shorter than 25 words. These counts show that short-

text filtering is not negligible, although the current primary analysis avoids an arbitrary minimum-word exclusion.

The study population must be interpreted narrowly. The supplied extracts contain web-submitted complaints with published narratives in selected product families. Complaint propensity, digital access, awareness of the CFPB, narrative consent/publication, company mix, product mix, and changes in reporting practices can all affect observed signals. Complaint intensity is therefore not population incidence.

4. Research Questions

RQ1. Can a fully specified statistical surveillance rule identify persistent abnormal increases in interpretable consumer-finance themes using only historical complaint information?

RQ2. How stable are first-alert dates across reasonable choices of baseline window, minimum support, growth threshold, and persistence?

RQ3. Do the strongest complaint-based alerts justify an early-warning interpretation when compared with independently dated public recognition benchmarks?

5. Methodology

5.1 Interpretable theme representation

Ten consumer-finance themes were represented by transparent phrase lexicons: COVID-19/pandemic; mortgage forbearance; overdraft/NSF fees; fraud/unauthorized transactions; account access/frozen account; debt validation/not owed; medical debt; late fees; dispute handling; and payment apps/transfers. A narrative contributes at most one count to a theme in a month. Themes can overlap because a single complaint may legitimately describe multiple problems. The purpose of this representation is not to claim semantic completeness; it is to create an auditable surveillance baseline whose trigger logic can be inspected directly.

5.2 Theme intensity and temporal information set

For theme k and month t, intensity is represented by the number of unique narratives containing the theme and, for visualization, the prevalence per 1,000 unique narratives in that month. Every alert is calculated from counts available no later than month t. External recognition dates are not used in theme construction or alert generation.

5.3 Primary emergence rule

For each theme-month, the historical baseline is the preceding 12 months. The primary screening statistic uses a one-sided Poisson upper tail with a half-count continuity adjustment distributed across the baseline window, lambda(k,t) = [sum of the preceding 12 monthly counts + 0.5]/12. A month qualifies

only if (i) the observed count is at least 20, (ii) the observed-to-baseline ratio is at least 2.0, and (iii) the Benjamini-Hochberg false-discovery-rate-adjusted q value is below 0.05 across all evaluated theme-month tests. The primary rule additionally requires two consecutive qualifying months. We distinguish the onset month (the first month in a qualifying run) from the confirmation month (the month in which the persistence requirement becomes knowable). This distinction prevents a two-month persistence rule from being backdated by one month.

The Poisson benchmark is deliberately simple and interpretable. Complaint counts can be overdispersed and serially dependent, so q values are treated as screening statistics rather than causal or structural parameters. Substantive interpretation therefore depends jointly on persistence, effect magnitude, minimum support, multiplicity control, and specification stability. The surveillance rule is designed as an auditable baseline, not as a claim that complaint generation follows an exact Poisson process.

5.4 Robustness grid

To stress-test threshold dependence, the rule is recomputed under 81 combinations: baseline windows of 6, 12, or 18 months; minimum growth ratios of 1.5, 2.0, or 2.5; minimum monthly counts of 10, 20, or 40; and persistence requirements of 1, 2, or 3 consecutive qualifying months. These choices are an exploratory robustness grid rather than a preregistered confirmatory family. For every specification we report both onset and confirmation timing; a persistent alert is not treated as operationally available before its confirmation month.

5.5 External recognition benchmark

Public dates are used only to assess interpretation after alert generation. The benchmark set is not a prospectively registered event universe and therefore cannot establish confirmatory early-warning performance. COVID-19 had been characterized as a pandemic by the World Health Organization on March 11, 2020, and the U.S. CARES Act became law on March 27, 2020, creating federal mortgage-forbearance rights. CFPB mortgage-servicing guidance followed on April 3, 2020, and the Bureau's complaint bulletin on coronavirus-related complaints appeared on May 21, 2020. These dates provide a conservative boundary test: an April-confirmed complaint alert cannot be described as advance warning of a shock and policy response already publicly recognized in March.

6. Results

6.1 Descriptive structure and taxonomy effect

Debt collection is the largest harmonized family (151,241 records), followed by credit card/prepaid card (103,286), checking/savings (75,331), and mortgage (61,751). The temporal product plot shows that the harmonized credit-card/prepaid series continues through late 2023; the unharmonized legacy label does not. This confirms that the apparent late-2023 disappearance in the original extract is administrative rather than behavioral.

6.2 Primary emergence results

Under the primary 12-month, twofold-growth, minimum-20, two-month-persistence specification, two of the ten predefined themes form persistent runs: COVID-19/pandemic language and mortgage forbearance. Both runs begin in March 2020, but because persistence requires a second qualifying month, the operational confirmation month is April 2020. COVID-related language has 298 unique-narrative mentions in March and 927 in April, following zero mentions in the preceding 12-month baseline. Mortgage-forbearance language has 107 mentions in March against a preceding 12-month mean of 44.8 (approximately 2.39 times the raw baseline), followed by 223 mentions in April. Thus March is an onset date, not an alert date available to a real-time analyst under the primary persistence rule.

Table 3
Primary persistent-emergence results

Theme	Onset month	Confirmation month	Onset count	12-month raw mean	Interpretation
COVID-19 / pandemic	2020-03	2020-04	298	0.0	Robust shock detection; not advance warning
Mortgage forbearance	2020-03	2020-04	107	44.8	Persistent contemporaneous emergence

6.3 Specification stability

Table 4
Robustness across 81 specifications

Theme	Specs with run	Share	Modal onset month	Modal-onset specs
COVID-19 / pandemic	81/81	100.0%	2020-03	54
Mortgage forbearance	60/81	74.1%	2020-03	36
Overdraft / NSF fees	36/81	44.4%	2023-01	21
Account access / frozen account	24/81	29.6%	2020-04	12
Payment apps / transfers	21/81	25.9%	2020-04	12
Late fees	15/81	18.5%	2020-04	12
Medical debt	12/81	14.8%	2022-03	6
Debt validation / not owed	9/81	11.1%	2021-03	9
Fraud / unauthorized transactions	3/81	3.7%	2021-03	3
Dispute handling	0/81	0.0%		0

The robustness grid sharpens rather than overturns the primary conclusion. COVID-19/pandemic language produces a qualifying run in all 81 specifications, and mortgage forbearance in 60 of 81 (74.1%). Because persistence varies across the grid, confirmation timing necessarily shifts: for COVID-related language, confirmations span March-May 2020 under the shorter-baseline specifications and July-September under 18-month baselines; mortgage-forbearance confirmations are concentrated in March-May 2020 with a smaller set later in 2020. The key robustness result is therefore the existence of a stable shock-related emergence pattern, not a single invariant alert date.

6.4 Detection versus early warning

The strongest signals do not support a strong early-warning claim. Under the primary two-month persistence rule they are confirmed in April 2020, after the World Health Organization's March 11 pandemic characterization and after the CARES Act became law on March 27. The CFPB's April 3 mortgage-servicing guidance further confirms that the relevant mortgage-forbearance problem was already within the formal regulatory response by the time the complaint alert became operationally confirmable. The later May 21 CFPB complaint bulletin shows that complaint narratives were useful for rapid contemporaneous surveillance, but it is not an appropriate benchmark for claiming that the complaints predicted the pandemic itself.

The secondary themes reinforce this caution. Overdraft/NSF fees generate alerts in only 36 of 81 specifications (44.4%) and do not alert under the primary specification. Twenty-seven of the 36 overdraft

alerts arise under one-month persistence, only six under two-month persistence, and three under three-month persistence; this pattern is threshold-sensitive rather than a robust persistent signal. Moreover, the modal onset is January 2023, well after the CFPB's December 2021 overdraft/NSF research and subsequent 2022 policy attention. Medical debt and several other themes trigger under still smaller fractions of the grid. These are negative boundary results: they show that the framework does not label every visible trend as a stable emergent theme and that post-recognition increases should not be relabeled as early warnings.

7. Robustness, Bias, and Measurement Threats

Taxonomy robustness is non-negotiable. Any analysis that ignores the August 2023 credit-card/prepaid-card split introduces an artificial structural break into both volume and text-composition series.

Duplicate handling matters because 14,802 excess exact duplicate narratives occur in the original extract. The primary text analysis therefore operates on unique narrative strings. This prevents identical copied text from mechanically multiplying a theme signal, but it may also down-weight genuinely repeated consumer experiences. The correct interpretation is robustness-oriented rather than ontological: no duplicate policy can perfectly distinguish repeated harm from repeated wording.

Short-narrative thresholds can change topic discovery and vocabulary composition. The corpus contains thousands of narratives below common 10-, 15-, and 25-word cutoffs. Because the present surveillance rule depends on explicit phrases rather than latent topics, the primary analysis retains short narratives rather than silently removing them.

The largest remaining measurement threat is selection into the complaint corpus itself. The series can change because the underlying incidence of harm changes, because complaint propensity changes, because consumers learn about the CFPB, because access to the web channel changes, or because firms and regulators alter the complaint process. The paper therefore uses the term 'complaint-based signal' and does not estimate population prevalence or causal effects.

8. Discussion

The empirical result is intentionally less dramatic than an 'AI predicts financial risk' narrative. A simple, auditable surveillance rule detects the sample's dominant shock-related language, but the persistence rule makes the real-time confirmation date later than the first visible rise. That distinction is not semantic: backdating confirmation to the onset month would mechanically exaggerate lead time. The corrected timing therefore strengthens the central boundary claim that robust emergence detection is not equivalent to prediction.

The taxonomy finding is arguably the most generalizable result for empirical work using administrative text. Longitudinal NLP studies frequently treat labels as ground truth while focusing methodological attention on embeddings or clustering. Here, a product-form redesign would create a larger apparent

discontinuity than many substantive market events. Measurement-system versioning should therefore be part of model governance for complaint intelligence.

The robustness grid provides a decision-support principle rather than a search for the most favorable parameterization. Signals that appear only under one-month persistence or a narrow subset of baseline choices should be treated as exploratory. COVID-related language is specification-insensitive; mortgage forbearance is moderately robust; overdraft/NSF and other secondary themes are materially more threshold-dependent. The paper therefore reports the entire grid rather than selecting the specification that produces the earliest date.

Finally, the negative early-warning finding is substantive rather than a failed prediction exercise. It identifies a necessary validation standard: future work that seeks an early-warning claim should prespecify an event universe, recognition rule, matching procedure, and null distribution before observing the complaint alerts. Without those elements, retrospective lead-time claims remain vulnerable to hindsight and event-selection bias.

9. Practical Implications

A defensible complaint-surveillance workflow should preserve original taxonomy labels; maintain a versioned harmonization layer; deduplicate or otherwise weight repeated text transparently; calculate alerts from historical information only; require minimum support, abnormality, multiplicity control, and persistence; report alert stability across reasonable specifications; and distinguish 'emergence detected' from 'advance warning validated.' This design is suitable as a research prototype for regulatory or financial-services analysts because every alert can be traced back to a visible counting and testing rule.

10. Limitations

The study has important limits. First, the complaint corpus is selected and web-only in the supplied extracts, so results do not estimate population harm. Second, lexicon themes trade semantic coverage for auditability; paraphrases outside the lexicons can be missed, while ambiguous phrases can create false positives. Third, the Poisson screening statistic does not fully model overdispersion, seasonality, dependence, or shifts in the monthly exposure population, so statistical significance should not be read as a generative-model claim. Fourth, the 81-specification grid is exploratory rather than preregistered. Fifth, external dates are interpretation benchmarks rather than a prospectively defined event universe; accordingly, the study does not perform or claim a confirmatory placebo lead-time test. Sixth, no independent human-rating study was conducted. Finally, the 2019–2023 window does not address more recent changes in complaint writing, including potential LLM-assisted narratives.

11. Conclusion

CFPB complaint narratives provide useful longitudinal signals, but reliable detection should not be conflated with prediction. After repairing a documented 2023 product-taxonomy discontinuity and

removing exact duplicate text for the primary surveillance series, a transparent emergence rule identifies persistent COVID-19/pandemic and mortgage-forbearance runs beginning in March 2020. Under the primary two-month persistence rule, however, those alerts are not operationally confirmed until April 2020, after broad public recognition of the pandemic and after enactment of federal mortgage-forbearance protections. The two themes remain the most stable across 81 reasonable threshold specifications, whereas overdraft/NSF and other secondary themes are substantially less stable. The strongest defensible contribution is therefore a taxonomy-aware, leakage-conscious surveillance framework plus a negative boundary result: complaint-based emergence can be robust without constituting advance warning.

Declarations

Data and code availability. The underlying source is the public CFPB Consumer Complaint Database. The submission package includes the analysis scripts and derived outputs used to reproduce the reported counts, monthly series, emergence tests, robustness summaries, and figures. Raw CFPB data are not redistributed in the journal package because the public source can be retrieved directly and complaint narratives may contain sensitive text.

Ethics and privacy. No participants were recruited, and no intervention was conducted. The study analyzes publicly released administrative complaint data. The manuscript does not reproduce individual complaint narratives or personally identifying text.

Funding and competing interests. Funding, author-contribution, and competing-interest information is supplied through the journal's submission interface in accordance with its double-anonymous review workflow.

AI Use Statement: AI-assisted tools were used only for grammatical support, formatting support, and organization of the manuscript. All analyses reported here were executed on the supplied data, numerical outputs were checked against generated result files, cited sources were independently verified, and the author(s) retain full responsibility for the manuscript and code.

References

1. Bastani K, Namavari H, Shaffer J (2019) Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Syst Appl 127:256–271. https://doi.org/10.1016/j.eswa.2019.03.001
2. Consumer Financial Protection Bureau (2020) Complaint Bulletin: Complaints mentioning coronavirus keywords. Consumer Financial Protection Bureau, May 21, 2020
3. Consumer Financial Protection Bureau (2021a) Complaint Bulletin: Mortgage forbearance issues described in consumer complaints. Consumer Financial Protection Bureau, May 4, 2021

4. Consumer Financial Protection Bureau (2021b) Data Point: Overdraft/NSF Fee Reliance Since 2015 - Evidence from Bank Call Reports. Consumer Financial Protection Bureau, December 1, 2021
5. Consumer Financial Protection Bureau (2022) Medical debt burden in the United States. Consumer Financial Protection Bureau, March 1, 2022
6. Consumer Financial Protection Bureau (2026) Consumer Complaint Database and product/issue documentation. Accessed September 2026
7. Gao P, Sun N, Wang X, Yang C, Zitikis R (2024) Natural language processing-based detection of systematic anomalies among the narratives of consumer complaints. J Oper Risk 19(2):79–104. https://doi.org/10.21314/JOP.2024.004
8. Sangaraju VR, Bolla BK, Nayak DK, Kh J, Based Embeddings (2022) Topic Modelling on Consumer Financial Protection Bureau Data: An Approach Using BERT. 2022 IEEE 7th International Conference for Convergence in Technology (I2CT), pp 1–6. https://doi.org/10.1109/I2CT54291.2022.9824873
9. Shin M, Kim J, Shin J (2026) The adoption and efficacy of large language models in US consumer financial complaints. Nat Hum Behav 10:669–680. https://doi.org/10.1038/s41562-026-02409-4

Figures

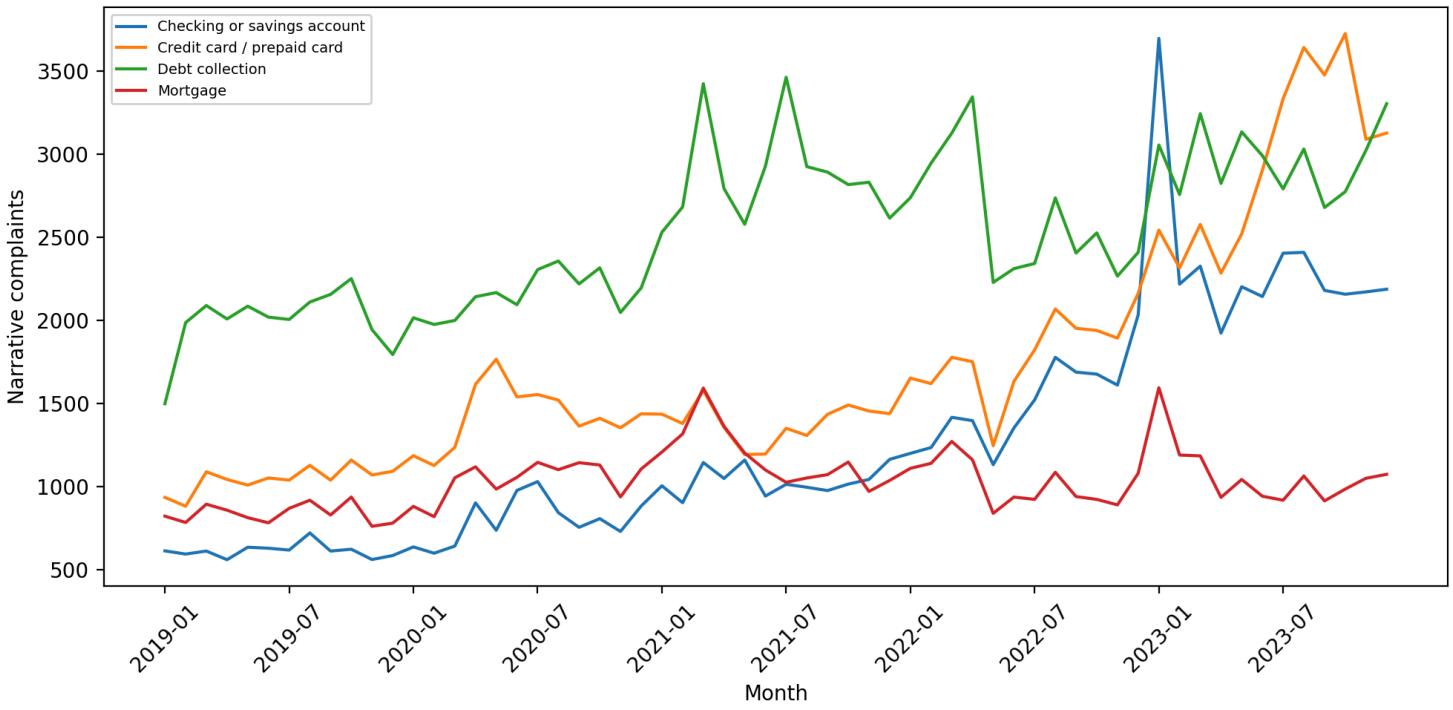


Figure 1

Monthly narrative complaint volume by harmonized product family, 2019–2023.

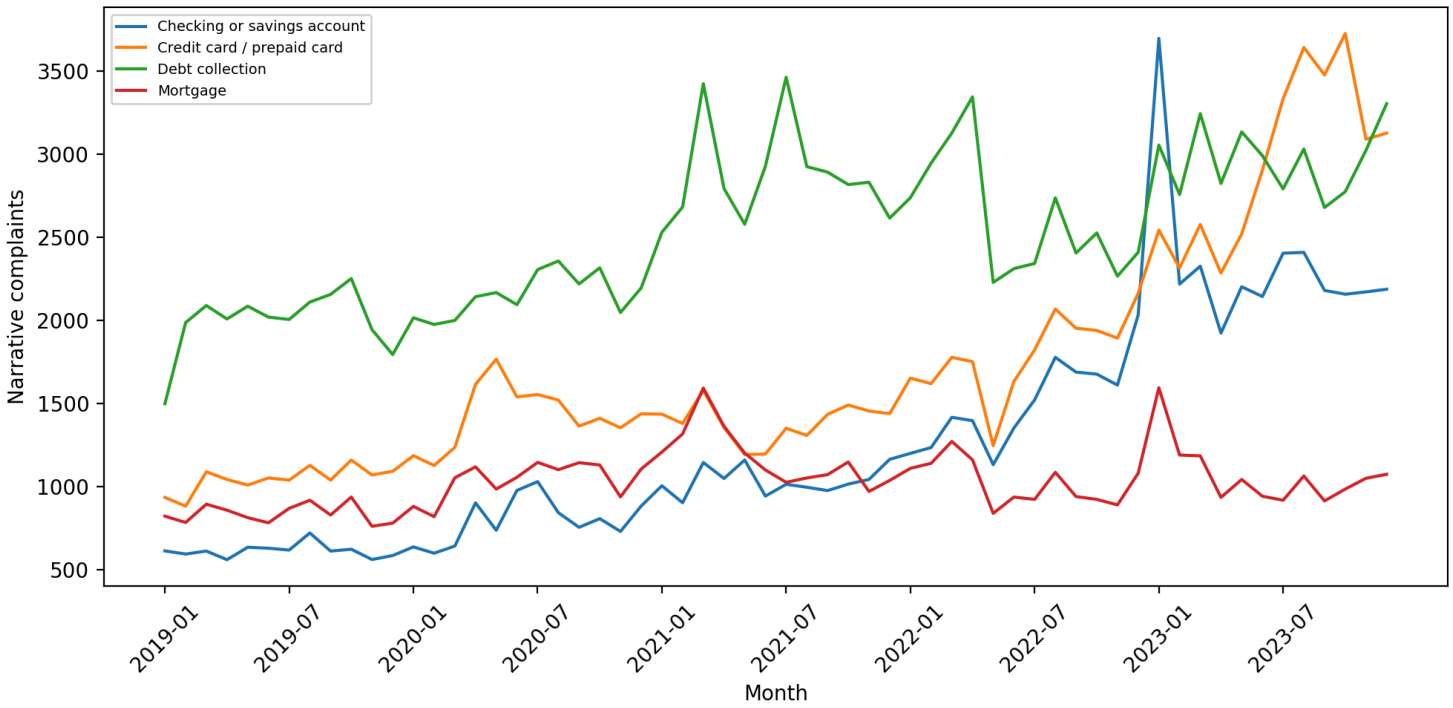


Figure 1

Monthly narrative complaint volume by harmonized product family, 2019–2023.

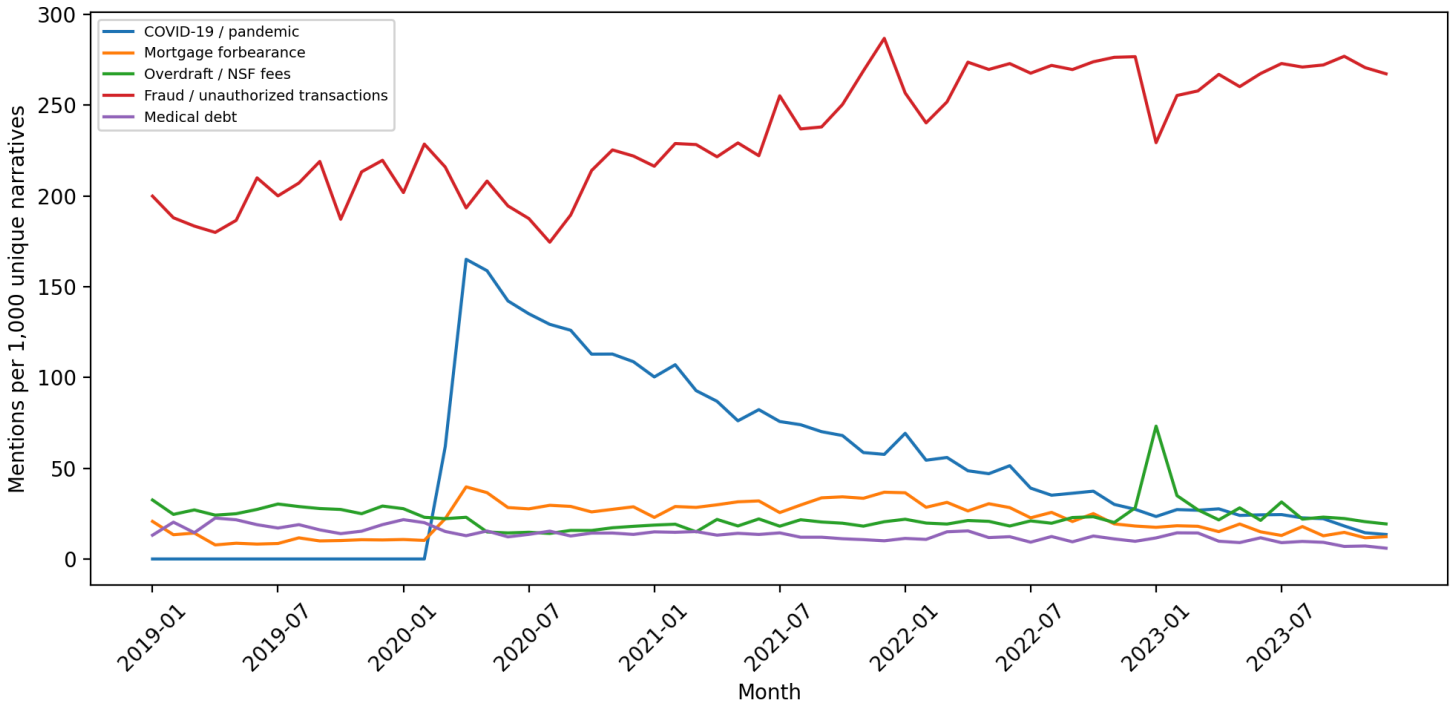


Figure 2

Selected theme prevalence per 1,000 unique narratives.

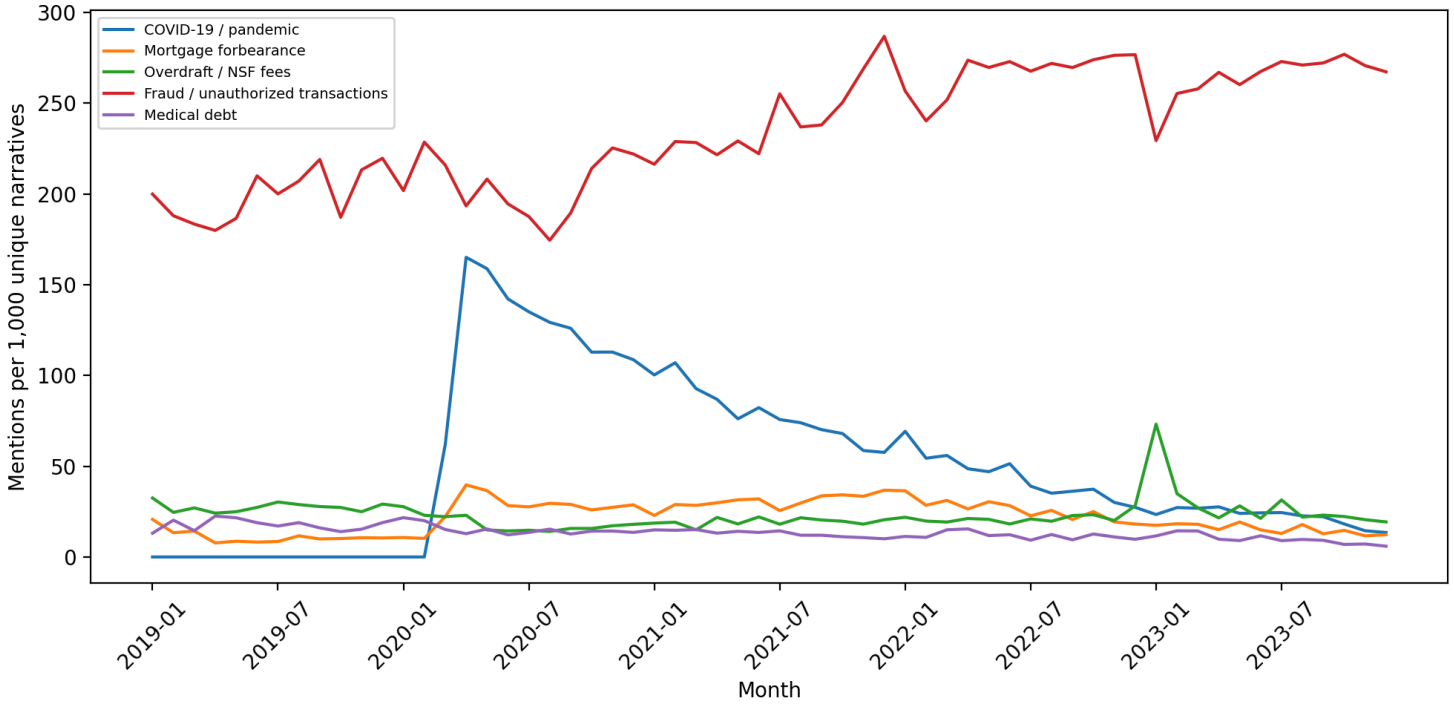


Figure 2

Selected theme prevalence per 1,000 unique narratives.

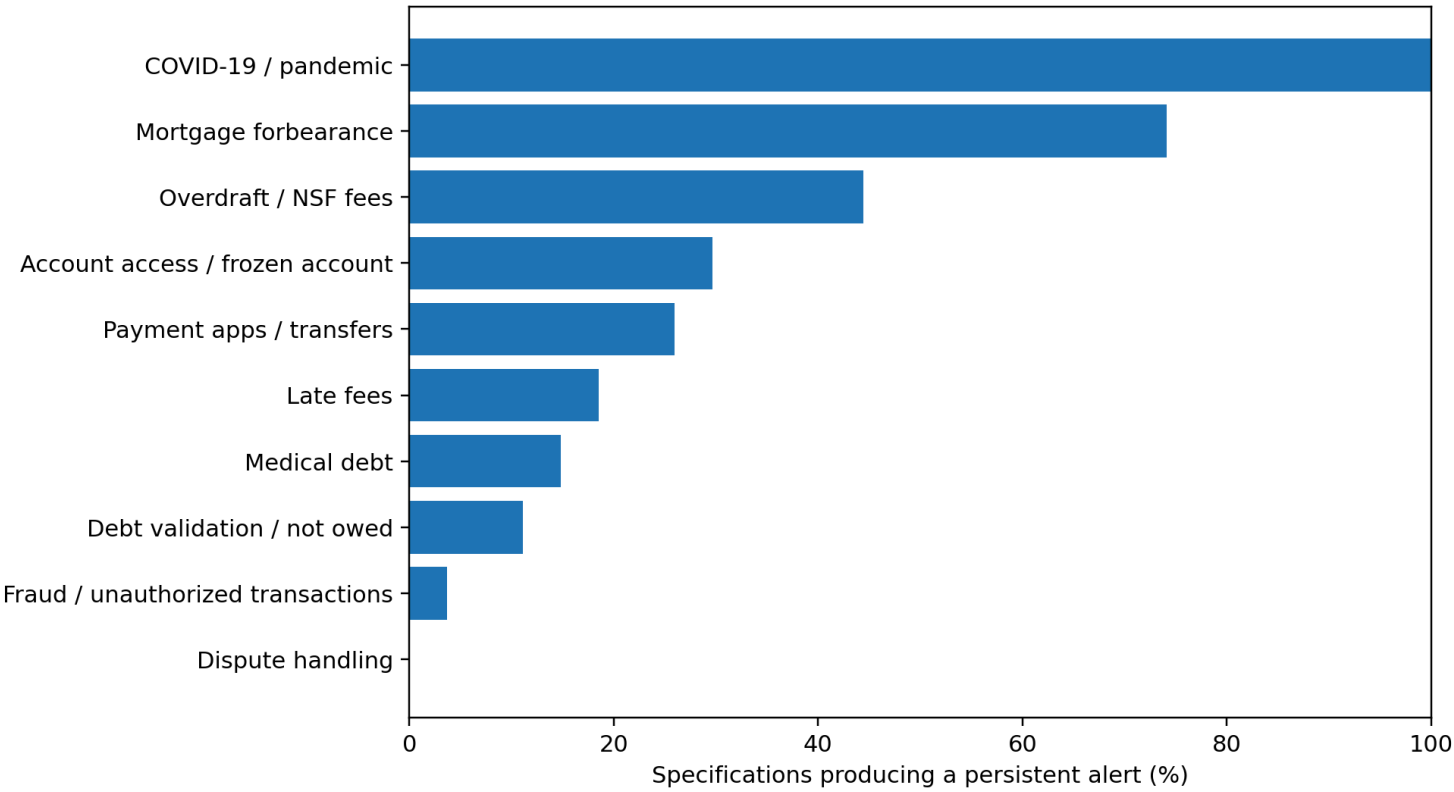


Figure 3

Fraction of 81 statistical specifications that generate a persistent alert for each theme.

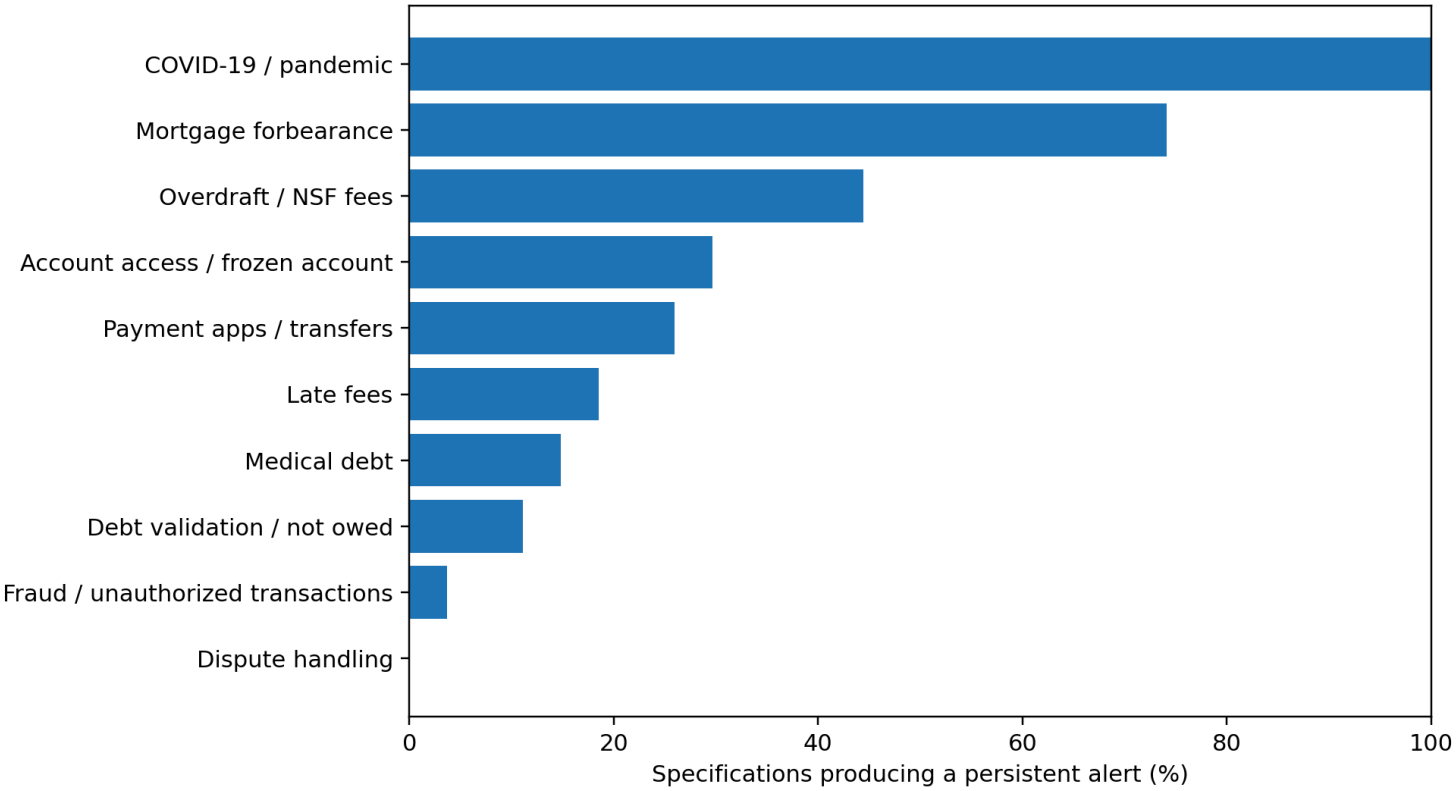


Figure 3

Fraction of 81 statistical specifications that generate a persistent alert for each theme.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixA.docx
- AppendixA.docx