Reverse-Engineering Survey Data from Public Social Discourse: A Computational Methodology for Large-Scale Opinion Mining

Published: 06 April 2026 by The FamilyBond Team

> ***Disclosure:*** *This is not a formal academic paper and has not been submitted to or published in any peer-reviewed journal or academic conference. The authors are not affiliated with any academic institution. This document is a detailed methodology write-up created to transparently describe and document the process used in a self-initiated research project. It is structured in an academic format for clarity and reproducibility, but it should not be cited or treated as peer-reviewed research.*

Abstract

Traditional survey methods constrain data collection through response bias, recruitment costs, small sample sizes, and point-in-time temporal limitations. This paper presents a generalizable computational methodology for reverse-engineering survey-equivalent quantitative data from organic public discourse in online topic-focused communities. The method combines AI-assisted source identification (Perplexity) and large-scale data collection (Apify) with a multi-pass extraction pipeline featuring regex-based numeric extraction, demographic inference, sentiment classification, and severity assessment. The methodology was validated and operationalized on digital parenting discourse as a proof of concept, assembling a corpus of 315,246 items (12,136 posts and 303,110 comments) spanning May 2025 through January 2026 from public discussion forums and community platforms. Across a sample first-phone age survey (N=2,543), the method recovered quantitative estimates with standard errors, confidence intervals, and effect sizes comparable to traditional survey research. Validation steps include multi-pass regex extraction with contextual filtering, plausibility range constraints (ages 4–18, screen time 0–12h), manual spot-checking, deduplication, sentiment classification within windowed contexts, severity coding on 3-point scales, and gender inference with documented accuracy metrics (~85–90% for name-based inference). The methodology demonstrates the feasibility of extracting survey-equivalent data from naturalistic online discourse and proves applicable to any domain where people publicly discuss experiences, behaviors, or opinions. This approach provides researchers with a scalable, ecologically valid complement to traditional survey methods.

Keywords: computational social science, social media mining, survey methodology, natural language processing, reverse-engineered survey, public discourse analysis

---

1. Introduction

Understanding public attitudes, behaviors, and concerns across diverse domains—from health and parenting to politics and consumer behavior—remains central to social science research. Traditional survey methodology, established through decades of refinement (Dillman, 1999; Tourangeau et al., 2012), provides the primary tool for capturing self-reported attitudes and behaviors at scale. However, conventional survey research carries well-documented limitations: response bias and social desirability effects systematically distort self-reported data; recruitment constraints limit sample sizes and statistical power; and cross-sectional designs capture only point-in-time snapshots rather than evolving norms.

Against these limitations, a potential complementary data source has emerged at scale: the organic discourse of millions of individuals discussing their experiences, concerns, and decisions in publicly accessible online communities. Unlike survey respondents, who answer researcher-generated questions on researcher-selected topics, people in public discourse communities self-select what to discuss and how to frame their experiences. This organic self-initiation may reduce social desirability bias while capturing genuine, urgent concerns that motivated communication in the first place.

The question addressed in this paper is fundamental: can survey-equivalent quantitative data be reliably extracted from organic online discourse, and can such extraction be performed using systematic, reproducible methodology applicable across diverse topics? This paper presents a comprehensive answer through a methodological approach termed "reverse-engineered survey construction from social discourse."

The contributions of this paper are threefold: (1) a systematic, reproducible methodology for constructing survey-equivalent datasets from unstructured social media text, operationalized and tested on a complete data collection pipeline; (2) detailed specification of validation and quality-control steps performed during actual research but rarely fully documented in published work; and (3) demonstration that this approach recovers quantitative estimates with statistical properties (effect sizes, confidence intervals, significance tests) comparable to conventional survey research.

---

2. Related Work

2.1 Survey Methodology and Its Limitations

Traditional survey methodology has been extensively documented in foundational work by Dillman (1999) and Tourangeau et al. (2012), establishing best practices for question design, sampling, and data quality. However, systematic limitations persist. Social desirability bias—the tendency for respondents to answer questions in ways they believe align with social norms—consistently inflates reports of socially valued behaviors while depressing reports of devalued ones (Tourangeau et al., 2012). Response rates have declined substantially over recent decades, raising questions about representativeness. Most critically, survey research is inherently cross-sectional; detecting norm evolution requires repeated surveys at substantial cost.

Computational approaches to opinion measurement offer an alternative paradigm. De Choudhury et al. (2013) demonstrated that aggregated patterns in social media discourse could serve as proxies for population-level attitudes in health and mental health domains. Lazer et al. (2009) argue for the integration of computational approaches with traditional survey methods, noting that big-data approaches can provide unprecedented sample sizes and temporal granularity. O'Connor et al. (2010) validated Twitter-based measurement of public opinion against conventional polling data, finding moderate-to-strong convergence across major political and social topics.

2.2 Natural Language Processing for Extraction from Social Text

Structured extraction of quantitative information from unstructured text has advanced substantially through developments in both regex-based pattern matching and neural language models. Jurafsky and Martin (2026) provide foundational work in information extraction; more recent work has focused on domain-adapted extraction pipelines (Finkel et al., 2005). The specific challenge of numeric value extraction from social media—recovering ages, frequencies, and quantities from informal text—has been addressed in computational linguistics through both rule-based (Eisenstein, 2013) and learning-based approaches (Ritter et al., 2011).

For the present work, regex-based extraction is preferred over neural approaches because: (1) the target patterns (numeric values, specific keywords, gendered pronouns) are simple and highly regular; (2) interpretability is essential for validation and debugging; and (3) regex patterns transfer reliably across similar domains without retraining.

2.3 Demographic Inference from Text and Metadata

Inference of demographic characteristics (gender, age, location) from social media data has been extensively studied. Burger et al. (2011) established that name-based gender classification achieves 85–90% accuracy on standard lexicons. Volkova et al. (2015) showed that combinations of linguistic and metadata-based signals can improve accuracy beyond name alone. Importantly, demographic inference introduces uncertainty that must be

explicitly quantified and reported; the present paper models this by documenting inference accuracy and reporting non-response rates for inferred variables.

2.4 Sentiment and Concern Classification in Social Discourse

Classification of sentiment, emotion, and concern-related language in social text has become a standard NLP task. Zhang & Liu (2016) provides comprehensive resources for lexicon-based sentiment analysis. Domain-adaptation of sentiment models requires calibration to domain-specific terminology and framing norms, as general-purpose sentiment lexicons may misclassify domain-specific language.

The multi-pass architecture deployed in the present study (broad pass → contextual validation → domain-specific extraction) follows the established pipeline approach documented in the computational social science literature (Grimmer & Stewart, 2013).

---

3. Methodology

3.1 Data Collection and Corpus Construction

Source channel identification was performed using Perplexity (an AI-powered research aggregation tool) to identify publicly accessible online communities with organic discourse relevant to the target domain. Selection criteria included: (1) community size sufficient for statistical analysis; (2) topical focus aligned with the research domain; (3) public accessibility requiring no authentication for read access; and (4) active discussion enabling temporal analysis.

Data collection was executed using Apify (a web scraping and automation platform) configured for scalable, distributed collection from the identified communities. All data were retrieved from publicly accessible content in open communities. The collection window for the proof-of-concept validation spanned recent data (May 2025 through January 2026 for the primary corpus), with extended historical data (extending to December 2018) collected to enable temporal trend analysis.

The resulting corpus comprised 315,246 total items disaggregated as 12,136 posts and 303,110 comments, distributed across two classes of public online communities: (1) a membership-based community platform (from which 9,302 posts and associated embedded comments were retrieved), and (2) open discussion forum communities spanning multiple topic-focused areas (from which approximately 335,000 items were collected across three batches). Table 1 summarizes the corpus structure.

Table 1. Corpus Structure by Source Type

Source Type	Items	Posts	Comments	Date Range	Notes
Community platform	~9,302	~9,302	Embedded	2025–2026	Membership-based, persistent identities
Discussion forum — small batch	329	—	—	2026-03-30	Test collection
Discussion forum — curated communities	10,999	—	—	2024–2026	Topic-specific sections
Discussion forum — large corpus (recent)	295,616	—	—	2025–2026	Primary analysis window
Discussion forum — large corpus (historical, pre-2025)	29,272	—	—	2018–2024	Historical trend analysis; may contain temporal artifacts and platform glitches; used only for trend estimation
Total	**~345,518**	**12,136**	**303,110**	**2018–2026**	—

3.2 Ethical Considerations

All data were sourced exclusively from publicly accessible communities, defined as those whose content is readable without account authentication and whose terms of service permit research use of public content. No data were collected from private communities or content requiring authentication. Personally identifiable information was minimized through algorithmic processing of usernames for demographic inference without retaining identifiers in analytic outputs. Author names and display names were used solely for gender inference and were not retained in datasets distributed or analyzed post-collection. No verbatim quotes from community discourse are reproduced in this paper; all examples are paraphrased descriptions of discourse patterns.

The study constitutes analysis of pre-existing public text and does not involve human subjects in the regulatory sense. Nevertheless, the ethical principles governing minimization of privacy risk were applied throughout, consistent with best practices in computational social science.

3.3 Topic Identification and Validation Pipeline

Topic identification and validation followed a multi-pass pipeline designed to maximize precision in extracting survey-relevant signals while minimizing false positives from topically adjacent but analytically irrelevant content.

Pass 1 — Broad keyword matching: Each item in the corpus was scanned against a library of predefined domain-relevant keywords (exactly which keywords depend on the target survey domain—for the proof-of-concept digital parenting application, keywords included terms related to device usage, age, and parental decision-making). Items matching at least one keyword trigger were retained for further processing. This pass prioritizes recall over precision, accepting false positives to minimize false negatives.

Pass 2 — Contextual validation: Retained items were evaluated against contextual validation rules designed to eliminate common false positive patterns. For example, when extracting reported behaviors from discourse, the validation rules distinguish between: (a) first-person parental experience reports ("I allow my child..."); (b) observed usage reports that lack parental framing ("she spent five hours..."); (c) third-party references lacking first-person parental framing; and (d) technical or off-topic mentions. Only pattern (a) is retained for extraction. This pass dramatically reduces false positives by requiring contextual alignment.

Pass 3 — Domain-specific extraction: Validated items are routed to domain-specific extraction modules corresponding to each survey topic. These modules apply targeted pattern matching to extract specific data elements (numeric values, categorical responses, sentiment signals) required for statistical analysis. The architecture is domain-agnostic: different target domains require different extraction rules, but the pipeline structure remains constant.

The multi-pass architecture was implemented as a multi-agent system in which specialized sub-agents were deployed in parallel across survey domains. This AI-assisted orchestration enabled concurrent processing of the full corpus against multiple extraction targets. A coordinating agent managed deduplication where items matched multiple topics.

3.4 Data Extraction and Validation Steps

The following validation and extraction steps were performed on all extracted values. These steps are detailed here because they represent quality-control practices that are essential to the methodology but are often incompletely documented in published work.

Multi-pass regex extraction with contextual validation: Numeric values (ages, hours, frequencies) were extracted using targeted regex patterns. Extraction was performed on multiple passes against different pattern variants to capture natural language variability (e.g., "12 years old," "12-year-old," "age 12," "when he was 12"). A secondary contextual validation pass confirmed that extracted numeric values appeared in sentences containing appropriate semantic markers, reducing false positives.

Plausibility range filtering: Only values falling within domain-specific plausibility ranges were retained. For child ages: 4–18 years. For screen time: 0–12 hours per day. For ages at life events: ranges calibrated to age-appropriate transitions. Values outside these ranges were excluded, eliminating obvious extraction errors.

Manual spot-checking of extracted values against source text: Approximately 10% of all extracted records were manually reviewed by independent annotators who verified that: (1) the extracted value appeared in the source text; (2) the extracted value was in the claimed semantic context; and (3) the value represented a plausible report. Disagreements triggered revision of the extraction rules. This step is computationally expensive but is essential for validating the entire extraction pipeline.

Deduplication across survey domains: Where items matched multiple extraction targets (e.g., a post mentioning both screen time limits and first-phone age), individual claims were

separated and tracked distinctly, with duplicate identical claims removed but related claims retained. Deduplication logic was domain-specific.

Filtering observed usage reports versus stated limits: For extraction of behavioral limits or norms, discourse describing observed child behavior (e.g., "he plays for 4 hours daily") was distinguished from discourse describing parental intention or policy ("I limit screen time to 1 hour"). Only the latter was retained for survey-equivalent analysis, as the former reflects observed behavior conditional on parental enforcement.

Filtering third-party references lacking first-person parental framing: References to other people's parenting practices, media reports of statistics, or clinical descriptions were excluded. Only first-person parental reports were retained.

Grade-level to age mapping with conservative midpoint estimates: Where discourse referenced school grades rather than ages, grades were mapped to age using standard U.S. educational norms (Kindergarten=5, Grade 1=6, ... Grade 12=18). When grade ranges were ambiguous (e.g., "middle school"), the conservative lower midpoint was used (Grade 6=12 years).

Sentiment classification using domain-adapted lexicon within windowed contexts: Sentiment was classified on a three-point scale (positive, neutral, negative/concern-laden) using a curated lexicon of domain-relevant positive and negative terms, evaluated within a ±150 to ±200 character context window surrounding each target mention. The window ensures that sentiment is classified in appropriate context rather than based on isolated words.

Severity classification on 3-point scale using escalatory linguistic markers: For concern-related extraction, severity was coded as Mild (vague or hypothetical concern), Moderate (described active worry or behavioral response), or Severe (described crisis, trauma, law enforcement involvement, or mental health services). Assignment followed presence of escalatory linguistic markers in the surrounding context (e.g., "my child was arrested," "my child attempted suicide," "I took my child to therapy").

Gender inference validation: Parent gender was inferred from display names using a curated lexicon of 300+ common English given names, achieving approximately 85–90% accuracy for names in the lexicon. Accounts using pseudonymous usernames (alphanumeric identifiers) were coded as gender-unknown. Child gender was inferred from gendered relational nouns ("my son," "my daughter") and pronouns in context. Inference accuracy is documented; non-response rates are reported explicitly.

Extra manual editing pass: Following automated extraction, all extracted values were reviewed in a second manual pass to identify systematic errors in the automated pipeline that might not be caught by spot-checking alone. This step identifies patterns (e.g., a regex variant that systematically extracts values three times too large) that can be corrected before analysis.

3.5 External Validation, Gap Analysis, and Survey Disposition

A critical step in the methodology—often omitted in computational social science pipelines—is comparison of reverse-engineered results against published external survey data on the same topics. For each survey domain, results were benchmarked against established surveys (e.g., Pew Research, Common Sense Media, national health surveys) to identify alignment and divergence.

Where reverse-engineered estimates aligned with published benchmarks within expected margins, confidence in the extraction pipeline was reinforced. Where significant gaps were identified between the reverse-engineered data and external survey results, a structured gap analysis was performed to determine whether the divergence could be explained by known methodological factors such as self-selection bias, community-specific norms, or differences in question framing.

Gaps that could be plausibly attributed to identifiable methodological factors were documented with explanations and retained as valid findings with appropriate caveats. However, in cases where the gap between reverse-engineered results and external benchmarks could not be satisfactorily explained—suggesting possible systematic extraction errors, unresolvable confounds, or insufficient data quality—the survey results were dropped entirely rather than published with unexplained discrepancies. In the proof-of-concept application, one of five survey domains (daily screen time limits) was dropped through this process because the extracted values diverged substantially from established benchmarks in ways that could not be attributed to known selection effects.

This disposition step is methodologically important: it demonstrates that the methodology includes built-in quality gates that prevent publication of unreliable results, and it distinguishes this approach from purely automated pipelines that lack human judgment checkpoints.

3.6 Statistical Analysis Framework

Statistical analyses were conducted independently for each survey domain using two-tailed tests at a significance threshold of $\alpha=.05$. Effect sizes are reported for all significant tests. Confidence intervals are reported at the 95% level. P-values at or below .001 are reported as $p<.001$; values above this threshold are reported to three decimal places. The following test types were employed:

- **Chi-square goodness-of-fit tests** (χ^2) to assess whether observed distributions departed from uniformity
- **Chi-square tests of independence** to assess associations between categorical variables, with Cramer's V as the effect size measure
- **One-way ANOVA** for continuous measure comparisons across three or more groups, with eta-squared (η^2) as effect size
- **Welch's two-sample t-tests** for continuous comparisons between two independent groups, with Cohen's d as effect size

- **Linear regression** for temporal trend analysis, with R^2 as the proportion of variance explained
- **Binomial tests** for comparisons of observed proportions against specified null values
- **Spearman rank correlation** for ordinal associations

---

4. Proof-of-Concept Validation: First-Phone Age Survey (N=2,543)

To demonstrate the methodology, we present results from one survey domain—age of first phone ownership—derived from the digital parenting discourse corpus. This survey serves as the primary methodological validation case.

4.1 Data Recovery and Distributional Properties

The first-phone survey extracted 3,376 records containing phone-age mentions from the full corpus of 345,518 items. Of these, 2,543 records (75.3%) contained a specific child age, enabling full quantitative analysis. The remaining 833 records (24.7%) contained phone-type and reason information without specific ages, retained for qualitative pattern description.

The 2,543 age-specific records exhibited a mean age of M=11.77 years (SD=3.12, Mdn=12.0, 95% CI [11.65, 11.89], range 4–18 years). The distribution was non-normal per the D'Agostino-Pearson K^2 test, K^2=139.47, df=2, p<.001, exhibiting a bimodal structure with primary clusters at age 12 (N=420, 16.5%) and age 14 (N=495, 19.5%), reflecting documented developmental transitions (middle school entry and high school entry).

4.2 Demographic Subgroup Comparisons

No significant gender difference in first-phone age was found when comparing sons (M=11.83 years, N=484) to daughters (M=11.83 years, N=497), t(976)=-0.023, p=.982, d=0.001. The effect size is essentially zero.

Comparison of inferred parent gender (male: M=12.05 years, N=40; female: M=11.79 years, N=406) similarly yielded no significant difference, t(47)=0.570, p=.572, d=0.095, though note that only 46 records (9.2% of the gender-inferred sample) were coded as male-authored due to the prevalence of pseudonymous accounts.

4.3 Temporal Trend Analysis

Analysis of 18 monthly periods with N≥3 observations, spanning December 2018 through March 2026, revealed a significant positive trend in mean first-phone age over time, β=+0.096 years/month, r=0.905, R^2=0.819, t(16)=8.513, p<.001. This corresponds to

approximately +1.15 years per calendar year. The high R^2 (0.82) indicates that time explains 82% of the variance in monthly mean first-phone age, demonstrating that temporal variation is robust rather than noise-driven.

This finding demonstrates a critical advantage of the reverse-engineered methodology: a single historical corpus enables retrospective trend analysis spanning years. Traditional survey research would require repeated cross-sectional surveys at substantial cost to detect comparable temporal patterns.

---

5. Discussion

5.1 Methodological Strengths Relative to Traditional Surveys

The reverse-engineered survey methodology offers several empirical advantages over conventional survey approaches, demonstrated in the present study:

Sample size: The first-phone survey alone (N=2,543) exceeds most published survey studies on comparable topics. Equivalent sample sizes through traditional survey research would require substantial recruitment costs and time investment.

Reduced social desirability bias: Because posts are self-initiated rather than response-induced, they reflect genuine concerns that motivated communication. While participants still engage in self-presentation, the content reflects actual preoccupations rather than responses to researcher-imposed framings.

Temporal granularity: A single historical corpus enables trend analysis across years. The present study recovered a temporal trend (R^2=0.82) from a seven-year corpus. Achieving equivalent temporal data through traditional surveys would require longitudinal data collection spanning the same period.

Multi-dimensional coding: The organic multi-label structure of social media posts enables simultaneous coding across multiple dimensions (demographics, sentiment, concern type) from single records, providing analytical richness that would require substantially longer traditional survey instruments.

Ecological validity: Data reflect naturally occurring discourse in the environments where people actually seek information and exchange experiences, rather than artificial survey contexts.

5.2 Transferability to Other Domains

The methodology presented here is not specific to digital parenting discourse. The same pipeline architecture can be applied to any domain where people publicly discuss experiences, behaviors, concerns, or opinions. Potential application domains include:

- **Health behavior and medical concerns:** Discourse in health forums discussing medication side effects, disease management, or treatment decisions
- **Political attitudes and political behavior:** Public discourse on political social media discussing voting intentions, policy preferences, or political concerns
- **Consumer behavior:** Product review forums discussing purchase decisions, product satisfaction, and consumer concerns
- **Workplace experience:** Employee discussion forums addressing job satisfaction, work conditions, and workplace concerns
- **Financial decision-making:** Investment and personal finance forums discussing financial choices and confidence in economic conditions

The domain-specific adaptation required is limited to Pass 1 keyword selection, contextual validation rules, and extraction target specifications. The overall pipeline structure remains constant.

5.3 Limitations

Self-selection bias: Participants in online discourse communities are not representative of broader populations. They are typically more engaged, more tech-literate, and often more concerned with the domain topic. Without validation against representative survey samples, the magnitude of selection bias is unknown. Reverse-engineered estimates should be interpreted as characterizing the engaged online community, with caution about generalizing to broader populations.

Demographic inference uncertainty: Parent and child gender are inferred from linguistic markers rather than directly reported. Name-based inference achieves 85–90% accuracy for names in lexicons but performs at chance for pseudonymous accounts. Gender-unknown non-response rates (79–84% across surveys) must be explicitly reported and limit gender-stratified analyses.

Temporal window and seasonal effects: The primary corpus spans nine months; seasonal effects are not controlled. Monthly sample sizes for some analyses are small, limiting monthly-level statistical power.

Platform-specific norms: Different online communities exhibit different discourse norms, as evidenced by documented differences between community types. These norms influence both what topics are discussed and the register in which they are discussed.

Repeated respondents: Individual community members may contribute multiple posts; each post is treated as an independent observation. Respondent duplication varies across communities and collection windows.

Language coverage: Data collection was restricted to English-language communities. Findings are not generalizable to non-English-speaking populations, which may differ substantially in norms and concerns.

---

6. Future Work

The methodology presented here opens several avenues for extended research:

Application to additional domains: The most direct extension is operationalizing this methodology across diverse domains—health, politics, consumer behavior, workplace experience—to establish the breadth of domains where reverse-engineered surveys are feasible and valuable.

Formalized convergent validity testing: The present study compared reverse-engineered results against published external surveys and dropped domains with unexplained divergence. A more formalized next step would be administering purpose-built probability-sampled surveys on identical topics to representative samples simultaneously with reverse-engineered collection, enabling direct quantification of selection bias magnitude and establishment of systematic correction factors.

Cross-cultural comparison: Extending the methodology to non-English-language communities would enable cross-cultural comparison of norms and concerns across domains, revealing whether patterns are universal or culturally specific.

Real-time monitoring dashboards: The capacity for near-real-time data collection suits this methodology to continuous monitoring architectures, providing stakeholders (policymakers, practitioners, platform safety teams) with updated views of evolving public concerns and emerging topics.

Integration with outcome data: Linking discourse patterns to behavioral outcomes through cohort studies or natural experiments would address whether behaviors and attitudes discussed in online discourse are associated with documented outcomes.

---

7. Conclusion

This paper has presented a systematic, reproducible methodology for reverse-engineering survey-equivalent quantitative data from organic public discourse in online topic-focused communities. The approach combines AI-assisted source identification (Perplexity) and

scalable data collection (Apify) with a multi-pass extraction pipeline featuring explicit validation steps, regex-based numeric extraction, demographic inference with documented accuracy, sentiment and severity classification, and rigorous statistical analysis.

The methodology was validated through proof-of-concept application to digital parenting discourse, recovering quantitative estimates on first-phone age (N=2,543) with standard errors, confidence intervals, and effect sizes. Validation steps included multi-pass regex extraction, contextual filtering, plausibility range constraints, manual spot-checking (10% of extractions), deduplication, sentiment classification within windowed contexts, severity coding on 3-point scales, and gender inference with documented accuracy metrics.

The temporal trend recovered from this data (β=+0.096 years/month, R^2=0.82, $p<.001$) demonstrates a critical methodological advantage: a single historical corpus enables retrospective trend detection that would require years of traditional longitudinal survey research.

The demonstrated capacity to recover survey-equivalent data at sample sizes exceeding conventional research (N=2,543), with temporal resolution and analytical richness comparable to traditional surveys, and at substantially lower cost, suggests that reverse-engineered survey methodology represents a promising complement to established survey research. Critical conditions for implementation include rigorous reporting of selection bias limitations, explicit documentation of demographic inference accuracy, and validation through convergent validity studies against probability-sampled benchmarks.

References

Dillman, D. A. (1999). *Mail and Internet surveys: The tailored design method* (2nd ed.). John Wiley & Sons. https://www.amazon.com/Mail-Internet-Surveys-Tailored-Design/dp/0471323543

Tourangeau, R., Rips, L. J., & Rasinski, K. (2012). *The psychology of survey response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 47–56). https://doi.org/10.1145/2464464.2464480

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., ... Macy, M. (2009). Computational social science. *Science*, 323(5915), 721–723. https://doi.org/10.1126/science.1167742

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the*

International AAAI Conference on Web and Social Media, 4(1), 122–129. https://doi.org/10.1609/icwsm.v4i1.14031

Jurafsky, D., & Martin, J. H. (2026). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3rd ed.). Online manuscript. https://web.stanford.edu/~jurafsky/slp3/

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 363–370). https://doi.org/10.3115/1219840.1219885

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 359–369). https://aclanthology.org/N13-1037/

Ritter, A., Clark, S., Mausam, & Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1524–1534). https://aclanthology.org/D11-1141/

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1301–1309). Association for Computational Linguistics. https://aclanthology.org/D11-1120/

Volkova, S., Bachrach, Y., Armstrong, M., & Sharma, V. (2015). Inferring latent user properties from texts published in social media. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 4296–4297). AAAI Press. https://doi.org/10.1609/aaai.v29i1.9271

Zhang, L., & Liu, B. (2016). Sentiment analysis and opinion mining. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning and data mining*. Springer. https://doi.org/10.1007/978-1-4899-7502-7_907-1

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297. https://doi.org/10.1093/pan/mps028

Data collection was performed using Apify (apify.com). Source channel identification was performed using Perplexity (perplexity.ai). All statistical analyses were conducted using standard parametric and non-parametric test implementations. Analysis date: March 31, 2026.