

A Computational Text Analysis of Stakeholder Framing in the Swachh Bharat Mission

Name: Aahwaan Khullar
School: Shiv Nadar School, Faridabad
Year: 2026

Working under the mentorship of:
Mr. Gaurav Thapak, Consultant, Ernst & Young
Ms. Amrita Saib, Director, Ernst & Young

Abstract

The Swachh Bharat Mission (SBM), launched by the Government of India in October 2014, is one of the largest sanitation programmes ever undertaken, aiming to eliminate open defecation and improve solid waste management nationwide. In the years since, SBM has been described in strikingly different terms depending on who is doing the describing. This project asks a specific question about that disagreement: does the way SBM is discussed differ in measurable, systematic ways depending on the type of source doing the discussing, and depending on which stakeholder group a given piece of writing is centrally about?

To answer this, I adopt a text analysis approach using the VADER lexicon-based model, and topic modeling using Latent Dirichlet Allocation. The results show a clear, consistent gradient in sentiment by source type: policy briefings and government press releases are framed most positively, while opinion writing and investigative journalism are framed most negatively, with academic papers sitting almost exactly at neutral. At the stakeholder level, citizens and public health outcomes are framed most positively, and marginalized castes and a critically-framed general public are framed most negatively. The single most interesting finding, however, is what did not happen: despite the corpus containing multiple documents reporting hundreds of sanitation-worker deaths from hazardous sewer cleaning, sanitation workers as a stakeholder group scored as sentiment-neutral. This project argues that this reflects the flat, procedural register in which worker-safety reporting tends to be written, not the actual severity of what is being reported, a finding with implications for how much weight sentiment analysis alone should be given in social-impact research.

Keywords: Swachh Bharat Mission; sanitation policy; framing; sentiment analysis; VADER; topic modeling; Latent Dirichlet Allocation; stakeholder analysis; computational text analysis

Table of Contents

Abstract	2
1. Introduction	4
1.1 Research Questions	4
1.2 Intent of the paper	4
1.3 Structure of the Paper	5
2. Literature Review	5
2.1 From the Total Sanitation Campaign to SBM	5
2.2 Measuring Success: Construction, Use, and Health	6
2.3 Sanitation Workers, Caste, and Manual Scavenging	7
2.4 Fund Management, Verification, and Accountability	8
2.5 Gender, Menstrual Hygiene, and School WASH	9
2.6 Theoretical Framework	9
2.7 Computational Approaches to Framing and Sentiment	10
2.8 Research Gap	11
2.9 Expectations (What I Expected to Find)	11
3. Methodology	12
3.1 Overall Approach: Why Text Analysis	12
3.2 Corpus Construction	12
3.3 Preprocessing Pipeline	14
3.4 Sentiment Analysis	15
3.5 Topic Modeling	15
3.6 Tools and Software Used	16
4. Timeline and Plan	17
5. Results	19
5.1 Corpus Composition	19
5.2 Word Frequency Findings	20
5.3 Sentiment by Source Type	21
5.4 Sentiment by Stakeholder Group	22
5.5 Topic Modeling Results	23
6. Discussion	25
6.1 Testing the Expectations	25
6.2 What Does This Mean?	26
7. Limitations of This Project	27
8. Conclusion	28
Acknowledgements	28
References	29
Appendix A: Full Source Corpus (55 Documents)	34
Appendix B: Note on AI Assistance	36
Appendix C: Stakeholder Tagging Codebook	37
Appendix D: My Reflection on the Project	38

1. Introduction

Sanitation policy in India has rarely wanted attention, but it has consistently wanted for agreement. The Swachh Bharat Mission (SBM), launched on 2 October 2014 with the stated aim of eliminating open defecation and improving solid waste management nationwide (Press Information Bureau [PIB], 2022), has been described within the same decade both as the world's largest and most successful behaviour-change campaign (All India Radio News [AIR News], 2024), and as a infrastructure-heavy programme that mistook toilets built for toilets used (Yadavar, 2019). Government communications tend toward the first framing. Investigative journalism and rural ethnography tend toward the second. Between the two sits a body of academic and technical literature that is considerably more measured than either (VerKuilen et al., 2023) and considerably harder to summarize by simply reading a handful of articles.

This disagreement in how SBM is discussed is not incidental to understanding its impact; arguably, it is part of the impact. A programme's public narrative shapes political will, budget allocation, and the willingness of affected communities to engage with it in the first place (McCombs & Shaw, 1972; Entman, 1993). Understanding how that narrative differs by source and by stakeholder is therefore a legitimate empirical question in its own right, separate from, though related to, the question of what SBM has actually achieved on the ground.

This project aims to find an answer to that question through text-analysis. Rather than selecting a small number of sources to read closely, which is the standard approach in a traditional literature review, this study assembled a broader corpus of sources across seven categories and applied computational methods to describe patterns across the corpus as a whole: which words and themes recur, how sentiment varies by who is writing, and how sentiment varies by which stakeholder group a document is centrally about. The intent was not to replace close reading with computation, but to use computation to surface patterns that close reading of a handful of sources would likely miss, and that a full close reading of dozens of sources would simply take too long to do reliably by hand.

1.1 Research questions

The main question for this research is *Does the way the Swachh Bharat Mission is discussed differ systematically depending on who is writing about it and who is being written about, and if so, in what direction?* More elaborately, it explores the following:

- Does sentiment toward SBM differ across types of sources: government, academic, journalistic, and civil-society?
- Does sentiment toward SBM differ depending on which stakeholder group (government, sanitation workers, rural households, women, marginalized castes, and others) is the subject of a source material?
- What themes structure discourse about SBM, and do they correspond to recognizable sub-debates identified through independent reading of the literature?
- How does the literature highlight the concerns of a vulnerable stakeholder group, or can the way an issue is written about mask its seriousness?

1.2 Intent of the project

Specifically, this project aimed to: build a source corpus that deliberately spans a wide range of institutional perspectives on SBM rather than a single viewpoint; apply a standard, citable natural language processing pipeline to that corpus rather than relying on manual, subjective reading alone; compare sentiment and thematic content systematically across source type and stakeholder group; and use the resulting patterns to make an evidence-based argument about how SBM's costs and benefits are discursively distributed that is, who gets talked about positively, who gets talked about critically, and who barely gets talked about with any emotional register at all, positive or negative.

1.3 Structure of the paper

Section 2 reviews the existing literature on SBM, sets out the theoretical and conceptual framework, identifies the research gap, and states the five expectations tested in this paper. Section 3 describes the methodology, including corpus construction, preprocessing, sentiment analysis, and topic modeling. Section 4 presents the results. Section 5 discusses them, first testing each expectation and then considering what the findings mean. Section 6 sets out the limitations of the project, followed by a conclusion. The full source corpus, a note on AI assistance, the stakeholder tagging codebook, and my reflection on the research process are provided in Appendices A to D, respectively.

2. Literature Review

This review brings together the bodies of scholarship that describe what the Swachh Bharat Mission set out to do, what it achieved, and where its record is contested. It traces the mission's policy origins and then reviews four strands of empirical work that also provide a foundation to structure and frame the corpus analysed in this paper: evaluations of toilet construction, use, and health outcomes ; the caste politics of sanitation labour; fund management and accountability; and gender and school sanitation. It then turns to the theoretical literature on framing and computational methods for measuring framing and sentiment in text, before setting out the conceptual framework, the research gap this study addresses, and the expectations it tests. Throughout, the review attends not only to what each literature finds but also to how it is written, because the register of a literature is itself part of what this paper studies.

2.1 From the Total Sanitation Campaign to the Swachh Bharat Mission

SBM is the most recent of several national attempts to end open defecation in India. The Central Rural Sanitation Programme (1986), the Total Sanitation Campaign (1999), and the Nirmal Bharat Abhiyan (2012) each combined household subsidies with awareness campaigns, and each fell well short of its goals (Hueso & Bell, 2013). Hueso and Bell (2013) describe the Total Sanitation Campaign as a policy failure that was obscured by reported progress: administrative data recorded rapid gains in latrine coverage that census and survey data did not support. This history matters for the present study because it shows that a gap between how sanitation programmes are officially reported and how independent researchers measure them was already well established before 2014.

SBM is organized into two sub-missions SBM-Gramin, covering rural India, and SBM-Urban and has proceeded in two phases (PIB, 2022). The first phase, running from 2014 to 2019,

targeted universal open-defecation-free (ODF) status by 2 October 2019, chosen deliberately to coincide with the 150th birth anniversary of Mahatma Gandhi (The Wire, 2019). The second phase, SBM 2.0, launched in 2021, shifted emphasis toward solid and liquid waste management and “garbage-free” city status, with a budget outlay of approximately ₹1,41,600 crore for the urban component alone over two and a half times the outlay of the mission's first phase (Ministry of Housing and Urban Affairs [MoHUA], 2021; Prime Minister’s Office [PMO], 2021). The scale of construction is not seriously disputed: official figures cite over 100 million rural toilets built during the first phase and a nationwide ODF declaration by October 2019 (The Wire, 2019; UNICEF India, 2020). What is disputed, sometimes sharply, is what that construction actually achieved for the people it was meant to serve.

Curtis (2019), drawing on interviews with seventeen officials at national, state, and district level, attributes the speed of SBM’s construction drive to high-level political backing and disruptive leadership, which reordered district administrators’ priorities around a single, measurable target. The same features that made rapid construction possible, a highly visible political goal and district-level pressure to report progress against it, are those that independent studies later link to over-reporting and coercion (Gupta et al., 2019; VerKuilen et al., 2023). The mission’s political prominence is therefore directly relevant to how it is discussed: a programme closely identified with national leadership can be expected to attract strongly evaluative language from both supporters and critics.

2.2 Measuring success: Toilet construction, toilet use and Hhalth outcomes

The central empirical debate in the SBM literature concerns the gap between toilets built and toilets used. VerKuilen et al. (2023) conclude that the government’s monitoring system likely overestimates both toilet coverage and the number of villages that are genuinely open-defecation-free. They identify financial constraints, inadequate water supply, distrust of government, cultural beliefs, and personal preference as the main reasons households continue to defecate in the open.

Independent household surveys support this conclusion while also documenting real progress. Gupta et al. (2019) revisited households in rural Bihar, Madhya Pradesh, Rajasthan, and Uttar Pradesh first surveyed in 2014 and found that, although latrine ownership rose sharply, open defecation fell by only about a quarter between 2014 and 2018 and remained common in all four states; they also reported that threats and coercion were used in parts of the construction drive. Exum et al. (2020), using repeated cross-sectional survey data from Rajasthan, similarly questioned whether the state’s 2018 declaration of open-defecation-free status reflected actual behaviour. Ethnographic and survey-based studies from Bihar, West Bengal, and Jharkhand consistently report that measured open-defecation rates remain higher than official coverage statistics imply (Sarkar & Biswas, 2019; Jain et al., 2020; Novotný et al., 2024).

The literature offers competing explanations for this gap. One emphasises toilet design: many households received cheaper single-pit latrines rather than the twin-pit design recommended for safe, low-maintenance long-term use (VerKuilen et al., 2023; Yadavar, 2019). A second emphasises caste. Coffey and Spears (2017) argue that beliefs about ritual purity and pollution make households in rural north India reluctant to use the affordable small-pit latrines promoted by government programmes, because emptying such pits has historically been work imposed on

Dalit castes; Gupta et al. (2019) likewise report that households preferred latrines with very large, expensive pits to avoid pit emptying. A third emphasises the relationship between citizens and the state. Jain et al. (2020) argue that research and behaviour-change campaigns wrongly treat reluctance to adopt latrines as a preference for open defecation. In their fieldwork in rural Bihar, residents were well aware of its risks, and non-use was tied instead to the size and disbursement of subsidies and to resentment of a perceived development bias against rural areas, which fed a broader distrust of government schemes. The coercive rather than persuasive character of some behaviour-change messaging (VerKuilen et al., 2023) is consistent with this account.

National survey data point in the same mixed direction. National Family Health Survey (NFHS) data (NFHS-5, 2019–21) is cited in the secondary literature as showing improved-sanitation-facility use rising from 49% to 70% nationally between NFHS-4 and NFHS-5, a period overlapping closely with SBM's first phase and real, substantial progress by any reasonable standard (Seth & Jain, 2021). The same source cautions that household-level survey methodology likely understates actual open defecation, since usage varies within a single household by age and by gender in ways a survey answered by one household member cannot fully capture (Seth & Jain, 2021).

Evidence on health outcomes is more positive but more dependent on modelling assumptions. A World Health Organization (WHO) assessment estimated that SBM-Gramin could avert more than 300,000 deaths from diarrhoea and protein-energy malnutrition between 2014 and October 2019 (WHO India, 2018); this was a projection that assumed universal use of safe sanitation, not an observed count. Dandabathula et al. (2019), analysing national disease-surveillance data for 2010–2018, found that acute diarrhoeal disease outbreaks in 2017 and 2018 were fewer than in any earlier year of the period. Varshney et al. (2025) have since published a protocol for a systematic review of SBM's health effects, indicating that the evidence base has not yet been consolidated. The conditional nature of estimates such as WHO's is easily lost when they are repeated in public communication, which is itself an example of the framing process this paper examines.

Taken together, this literature supports neither a simple success narrative nor a simple failure narrative. Coverage expanded on an unprecedented scale, measured use lagged behind coverage, and the size of the gap varied by state, caste, and household. This mixed record is the empirical background against which the sentiment of different source types in this study should be read.

2.3 Sanitation workers, caste, and the manual scavenging question

A separate strand of SBM-adjacent research, less often connected to the toilet-usage debate in public discussion, concerns the safety of sanitation workers. Manual scavenging, the cleaning of human excreta by hand from dry latrines, sewers, or septic tanks, has been prohibited in India since the Employment of Manual Scavengers and Construction of Dry Latrines (Prohibition) Act, 1993, and the prohibition was extended in 2013 to cover hazardous cleaning of sewers and septic tanks (Prohibition of Employment as Manual Scavengers and their Rehabilitation Act, 2013). In *Safai Karamchari Andolan v. Union of India* (2014), the Supreme Court directed governments to identify and rehabilitate manual scavengers and to compensate the families of

workers who died in sewer work, and the Court has since repeatedly ordered an end to manual sewer entry (Outlook, 2026).

The literature is clear that this labour is structured by caste. Human Rights Watch (2014) documented that manual scavenging is performed overwhelmingly by Dalit communities, and that caste-based pressure and threats are used to keep families in the work. Doron and Raja (2015) argue more broadly that sanitation in India cannot be separated from the caste and class relations that determine who handles waste and who controls public space. This connects the worker-safety literature directly to the caste explanation for low toilet use discussed in the previous section (Coffey & Spears, 2017).

Despite the legal prohibition, government data placed before Parliament records several hundred deaths from hazardous sewer and septic-tank cleaning since 2019 alone (Dar, 2026; Deccan Herald, 2020). Advocacy organisations, including the Safai Karmachari Andolan, report figures meaningfully higher than the government's own count and dispute its classification methodology: as of the most recent reporting reviewed for this project, none of the recorded deaths had been formally classified as manual-scavenging-related, a dispute with direct implications for legal accountability and for the compensation owed to victims' families (Outlook, 2026; SabrangIndia, 2026).

This strand of the literature is written in a noticeably different register from the toilet-usage literature. It is more procedural and more statistic-driven, and it is organised around parliamentary reply language and scheme names such as the government's NAMASTE mechanisation scheme (Dar, 2026). This stylistic observation becomes directly relevant to the sentiment-analysis findings in Section 4.

2.4 Fund management, verification and institutional accountability

A third, largely independent strand of the literature, drawn primarily from Comptroller and Auditor General (CAG) audit reports across several states, documents fund-management failures under SBM: delayed disbursement of allocated funds, unspent allocations sitting idle in state or municipal accounts, procurement irregularities (Comptroller and Auditor General of India, 2018, 2025; The Sentinel, 2025), and, in at least one case reviewed for this project, an Enforcement Directorate investigation into the alleged diversion of tens of crores of rupees in mission funds toward unapproved municipal works (Deccan Herald, 2022). This strand recurs across multiple states Uttar Pradesh, Delhi, Assam, and Karnataka among the sources reviewed frequently enough to constitute a distinct, recognizable thread in SBM discourse rather than a set of isolated local failures.

A related concern in the literature is verification. Because open-defecation-free status was declared by the same administrative system that was responsible for achieving it, independent studies have questioned how reliable those declarations are (Exum et al., 2020; VerKuilen et al., 2023). Accountability concerns therefore appear in two forms: financial, in audit findings about how money was spent, and statistical, in disputes over what official coverage and open-defecation-free figures actually measure. Both are closely tied to how the mission is described, since each concerns the difference between what is officially reported and what independent scrutiny finds.

2.5 Gender, menstrual hygiene, and school WASH

A fourth strand concerns menstrual hygiene management and school water, sanitation, and hygiene (WASH) facilities, formally incorporated into SBM's guidelines from 2017 onward through initiatives such as “Swachh Bharat: Swachh Vidyalaya” and the Ministry of Jal Shakti's Guidelines on Gender Issues in Sanitation (Ministry of Drinking Water and Sanitation, 2017; Sinha & Paul, 2018). This literature documents genuine institutional progress, formal recognition of menstrual hygiene as a sanitation-policy concern, requirements for gender-segregated school toilets alongside continuing implementation gaps in water access, privacy, and disposal facilities documented in household- and school-level surveys from Odisha, Assam, and Maharashtra (Behera et al., 2022; Kabir et al., 2020; Borah, 2026). Notably, this literature is generally written in a more constructive, policy-progress register than the worker-safety or fund-audit literature, even when it is reporting on comparable implementation shortfalls, a second stylistic observation relevant to the sentiment findings in Section 4.

2.6 Theoretical framework: framing, salience, and institutional voice

This project's central premise, that who is speaking and about whom shapes how an issue is discussed independently of the underlying facts, draws on a long tradition of framing research. Communication scholar Robert Entman describes framing as fundamentally a matter of “selection and salience”: to frame is to select some aspects of a perceived reality and make them more salient in a text, in a way that promotes a particular problem definition, causal interpretation, or moral evaluation (Entman, 1993). Earlier, sociologist Erving Goffman argued that people, and by extension institutions, organise their understanding of events through interpretive frameworks that determine what counts as relevant, rather than passively reporting a neutral reality (Goffman, 1974). Agenda-setting theory adds that media and institutional sources shape not only what audiences think about an issue but which issues, and which aspects of them, are considered worth discussing at all (McCombs & Shaw, 1972).

Later work distinguishes between frames in communication, the words, emphases, and evaluations a speaker uses to present an issue, and frames in thought, the way audiences come to understand it (Scheufele, 1999; Chong & Druckman, 2007). This study is concerned only with the first: it measures how sources write about SBM, not how readers interpret them. Content-analytic research has also identified a small set of generic news frames that recur across topics, including attribution of responsibility, conflict, human interest, economic consequences, and morality (Semetko & Valkenburg, 2000). Several of these can be expected in SBM coverage: government communication tends to emphasise achievement and civic responsibility, investigative reporting uses responsibility framing to attribute failure, and reporting on worker deaths may rely more on statistics than on human-interest framing.

Applied to SBM, this framework predicts the kind of source-type gradient this project set out to test. Government communications, institutionally positioned to promote a favourable evaluation of the mission, would be expected to select and foreground different aspects of its reality than investigative journalism, which is conventionally oriented toward identifying institutional failure. This project's contribution is to test that prediction computationally across a real corpus rather than asserting it impressionistically.

2.7 Computational approaches to framing and sentiment

Automated text analysis has become a standard tool for studying political and policy communication at scale. Grimmer and Stewart (2013) argue that such methods make large bodies of text analysable but cannot replace careful reading, and that every automated measure must be validated against the texts it summarises. Topic models such as Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003) have been used to study framing directly: DiMaggio, Nag, and Blei (2013) showed that topics extracted from newspaper coverage of U.S. government arts funding corresponded closely to the frames sociologists had identified in that debate, supporting the use of topic models as a way of measuring how an issue is framed.

Dictionary-based measures of tone have a similar history in political communication (Young & Soroka, 2012), but their limitations are well documented. Boukes et al. (2020) found that widely used off-the-shelf sentiment tools agree poorly both with one another and with human coders when applied to news text, and van Atteveldt, van der Velden, and Boukes (2021) found that dictionary approaches performed worse than manual annotation and machine-learning methods in measuring the sentiment of economic news. VADER (Hutto & Gilbert, 2014), the tool used in this study, was developed and validated mainly on short, informal social-media text; its performance on formal, bureaucratic prose is much less established.

Two implications follow for this study. First, sentiment scores are treated here as a measure of the tone of the writing, not of the events being described, and are interpreted alongside topic modelling and close reading, as Grimmer and Stewart (2013) recommend. Second, the known weakness of lexicon-based tools on procedural language is not only a limitation but a testable property: if worker-safety reporting is written in a flat, statistical register, a lexicon-based tool may score it closer to neutral than its subject matter would suggest. Expectation E4 (Section 2.10) tests whether this occurs in the SBM corpus.

Putting the framing theory from Section 2.6 in more everyday terms: every source in this corpus has a reason for writing about SBM, and that reason shapes what it chooses to say.

- Government and policy sources have an institutional stake in SBM being perceived as successful, and communicate accordingly favouring achievement statistics, participation numbers, and forward-looking announcements over documented shortfalls.
- Investigative journalism and opinion writing operate under a different set of professional norms, in which identifying and naming a problem is itself the point of the piece, which should be expected to pull sentiment in the opposite direction.
- Academic research operates under norms of balance and evidentiary caution, and reports on a genuinely mixed empirical record, described in Sections 2.2 through 2.5 so should be expected to sit closer to neutral than either institutional or journalistic sources, rather than skewing toward either pole.
- Separately from who is writing, how seriously a topic is treated in a piece of writing is not the same thing as how serious the topic actually is. A death count reported in bureaucratic, procedural language does not necessarily register as “negative” to a sentiment-scoring tool in the way an angry opinion column does, even where the

underlying subject matter is, in human terms, far more serious (Hutto & Gilbert, 2014; Boukes et al., 2020).

Together, these four ideas motivate the five expectations set out in [Section 2.10](#), and structure how the findings in Section 4 are interpreted in Section 5.1 and discussed in Section 5.2.

2.8 Research Gap

The literature reviewed above has three features that together define the gap this study aims to bridge. First, SBM research has focused almost entirely on the mission's outcomes, measured either through large surveys of coverage, use, and disease (Gupta et al., 2019; Dandabathula et al., 2019; VerKuilen et al., 2023; Varshney et al., 2025) or through close qualitative study of particular villages and districts (Jain et al., 2020). How SBM is discussed has not itself been studied systematically, even though the gap between official and independent accounts of the mission is one of the most consistent findings in that literature (Hueso & Bell, 2013; VerKuilen et al., 2023).

Second, computational studies of framing have more often examined a single type of source, typically newspaper coverage (e.g., DiMaggio et al., 2013), than compared institutionally different sources writing about the same policy. Third, methodological work on automated sentiment analysis has documented its weaknesses mainly in news and social-media text (Boukes et al., 2020; van Atteveldt et al., 2021) and has paid little attention to the specific problem this paper identifies: that serious harms reported in procedural language may register as neutral.

This study addresses these gaps by applying computational text analysis across a deliberately heterogeneous corpus of government, academic, journalistic, and civil-society sources to ask how SBM's discourse itself, as distinct from its measured outcomes, varies systematically by who is producing it and who it concerns, using a corpus I assembled and a pipeline I built myself in Python.

2.9 Expectations

Before building the corpus, I set out five directional expectations, grounded in the framework discussed in Sections above:

- E1: Government and policy sources will frame SBM more positively than journalism and opinion sources.
- E2: Sentiment will differ meaningfully depending on which stakeholder group a document concerns.
- E3: Topics recovered through unsupervised topic modeling will correspond to sub-debates I identify independently through background reading.
- E4: Stakeholder groups facing documented serious harms, such as sanitation workers, will register strongly negative sentiment scores, proportional to the severity of what is reported about them.
- E5: Academic sources, drawing on a mixed empirical record, will sit closer to neutral sentiment than either institutional or journalistic sources.

Section 5.1 returns to each of these expectations directly, after the results are presented in Section 4.

3. Methodology

3.1 Overall Approach: Why Text Analysis

The goals of the study was to test, rather than simply assert, the intuition that SBM is discussed differently by different kinds of sources. Before settling on computational text analysis, two alternatives were considered. First, conducting a small number of interviews with people directly affected by SBM sanitation workers, rural households, local officials which would have produced richer individual testimony. However, this would not allow a systematic comparison of how the mission is discussed across a wide range of published sources, which was the prime research question. Second was a purely qualitative literature review, reading and synthesizing a smaller number of sources by hand, in the traditional style. This was found to be the unsuitable tool for the specific question asked in this research: a traditional review is well suited to synthesizing what a literature concludes, but poorly suited to systematically comparing how that literature's sentiment and word choice differ by source type and stakeholder focus at a scale of 50 or more documents, since a human reader cannot easily hold that many documents' word-level patterns in mind at once, and a small hand-picked sample of sources is vulnerable to unconscious selection bias in exactly the dimension I was trying to measure. Computational text analysis, applied to a deliberately broad corpus, was the approach that let me test my actual research questions directly and reproducibly.

3.2 Corpus construction

The corpus consists of 55 English-language documents, collected to represent seven source categories: academic papers (17), news reports (11), government press releases (7), policy briefings (6), investigative journalism (6), opinion and analysis pieces (4), and NGO or development-bank reports (4). Table 1 summarizes this breakdown.

Table 1. Corpus composition by source type

Source type	Count	Share
Academic papers	17	31%
News reports	11	20%
Government press releases	7	13%
Policy briefings	6	11%
Investigative journalism	6	11%
Opinion / analysis	4	7%
NGO / development-bank reports	4	7%

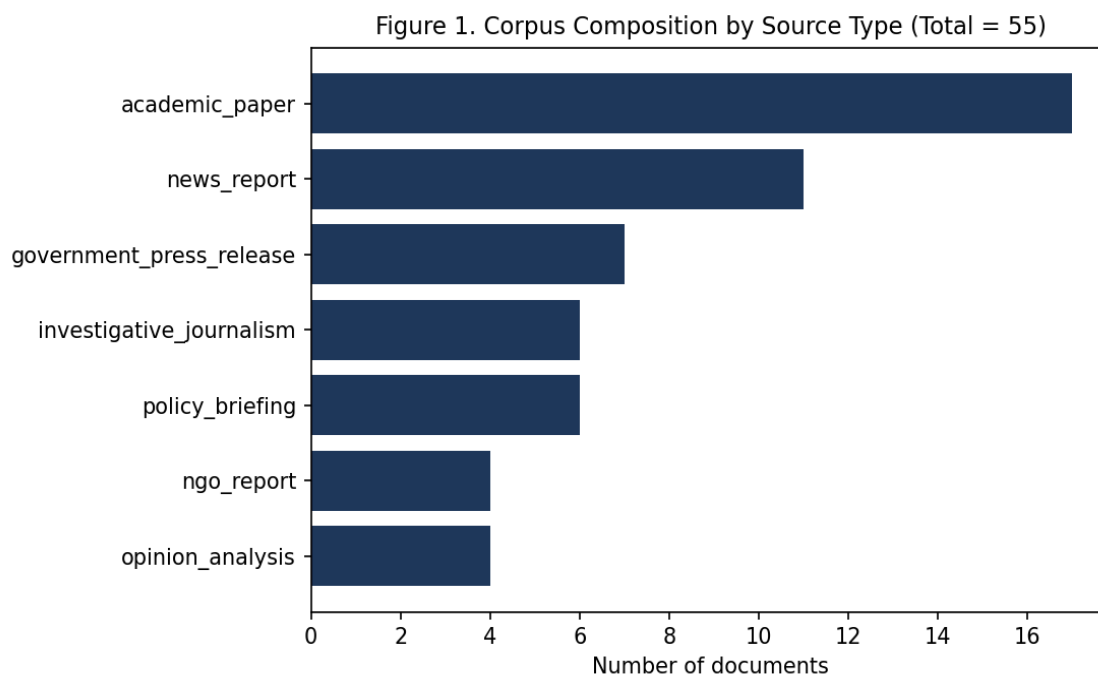


Figure 1. Corpus composition by source type ($n = 55$).

Sources were identified through targeted searches covering each of the four thematic strands discussed in Section 2- rural toilet usage, sanitation-worker safety, fund management, and gender or school WASH plus general SBM coverage, deliberately favouring breadth of institutional perspective over depth in any single strand. Documents were included where they substantively discussed SBM's implementation, impact, or governance, rather than mentioning the mission only in passing.

Each document was coded with its source type, publication date, source URL, and one or more stakeholder tags drawn from a fixed vocabulary of fifteen categories: government, sanitation workers, rural households, urban local bodies, women, students, marginalized castes, marginalized communities, public health, policy experts, NGO sector, general public, citizens, infrastructure, and local administration. A stakeholder tag was applied wherever a document's central subject, not merely a passing mention, concerned that group's treatment, experience, or outcomes under SBM. A document could receive more than one tag where it substantively addressed more than one group; for instance, a study on female sanitation workers would be tagged both 'women' and 'sanitation_workers'. The full tagging codebook is reproduced in Appendix C.

A material methodological limitation here is that each document is represented in the corpus by a concise summary of its content rather than by the full original text. This reflects a practical constraint on reproducing full copyrighted material news articles, paywalled reports, and academic papers in a working dataset that could be freely shared and reviewed. The summaries were written to capture each source's key claims, findings, and framing as accurately as possible, and the pipeline described below was applied identically to every document regardless of source type, so the comparative patterns reported in Section 4 should still be broadly informative.

However, the analysis in this paper should be read as a validated pipeline run on representative summaries rather than a final, publication-grade result; re-running an identical pipeline on full-length original source text is a direct and comparatively low-effort extension of this work, discussed further in Appendix D.

Stakeholder tagging was performed by me, without independent cross-validation by a second coder; inter-rater reliability has therefore not been formally established, and the stakeholder-level comparisons in Section 4.4 should be read as descriptive rather than statistically validated categorizations. This is discussed further as a limitation in Section 6.

3.3 Preprocessing pipeline

Before any analysis could be run, every document's text needed to be converted into a standardized form a computer could work with consistently. This involved several sequential steps, each addressing a different kind of noise in raw text.

Cleaning

Each document's text was first converted to lowercase, so that “Sanitation” and “sanitation” would be counted as the same word rather than as two different words. URLs and standalone numeric digits were removed using regular-expression pattern matching, since these do not carry topical meaning for word-frequency or topic-modeling purposes, though, as discussed in Section 4.2, the presence of numeric and budget-related vocabulary in the underlying text still shows up indirectly through terms like “crore.”

Tokenization

Tokenization is the process of splitting a continuous string of text into individual units, or tokens, roughly corresponding to words. This study used NLTK's `word_tokenize` function (Bird, Klein, & Loper, 2009) rather than a simple whitespace split, because `word_tokenize` correctly separates punctuation from adjacent words and handles contractions in a linguistically sensible way for example, splitting “don't” into “do” and “n't” rather than treating the whole contraction as one unrecognizable token, and correctly separating a trailing comma or period from the word it follows. A naive split on whitespace would instead produce tokens like “crore,” or “waste.” with punctuation still attached, artificially inflating the apparent vocabulary size and undercounting true word frequency.

Stopword removal

Not every word carries useful topical information. Words like “the,” “and,” “is,” and “of” appear extremely frequently in English text regardless of subject matter, and including them in a word-frequency count or topic model would simply produce a list dominated by grammatical scaffolding rather than content. This project removed NLTK's standard list of 179 English stopwords, and additionally removed a small set of corpus-specific high-frequency terms “india,” “mission,” “swachh,” and “bharat” that appear in nearly every document in this corpus by definition, since every document is about the Swachh Bharat Mission, and would otherwise dominate frequency-based results without conveying any information that distinguishes one document from another.

Stemming and lemmatization

Two related but distinct techniques were used to handle the fact that a single underlying concept can appear in many grammatical forms “clean,” “cleaned,” “cleaning,” and “cleans” are all forms of the same root idea, but would be counted as four separate words without some form of normalization. Stemming, implemented here using the Porter algorithm (Porter, 1980), works by applying a fixed sequence of suffix-stripping rules for example, removing common endings like “-ing,” “-ed,” and “-ational” under specific conditions to reduce a word to an approximate root form. This is fast and effective for grouping related words together, but the resulting stem is not always a real dictionary word: for example, “sanitation” stems to “sanit,” and “argument” to “argu.” Lemmatization, by contrast, uses a dictionary-based approach, via NLTK’s WordNet-based lemmatizer (Miller, 1995), to reduce a word to its genuine dictionary base form, or lemma for example, reducing “was” to “be” and “children” to “child.” Because a word’s correct lemma depends on its grammatical role in a sentence, for example “better” as an adjective lemmatizes differently than “better” as a verb, this project first applied part-of-speech tagging to each token using NLTK’s averaged perceptron tagger, then mapped each tag to the format the WordNet lemmatizer expects, before lemmatizing. Both stemmed and lemmatized versions of the corpus were computed; lemmatized tokens were used for the primary word-frequency and topic-modeling analysis reported in Section 4, since real dictionary words are considerably easier for a human reader to sanity-check than truncated stems, which matters directly for verifying that topic-model output is sensible rather than noise.

3.4 Sentiment analysis

Sentiment analysis assigns a numerical score representing how positive or negative a piece of text is. This study used VADER, standing for Valence Aware Dictionary and sEntiment Reasoner, a lexicon- and rule-based sentiment model developed specifically for the mixed formal-to-informal register found in news, social, and opinion text, and validated directly against human sentiment ratings (Hutto & Gilbert, 2014). VADER works by combining a large pre-built dictionary of words rated for sentiment intensity by human annotators with a set of grammatical rules that adjust a raw score based on context for example, recognizing that “not good” should score negatively despite containing the positive word “good” (negation handling), that “very poor” should score more intensely negative than “poor” alone (degree-modifier handling), and that capitalization or exclamation marks can intensify a sentiment. This made VADER a better fit for this corpus than a simple hand-built positive or negative word-count lexicon, which cannot handle negation or intensification at all and would, for example, score “not effective” as positive purely because it contains the word “effective.” VADER was applied to each document’s original, non-lemmatized text rather than to the cleaned token list, since its scoring depends on exactly the punctuation, capitalization, and word order that the earlier preprocessing steps deliberately strip out. Each document received a single compound sentiment score, a normalized value ranging from -1, most negative, to +1, most positive, calculated by summing the valence scores of each word in the text, adjusted for the contextual rules above, and normalizing the result to fall within that fixed range.

3.5 Topic modeling

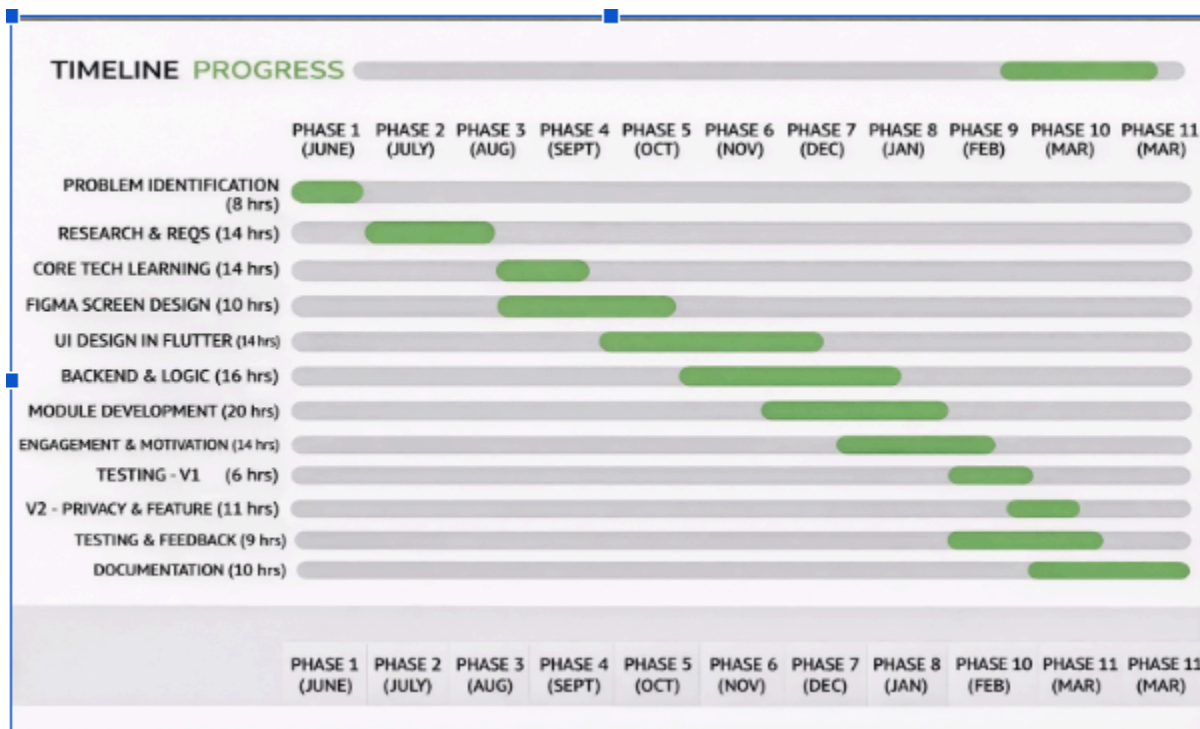
Where sentiment analysis asks how positively or negatively a document reads, topic modeling asks a different question: without being told in advance what the topics are, can an algorithm discover which words tend to occur together across the corpus, and therefore infer what a document is substantively about? This project used Latent Dirichlet Allocation, or LDA, a widely used topic-modeling algorithm (Blei, Ng, & Jordan, 2003). LDA's underlying idea, described here in plain terms, is that every document can be thought of as a mixture of a small number of latent, or hidden, topics, and every topic can be thought of as a probability distribution over words some words are highly characteristic of a given topic, and others are not. The algorithm does not know the topics in advance; instead, it works backward from the observed word patterns across the whole corpus to infer both a plausible set of topics and how much each document draws on each topic, using a statistical technique that assumes a Dirichlet probability distribution over topic proportions, which gives the method its name.

Before applying LDA, documents were converted into numeric vectors using TF-IDF weighting, standing for Term Frequency–Inverse Document Frequency (Salton & Buckley, 1988), a standard technique that weights each word in a document by how frequently it appears in that document, discounted by how common that word is across the whole corpus. The effect is that a word appearing in nearly every document, which after stopword removal still includes words like “sanitation” given this corpus's subject matter, receives a lower weight than a word that is frequent in a smaller subset of documents but rare elsewhere, such as “menstrual,” “manhole,” or “audit,” since the latter kind of word is far more informative about what specifically distinguishes one document from another. LDA was then run on these TF-IDF-weighted vectors, using scikit-learn's implementation, to extract five topics. The number of topics, five, was chosen to correspond to the number of distinct thematic strands identified independently through the background reading summarized in Section 2, providing a check on whether the unsupervised method would recover a structure resembling one arrived at through ordinary human reading, discussed directly as part of testing Expectation E3 in Section 5.1.

3.6 Tools and software used

All analysis was conducted in Python 3, primarily using the pandas library for data handling, NLTK, the Natural Language Toolkit, for tokenization, stopword removal, part-of-speech tagging, stemming, and lemmatization, vaderSentiment for sentiment scoring, and scikit-learn for TF-IDF vectorization and LDA topic modeling. Code was written and run in Jupyter notebooks, executed through Google Colab, which provides free cloud-based access to a Python environment with internet access, needed to install packages and download NLTK's language-data files, without requiring any local software installation. Corpus construction and management used Microsoft Excel and CSV files, and visualizations were produced using matplotlib. The complete, executable notebook, corpus file, and generated result files accompany this paper.

4. Timeline and Plan



Total Hours: 148 hours

Phases

Breakdown of time commitment

Phase 1: Identifying the Problem & Initial Understanding (8 hours)

- Understanding the context (4 hrs): Explored the history of the Swachh Bharat Mission, from the Total Sanitation Campaign to SBM 2.0, and noticed how sharply official and independent accounts of its success differed.
- Problem framing (4 hrs): Defined the core question: does the way SBM is discussed differ systematically by who is writing and who is being written about? Considered interviews and a traditional literature review before choosing computational text analysis.

Phase 2: Background Reading and Literature Review (20 hours)

- Reading on SBM outcomes (8 hrs): Reviewed academic studies, household surveys, NFHS data, and health-outcome estimates on toilet construction, toilet use, and the gap between the two.
- Identifying thematic strands (6 hrs): Identified four recurring sub-debates: rural toilet usage, sanitation-worker safety and caste, fund management and CAG audits, and gender and school WASH.
- Framing theory and computational methods (6 hrs): Studied framing and agenda-setting theory

(Entman, Goffman, McCombs & Shaw) and research on topic modeling, sentiment analysis, and the limitations of lexicon-based tools.

Phase 3: Research Design and Expectations (8 hours)

- Research questions and gap (3 hrs): Finalized four research questions and defined the research gap: SBM's discourse, as distinct from its outcomes, had not been studied systematically.
- Setting expectations (2 hrs): Wrote five directional expectations (E1–E5) before any data was collected, so results could be tested against them honestly.
- Mentor review (3 hrs): Discussed research design, methodology, and scope with mentors Gaurav Thapak and Amrita Saib at Ernst & Young.

Phase 4: Codebook and Corpus Structure Design (6 hours)

- Source-type categories (2 hrs): Defined seven source categories: academic papers, news reports, government press releases, policy briefings, investigative journalism, opinion pieces, and NGO/development-bank reports.
- Stakeholder tagging codebook (4 hrs): Created a fixed vocabulary of fifteen stakeholder groups and rules for when a document receives a tag, including multiple tags for documents covering more than one group.

Phase 5: Corpus Collection and Coding (30 hours)

- Targeted source search (10 hrs): Searched for sources across all four thematic strands, deliberately looking for underrepresented angles such as worker safety, fund audits, and menstrual hygiene so the corpus would not skew toward the easiest storyline to find.
- Reading and summarizing (12 hrs): Read each source and wrote a concise summary capturing its key claims, findings, and framing, for 55 documents in total.
- Coding and tagging (5 hrs): Recorded source type, publication date, URL, and stakeholder tags for every document in Excel.
- Quality check and balancing (3 hrs): Reviewed summaries side by side, caught a duplicate story on a UN manual-scavenging dispute, replaced it with NFHS-5 survey data, and checked that categories were reasonably balanced.

Phase 6: Learning Core Technologies (Python & NLP) (12 hours)

- Python and data handling (5 hrs): Learned pandas for data handling and set up Jupyter notebooks in Google Colab to run code without local installation.
- NLP concepts (7 hrs): Learned tokenization, stopword removal, stemming, lemmatization, sentiment scoring, TF-IDF weighting, and Latent Dirichlet Allocation.

Phase 7: Building the Preprocessing Pipeline (10 hours)

- Cleaning and tokenization (3 hrs): Converted text to lowercase, removed URLs and digits, and tokenized using NLTK's word_tokenize.
- Stopword removal (2 hrs): Removed NLTK's 179 standard English stopwords plus corpus-specific terms such as "swachh," "bharat," "mission," and "india."
- Stemming and lemmatization (5 hrs): Implemented Porter stemming and POS-tagged WordNet lemmatization, and chose lemmatized tokens for the main analysis because they are easier to check by eye.

Phase 8: Sentiment Analysis (8 hours)

- Prototype lexicon (3 hrs): Built a first version using a small hand-made list of positive and negative words. Results looked plausible but could not be defended methodologically.
- Migration to VADER (5 hrs): Replaced the prototype with VADER, a validated sentiment model, applied it to the original text, and calculated mean scores by source type and stakeholder group.

Phase 9: Topic Modeling (8 hours)

- TF-IDF and LDA setup (4 hrs): Converted documents into TF-IDF vectors and ran a five-topic LDA model using scikit-learn.
- Refinement (4 hrs): The first run on raw tokens produced topics that were hard to interpret, so switched to lemmatized tokens before vectorization, which produced clearer topics.

Phase 10: Results Interpretation and Visualization (10 hours)

- Visualizations (3 hrs): Created charts in matplotlib for corpus composition, word frequency, sentiment by source type and stakeholder, and topic distribution.
- Testing expectations (3 hrs): Compared results against E1–E5 and found four supported or partially supported and one not supported.
 - Investigating the sanitation-worker finding (4 hrs): The near-neutral score for sanitation workers first looked like a coding error. Re-reading the underlying documents showed it reflected the flat, procedural way worker deaths are reported, and with mentor encouragement this became a key finding.

Phase 11: Full-Text Extraction Tool (6 hours)

- Extraction tool (4 hrs): Built a separate tool using pdfplumber to extract full text from source PDFs, so summaries can gradually be replaced with original text.
- Manual transcription (2 hrs): Set up a manual process for sources on websites that block automated scraping.

Phase 12: Documentation (22 hours)

- Writing the paper (15 hrs): Wrote the full paper, covering introduction, literature review, methodology, results, discussion, limitations, and conclusion.
- References and appendices (3 hrs): Compiled references, the full 55-document corpus list, the stakeholder codebook, the AI-assistance note, and a personal reflection.
- Mentor review and revisions (4 hrs): Revised the paper based on review and edits from Gaurav Thapak.

5. Results

5.1 Corpus composition

The final corpus contains 55 documents across seven source-type categories. Academic papers form the largest single category (17 documents, 31% of the corpus), followed by news reports (11 documents, 20%), government press releases (7 documents, 13%), policy briefings and

investigative journalism (6 documents each, 11% each), and opinion/analysis and NGO/development-bank reports (4 documents each, 7% each). This distribution reflects a deliberate choice to weight the corpus toward academic and journalistic sources, which together provide the most substantive, evidence-based content, while still ensuring institutional (government) and civil-society (opinion, investigative, NGO) perspectives are represented in sufficient numbers for meaningful comparison.

5.2 Word frequency

Across the corpus, the most frequent content-bearing lemmatized terms were toilet (17 occurrences), sanitation (16), open (11, almost entirely as part of the phrase “open defecation”), crore (11, the Indian numbering unit for ten million, reflecting how much of the corpus concerns budget figures), study (10), defecation (10), waste (9), state (9), coverage (9), and government (9).

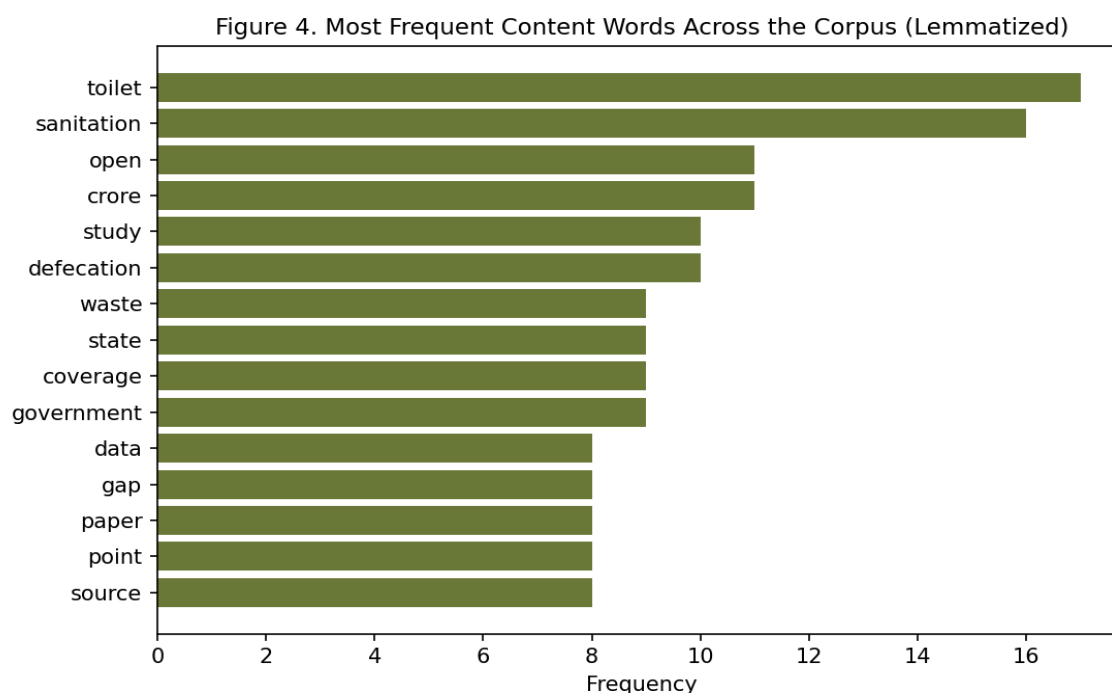


Figure 2. Most frequent content words across the corpus, lemmatized ($n = 55$).

The prominence of budget-related and survey-related vocabulary, specifically crore, study, coverage, data, and gap, alongside the expected substantive terms toilet, sanitation, defecation, and waste, is itself informative: a meaningful share of SBM discourse in this corpus is about measurement and accountability, meaning what has been counted, funded, or audited, rather than purely descriptive coverage of the programme's day-to-day activities. This is consistent with the fund-management strand of the literature identified in Section 2.4, and suggests that strand is not a marginal concern within the broader discourse but a recurring and substantial one.

5.3 Sentiment by source type

Mean sentiment varied substantially and, notably, monotonically by source type, from most positive to most negative: policy briefings (+0.565), government press releases (+0.384), NGO and development-bank reports (+0.243), academic papers (+0.052, effectively neutral), news reports (-0.076), investigative journalism (-0.097), and opinion and analysis pieces (-0.422).

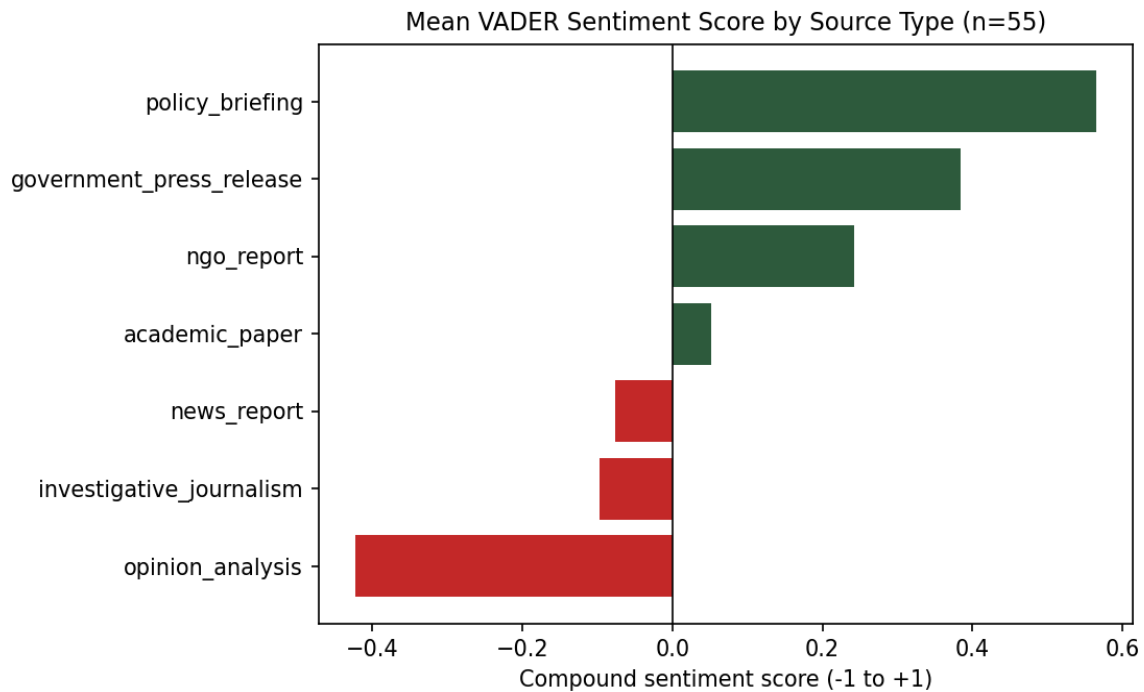


Figure 3. Mean VADER sentiment score by source type (n = 55).

Table 2. Mean sentiment by source type

Source type	Mean sentiment	n
Policy briefings	+0.565	6
Government press releases	+0.384	7
NGO / development-bank reports	+0.243	4
Academic papers	+0.052	17
News reports	-0.076	11
Investigative journalism	-0.097	6

Opinion / analysis	-0.422	4
--------------------	--------	---

This gradient is broadly consistent with the institutional position of each source type, as anticipated in Sections 2.6 and 2.8. Government and policy documents are written to communicate progress and are produced by parties with a direct interest in the programme's success being legible to the public and to political stakeholders; opinion and investigative journalism are, by genre convention, oriented toward identifying and stating problems rather than affirming successes. What is particularly notable is that academic papers sit almost exactly at neutral rather than skewing toward either pole, which is consistent with a literature that, as Section 2 describes at length, contains genuinely mixed findings, real coverage gains alongside real usage gaps, rather than a literature that has settled decisively on either a success or a failure narrative.

5.4 Sentiment by stakeholder group

Sentiment also varied by which stakeholder group a document primarily concerned, though the pattern here is less a clean institutional gradient than the source-type result, and arguably more informative for that reason.

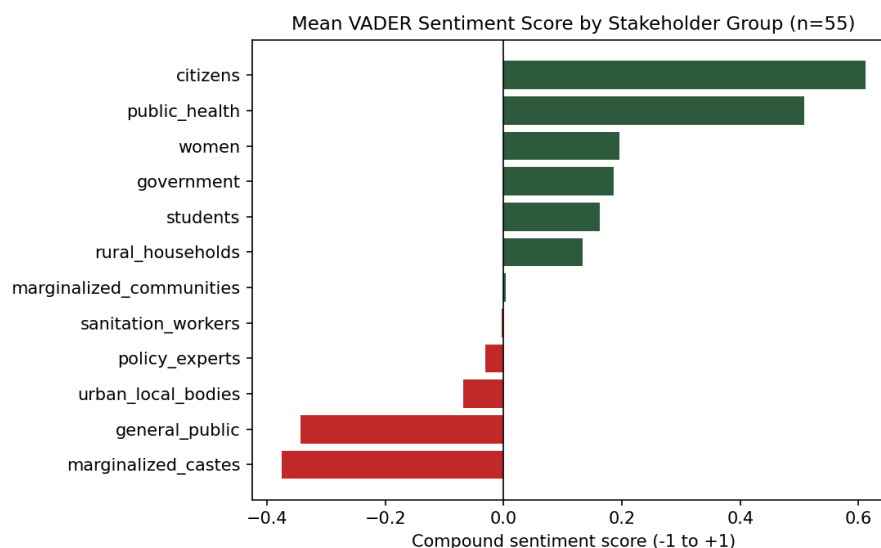


Figure 4. Mean VADER sentiment score by stakeholder group (n = 55; groups with fewer than 2 documents excluded).

Table 3. Mean sentiment by stakeholder group

Stakeholder group	Mean sentiment	n
Citizens	+0.612	2
Public health	+0.508	3

Women	+0.196	10
Government	+0.187	31
Students	+0.163	6
Rural households	+0.133	15
Marginalized communities	+0.003	2
Sanitation workers	-0.003	10
Policy experts	-0.031	4
Urban local bodies	-0.069	14
General public	-0.343	3
Marginalized castes	-0.376	5

Citizens (+0.612) and public health outcomes (+0.508) register as the most positively framed groups, followed by women (+0.196), government (+0.187), students (+0.163), and rural households (+0.133); all net positive. At the other end, marginalized castes (-0.376) and the general public in a critically-framed context (-0.343) are the most negatively framed groups. Two results merit specific comment.

First, women's positive sentiment is driven substantially by the menstrual-hygiene and school-WASH literature discussed in Section 2.5, which is written in a constructive, policy-progress register even when documenting real implementation gaps, a genre effect worth distinguishing clearly from any claim that women's actual outcomes under SBM are uncomplicatedly positive; the sentiment score reflects how the topic is written about, not a verified measurement of outcomes.

Second, and more strikingly, sanitation workers register at effectively neutral sentiment (-0.003) despite the corpus containing multiple documents reporting hundreds of worker deaths from hazardous sewer and septic-tank cleaning (Deccan Herald, 2020; ETV Bharat 2025; Dar, 2026). This is not evidence that sanitation workers' circumstances are neutral; it is evidence that coverage of their circumstances is frequently written in a flat, statistical, procedural register, meaning death counts, scheme names, and parliamentary reply language, that VADER, like most lexicon-based sentiment tools, does not read as strongly negative in the way that emotionally charged opinion language is read as strongly negative (Hutto & Gilbert, 2014). This is a genuine finding about how a topic is written, not a claim about how serious the underlying issue is, and the distinction is discussed further in Section 5.2.

5.5 Topic modeling results

Five-topic LDA over the corpus produced the following top terms per topic, ranked by weight within each topic:

- Topic 1 (10 documents): still, household, toilet, problem, survey, access, sanitation, campaign corresponds closely to the rural toilet-access and usage-gap literature discussed in Section 2.2.
- Topic 2 (12 documents): toilet, sanitation, instead, year, municipal, defecation, open, policy corresponds to municipal and policy-level sanitation and waste management, connecting Sections 2.1 and 2.4.
- Topic 3 (13 documents): government, policy, number, death, sbm, issue, crore, bihar, manual a mixed government-policy and worker-safety cluster, overlapping Sections 2.3 and 2.4.
- Topic 4 (8 documents): death, useful, funding, notably, city, figure, count, crore overlaps substantially with Topic 3, both concentrated around worker deaths and funding figures.
- Topic 5 (12 documents): school, guideline, paper, study, level, student, gap, release, solid corresponds closely to the menstrual hygiene and school WASH cluster discussed in Section 2.5.

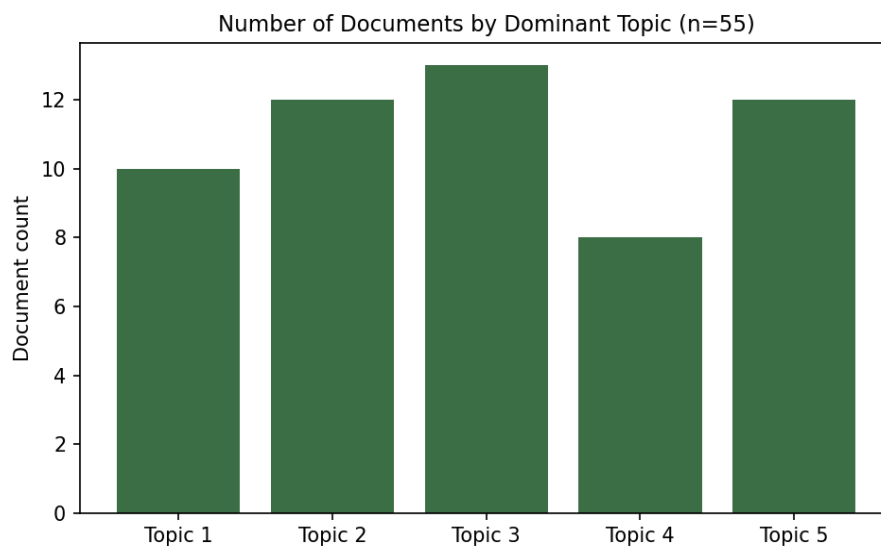


Figure 5. Number of documents assigned to each dominant topic ($n = 55$).

Topics 1, 2, and 5 are cleanly interpretable and correspond closely to sub-debates identified independently during the background research summarized in Section 2, before any modeling was run rural usage, municipal and waste policy, and school WASH, respectively. Topics 3 and 4 show meaningful overlap: both are organized around death counts, funding figures, and government-response language, which most plausibly reflects that worker-safety reporting and fund-audit reporting in this corpus share a great deal of vocabulary, such as crore, government, figures, and official responses, even when their underlying subject matter differs. This overlap, rather than indicating a methodological failure, is itself a modest and genuine finding: worker-safety and fund-accountability discourse in this corpus are not sharply distinguished from one another in practice, which may reflect how frequently the two issues are reported on together, for instance in a single report on inadequate mechanisation funding that also documents worker deaths resulting from that funding gap.

6. Discussion

6.1 Testing the expectations

Section 2.10 set out five directional expectations before the analysis was run. Table 4 summarizes how each held up against the results presented in Section 4.

Table 4. Summary of five expectations

Expectation	Findings	Result
E1: Govt./policy sources more positive than journalism/opinion	+0.56 to +0.38 vs. -0.42, -0.10	Supported
E2: Sentiment differs by stakeholder group	Range +0.61 to -0.38	Supported
E3: Topics correspond to independently-identified sub-debates	3 of 5 topics separated cleanly; 2 overlapped	Partially supported
E4: Documented serious harm shows as strongly negative sentiment	Not confirmed — sanitation workers near-neutral (-0.003)	Not supported
E5: Academic sources sit closer to neutral	+0.052, closest to zero of all seven types	Supported

Four of the five expectations were supported, while E3 was only partially supported and E4 was not supported. Taken together, these results answer the main research question in a fairly consistent direction: the way SBM is discussed does differ systematically according to both who is writing and who is being written about.

The source-type results provide the clearest evidence for the first part of this question. Government and policy sources were substantially more positive than investigative journalism and opinion writing, while academic sources remained close to neutral. This suggests that institutional position is associated with a measurable difference in the language used to discuss SBM.

The stakeholder results answer the second part of the question. Sentiment varied across stakeholder groups, with citizens and public-health outcomes among the more positively framed groups and marginalized castes among the most negatively framed. The sanitation-worker result adds an important qualification: a stakeholder can be discussed in connection with severe harm without the language itself receiving a strongly negative sentiment score.

This means that the answer is not simply that some groups are discussed positively and others negatively. The form of the writing matters as well. Government announcements, policy documents, investigative reporting, and academic research follow different conventions, and those conventions affect what a sentiment model detects.

The topic-modeling results provide a third part of the answer. Three of the five topics corresponded closely to debates identified independently through background reading, while two overlapped substantially. This gives some support to the idea that the major divisions in the

literature are also visible in the language of the corpus, while showing that those divisions are not completely separate.

The most important result is therefore not that sentiment analysis can tell us whether SBM is successful. It cannot. Rather, the analysis shows that computational text methods can detect systematic differences in how the mission and its stakeholders are described, while also revealing cases where the apparent sentiment of the writing does not reflect the seriousness of the underlying issue.

6.2 Implications

The results give a fairly clear answer to my main research question. The way the Swachh Bharat Mission is discussed does change depending on both who is writing about it and who is being written about. The biggest difference appeared between different types of sources. Government releases and policy briefings were much more positive overall, while opinion pieces and investigative journalism were much more negative. Academic sources were much closer to neutral. There were differences between stakeholder groups as well. Citizens and public health were among the more positively discussed groups, while marginalized castes and the general public were among the more negatively discussed. The topic modeling also found several of the same areas of debate that came up during my background research, such as rural sanitation, waste management, sanitation workers, funding, and school and menstrual hygiene.

Overall, this suggests that there is not one single way of talking about SBM. The way it is presented depends partly on the type of source and partly on the group being discussed.

Taken together, these findings support a central claim of this paper: SBM's benefits and costs are not evenly distributed across the groups this corpus terms stakeholders, and that unevenness is legible directly in how the mission is discussed, not only in what the mission has measurably delivered. Framings cluster by two variables simultaneously, who is writing, meaning institutional position, and who is being written about, meaning stakeholder group, and these two variables interact with each other. Government and policy sources write positively about government and citizens; independent journalism writes more critically, and disproportionately about groups, specifically marginalized castes and sanitation workers, whose treatment under the mission is most contested in the underlying literature.

The sanitation-worker finding discussed in Section 4.4 is the clearest illustration of why sentiment analysis alone, without qualitative interpretation, risks producing a misleading headline number if reported without context. A naive reading of a near-zero sentiment score for sanitation workers would suggest their situation is unremarkable; the underlying documents, hundreds of recorded deaths, disputed classification of those deaths, a legal ban on the practice that caused them not being effectively enforced (Dar, 2026; Outlook, 2026), suggest close to the opposite. The register in which a serious issue is reported, procedural versus emotive, is doing real work here that a compound sentiment score alone cannot surface. This argues for pairing computational sentiment analysis with topic modeling and close reading of representative documents within each topic, rather than treating a sentiment score as a complete answer on its own, a methodological lesson this project takes seriously in its own conclusions.

The finding that academic sources cluster near neutral, rather than at either extreme, is consistent with a mature but unresolved literature, one that has documented both real gains, such as coverage expansion and improved health outcomes in several of the studies reviewed in Section 2 (Dandabathula et al., 2019; Varshney et al., 2025), and real shortfalls, such as usage gaps, toilet design flaws, and insufficient behaviour-change investment (VerKuilen et al., 2023), without either finding displacing the other. This is arguably the correct place for a body of empirical research to sit on a programme of SBM's scale and complexity, and its distinctness from the more polarized government and opinion or journalism framings is itself informative about where in the discourse a reader should look for a more calibrated, less institutionally-motivated picture of SBM's actual record.

It is also worth returning briefly to the framing-theory grounding set out in Section 2.6. Entman's account of framing as selection and salience (Entman, 1993) predicts exactly the source-type gradient this project found: institutions select and foreground the aspects of reality that serve their communicative purpose. What this project adds to that theoretical account is a specific, previously under-discussed mechanism by which framing can obscure severity rather than simply bias direction: even sources that are not deliberately minimizing an issue can, through genre convention and bureaucratic register alone, produce text that a sentiment-scoring tool, and quite possibly an inattentive human reader skimming the same text, does not register as urgent.

7. Limitations of This Project

- Corpus text fields are AI-assisted summaries rather than full original source documents. This is disclosed directly in Section 3.2 rather than only here, because it materially affects how every subsequent result should be interpreted. Results should be read as a validated pipeline output on representative summaries, not a final analysis.
- Corpus size, 55 documents, supports descriptive analysis but is modest for statistical inference. Stakeholder groups with only two or three tagged documents, specifically citizens, marginalized communities, and general public, should be read as indicative rather than robust, since a single unusual document can shift a small group's average considerably.
- Regional coverage within the corpus skews toward northern and western Indian states, specifically Uttar Pradesh, Bihar, Madhya Pradesh, Rajasthan, and Maharashtra. This is a skew also present in the underlying academic literature itself, as documented directly in one of the systematic reviews included in this corpus (VerKuilen et al., 2023), and is not unique to this corpus's construction, though it remains a real limitation on how broadly these findings should be generalized.
- English-language sources only. SBM discourse in Hindi and regional languages, which is likely to differ meaningfully from English-language coverage both in volume and in tone, is not captured by this corpus at all.
- VADER and similar rule-based sentiment tools can misread sarcasm, complex negation, and domain-specific or bureaucratic register (Hutto & Gilbert, 2014). The sanitation-worker finding in Section 4.4 is a direct illustration of this limitation in action, not an exception to it, and should inform how much weight any single sentiment score in this paper is given.
- Stakeholder tags were assigned by a single coder, myself, without independent cross-checking by a second coder. Inter-rater reliability has not been established for these

tags, and stakeholder-level comparisons should be treated as descriptive rather than statistically validated.

- Topic modeling at this corpus size, 55 documents composed largely of short summaries rather than full-length text, produces topics with some genuine overlap, specifically Topics 3 and 4 in Section 4.5. A larger, full-text corpus would likely sharpen topic separation considerably.
- This project reflects a snapshot of published discourse rather than a longitudinal study; it cannot speak to whether sentiment toward SBM by source type or stakeholder group has shifted over the mission's ten-plus-year history, only to the pattern present in the corpus as collected.

8. Conclusion

This project set out to ask how the Swachh Bharat Mission is discursively framed across source types and stakeholder groups, using computational text analysis of a 55-document corpus. The analysis finds a consistent institutional gradient in sentiment, with government and policy sources positive and journalism and opinion sources negative, and a more substantively important stakeholder-level divide, in which marginalized castes and the general public register as the most negatively framed groups, and citizens, public health, and women as the most positively framed. The most important qualification to all of this is that sanitation workers' near-neutral sentiment score reflects the flat, procedural register in which their circumstances are typically reported, not the actual severity of those circumstances, a distinction this project argues matters considerably for how computational text analysis results should be reported and interpreted in social-impact research more broadly, not only in this specific corpus.

Three extensions would materially strengthen this work. First, and most importantly, replacing the corpus's current summarized text with full original source documents, using the extraction tool described in Appendix D, would allow sentiment and topic-modeling results to be computed on authors' actual language rather than on paraphrase, which is likely to shift specific scores even if the broad patterns identified here hold up. Second, expanding the corpus, particularly toward underrepresented southern and eastern states and toward Hindi-language sources, would test whether these findings generalize beyond the current corpus's regional and linguistic scope. Third, independent cross-validation of stakeholder tagging by a second coder would allow the stakeholder-level findings, currently descriptive, to be reported with a formal inter-rater reliability measure rather than resting on my judgment alone.

None of these extensions would be expected to overturn the central finding of this paper, that SBM's framing, like its documented impact, is uneven across who is speaking and who is spoken about, but each would allow that finding to be stated with meaningfully more confidence than the present analysis alone can support.

Acknowledgements

Thank you to my mentors at Ernst & Young, Gaurav Thapak (GT), Consultant, and Amrita Saib (AS), Director, for their guidance on the research design, methodology, and scope of this project,

and for encouraging me to treat an unexpected result, the sanitation-worker sentiment finding, as worth investigating rather than as an error to be corrected away. GT also reviewed and edited the paper.

References

All India Radio News [AIR News]. (2024, October 2). PM Modi addresses Swachhata Hi Seva 2024 programme in New Delhi. All India Radio News. <https://www.newsonair.gov.in/pm-modi-addresses-swachhata-hi-seva-2024-programme-in-new-delhi>

Behera, M. R., Parida, S., Pradhan, H. S., Priyabadini, S., Dehury, R. K., & Mishra, B. (2022). Household sanitation and menstrual hygiene management among women: Evidence from household survey under Swachh Bharat (Clean India) Mission in rural Odisha, India. *Journal of Family Medicine and Primary Care*, 11(3), 1100–1108. https://doi.org/10.4103/jfmpc.jfmpc_1593_21

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Borah, P. B. (2026). Menstrual health and hygiene practices in higher educational institutions: A case study of female students in Dibrugarh, Assam, India. *International Journal of Community Medicine and Public Health*, 13(6), 3077–3082. <https://doi.org/10.18203/2394-6040.ijcmph20261799>

Boukes, M., van de Velde, B., Araujo, T., & Vliegthart, R. (2020). What's the tone? Easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures*, 14(2), 83–104.

The Sentinel. (2025, March 27). CAG audit flags Assam govt's failure in utilizing Rs 350 cr SWM funds. The Sentinel. <https://www.sentinelassam.com/north-east-india-news/assam-news/cag-audit-flags-assam-govts-failure-in-utilizing-rs-350-cr-swm-funds>

Chong, D., & Druckman, J. N. (2007). Framing theory. *Annual Review of Political Science*, 10, 103–126.

Coffey, D., & Spears, D. (2017). *Where India goes: Abandoned toilets, stunted development and the costs of caste*. HarperCollins India.

Comptroller and Auditor General of India. (2018). Report of the Comptroller and Auditor General of India on social, general and economic sectors (non-PSUs) for the year ended March 2017, Government of NCT of Delhi (Report No. 3 of 2018). <https://cag.gov.in/en/audit-report/details/43837>

Comptroller and Auditor General of India. (2025). Report of the Comptroller and Auditor General of India on solid waste management in urban areas, Government of Uttar Pradesh (Report No. 1 of 2025). <https://cag.gov.in/en/audit-report/details/121642>

Curtis, V. (2019). Explaining the outcomes of the ‘Clean India’ campaign: Institutional behaviour and sanitation transformation in India. *BMJ Global Health*, 4(5), e001892. <https://doi.org/10.1136/bmjgh-2019-001892>

Dandabathula, G., Bhardwaj, P., Burra, M., Rao, P. V. V. P., & Rao, S. S. (2019). Impact assessment of India’s Swachh Bharat Mission – Clean India Campaign on acute diarrheal disease outbreaks: Yes, there is a positive change. *Journal of Family Medicine and Primary Care*, 8(3), 1202–1208. https://doi.org/10.4103/jfmprc.jfmprc_144_19

Dar, U. A. (2026, August 5). Nearly 500 sewer, septic tank cleaning deaths since 2019, Centre tells LS. *Business Standard*. https://www.business-standard.com/india-news/sanitation-workers-sewer-septic-tank-cleaning-deaths-since-2019-126080500881_1.html

Deccan Herald. (2020, September 21). 288 people died cleaning sewers, septic tanks in last 3 years: Government. *Deccan Herald*. <https://www.deccanherald.com/amp/national/288-people-died-cleaning-sewers-septic-tanks-in-last-3-years-government-890812.html>

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6), 570–606.

Doron, A., & Raja, I. (2015). The cultural politics of shit: Class, gender and public space in India. *Postcolonial Studies*, 18(2), 189–207.

Deccan Herald (2022). ED probes BBMP for ‘misusing’ Rs 92 crore of Swachh Bharat funds. *Deccan Herald*. <https://www.deccanherald.com/india/karnataka/bengaluru/ed-probes-bbmp-for-misusing-rs-92-crore-of-swachh-bharat-funds-1068395.html>

Employment of Manual Scavengers and Construction of Dry Latrines (Prohibition) Act, 1993, No. 46, Acts of Parliament, 1993 (India).

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4), 51–58. <https://doi.org/10.1111/j.1460-2466.1993.tb01304.x>

Exum, N. G., Gorin, E. M., Sadhu, G., Khanna, A., & Schwab, K. J. (2020). Evaluating the declarations of open defecation free status under the Swachh Bharat (‘Clean India’) Mission: Repeated cross-sectional surveys in Rajasthan, India. *BMJ Global Health*, 5(3), e002277. <https://doi.org/10.1136/bmjgh-2019-002277>

ETV Bharat. (2025, December 24). 52 sanitation workers lost lives due to hazardous cleaning of sewer, septic tanks in a year. *ETV Bharat*.

<https://www.etvbharat.com/en/bharat/52-sanitation-workers-lost-lives-due-to-hazardous-cleaning-of-sewer-septic-tanks-enn25122405001>

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Harvard University Press.

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.

Gupta, A., Khalid, N., Desphande, D., Hathi, P., Kapur, A., Srivastav, N., Vyas, S., Spears, D., & Coffey, D. (2019). Changes in open defecation in rural north India: 2014–2018 (IZA Discussion Paper No. 12065). IZA Institute of Labor Economics.

Hueso, A., & Bell, B. (2013). An untold story of policy failure: The Total Sanitation Campaign in India. *Water Policy*, 15(6), 1001–1017.

Human Rights Watch. (2014). *Cleaning human waste: “Manual scavenging,” caste, and discrimination in India*. Human Rights Watch.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.

Jain, A., Wagner, A., Snell-Rood, C., & Ray, I. (2020). Understanding open defecation in the age of Swachh Bharat Abhiyan: Agency, accountability, and anger in rural Bihar. *International Journal of Environmental Research and Public Health*, 17(4), 1384. <https://doi.org/10.3390/ijerph17041384>

Kabir, Y., Kulkarni, A., Tahiliani, B., Chandrasekar, R., & Arora, R. (2020). Menstrual hygiene management programming in Maharashtra: The journey of a decade (WASH Field Note FN/33/2020). UNICEF. https://susana.org/_resources/documents/default/3-4000-7-1608569262.pdf

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2), 176–187.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11), 39–41.

Ministry of Drinking Water and Sanitation. (2017, April 3). Guidelines on gender issues in sanitation. Government of India. <http://www.mdws.gov.in/sites/default/files/Guidelines%20on%20Gender%20issues%20in%20Sanitation.pdf>

Ministry of Housing and Urban Affairs [MoHUA], Government of India. (2021). Swachh Bharat Mission-Urban 2.0 guidelines. <https://mohua.gov.in>

The Wire. (2019, October 2). Modi declares rural India open defecation free, questions remain. The Wire.
<https://thewire.in/government/modi-declares-rural-india-open-defecation-free-questions-remain>

Novotný, J., Borde, R., Ficek, F., & Kumar, A. (2024). The process, outcomes and context of the sanitation change induced by the Swachh Bharat Mission in rural Jharkhand, India. BMC Public Health, 24, 997. <https://doi.org/10.1186/s12889-024-18388-y>

Outlook. (2026). India's sewer death crisis: Why fatalities don't count as manual scavenging. Outlook.
<https://www.outlookindia.com/national/indias-sewer-death-crisis-why-fatalities-dont-count-as-manual-scavenging>

Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3), 130–137.

Press Information Bureau [PIB]. (2022). Swachh Bharat Mission (Grameen & Urban): Himachal Pradesh – India's second open defecation free state [Explainer]. Government of India.
<https://static.pib.gov.in/WriteReadData/specificdocs/documents/2022/sep/doc2022926109401.pdf>

Prime Minister's Office [PMO]. (2021, October 12). Cabinet approves the continuation of Swachh Bharat Mission (Urban) [SBM U] till 2025-26 for sustainable outcomes. Government of India.
https://www.pmindia.gov.in/en/news_updates/cabinet-approves-the-continuation-of-swachh-bharat-mission-urban-sbm-u-till-2025-26-for-sustainable-outcomes/

Prohibition of Employment as Manual Scavengers and their Rehabilitation Act, 2013, No. 25, Acts of Parliament, 2013 (India).

Safai Karamchari Andolan v. Union of India, (2014) 11 SCC 224 (India).

SabrangIndia. One death every few days, zero accountability: Sanitation workers protest at Jantar Mantar. (2026, August). One death every few days, zero accountability: Sanitation workers protest at Jantar Mantar. SabrangIndia.
<https://sabrangindia.in/one-death-every-few-days-zero-accountability-sanitation-workers-protest-at-jantar-mantar/>

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.

Sarkar, A. K., & Biswas, S. (2019). Toilet usage situation under 'Swachh Bharat Mission' in rural West Bengal: Need integration in strategies of intervention. Social Work Chronicle, 8(1), 1–15. <https://publishingindia.com/storage/PDFBrochures/5540.pdf>

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of Communication*, 49(1), 103–122.

Semetko, H. A., & Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of Communication*, 50(2), 93–109.

Seth, P., & Jain, P. (2021). NFHS-5 data can help India become a truly open defecation free country. *The Wire*. <https://thewire.in/government/nfhs-5-sanitation-open-defecation>

Sinha, R. N., & Paul, B. (2018). Menstrual hygiene management in India: The concerns. *Indian Journal of Public Health*, 62(2), 71–74. https://doi.org/10.4103/ijph.ijph_135_18

UNICEF India. (2020). Financial and economic impacts of the Swachh Bharat Mission. <https://www.unicef.org/india/media/1206/file/Financial-and-Economic-Impacts.pdf>

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2), 121–140.

Varshney, K., Khatri, C., & Lamichhane, P. (2025). Impacts of Swachh Bharat Mission on health and communicable disease control: A systematic review protocol. *Health Science Reports*, 8(2). <https://onlinelibrary.wiley.com/doi/full/10.1002/hsr2.70425>

VerKuilen, A., Sprouse, L., Beardsley, R., Lebu, S., Salzberg, A., & Manga, M. (2023). Effectiveness of the Swachh Bharat Mission and barriers to ending open defecation in India: A systematic review. *Frontiers in Environmental Science*, 11, 1141825. <https://www.frontiersin.org/journals/environmental-science/articles/10.3389/fenvs.2023.1141825/full>

World Health Organization [WHO] India. (2018, July 27). Health gains from the Swachh Bharat initiative [News release]. <https://www.who.int/india/news/item/27-07-2018-health-gains-from-the-swachh-bharat-initiative>

Yadavar, S. (2019, January). 4 yrs of Swachh Bharat: Toilet use, construction spree don't match. *The Quint* (republished from IndiaSpend). <https://www.thequint.com/amp/story/news/india/4-yrs-of-swacch-bharat-toilet-use-construction-spreedont-match>

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2), 205–231.

Individual citations for all 55 corpus source documents, including publication, date, and URL, are listed in full in Appendix A and in the accompanying source corpus file. Author names were not consistently available at the point of corpus construction for all sources and should be verified against original publications before any formal submission or publication of this paper.

Appendix A: Full Source Corpus (55 Documents)

The table below lists every document in the corpus analyzed in this paper, with its source type and publication date. Full titles, source names, stakeholder tags, and source URLs are included in the accompanying corpus file (sbm_corpus.csv / sbm_corpus.xlsx), which should be treated as part of this paper's supporting material.

ID	Title	Source Type	Date
1	Impact of Swachh Bharat Abhiyan (research review)	Academic paper	2022
2	Swachh Bharat Mission: A Case Study	Academic paper	2022
3	Impacts of SBM on Health and Communicable Disease Control (systematic	Academic paper	2025
4	Financial and Economic Impacts of the Swachh Bharat Mission (Gramin)	NGO / dev. bank report	2020
5	A Study on Implication of Swachh Bharat among Students	Academic paper	2023
6	Swachh Bharat Abhiyan: SWOT Analysis	Academic paper	2016
7	Effectiveness of the Swachh Bharat Mission and barriers to ending open	Academic paper	2023
8	Swachh Bharat Mission - Current Affairs Briefing	Policy briefing	2024
9	The Reality of the Swachh Bharat Mission	Opinion / analysis	2024
10	Why the Swachh Bharat Abhiyan Failed	Opinion / analysis	2025
11	How the Swachh Bharat Mission is heading for failure	Investigative journalism	2017
12	The failure of Swachh Bharat Abhiyan: we need a better way to manage t	Opinion / analysis	2018
13	Swachh Bharat has made little difference, govt should factor in critic	Opinion / analysis	2018
14	PM Modi addresses Swachhata Hi Seva 2024 programme	Government release	2024-10-02
15	Union Minister C R Patil urges people to undertake Shramdaan for Swach	Government release	2024-10-01
16	India rejects UN official's criticism (manual scavenging and caste)	News report	2017
17	One Death Every Few Days: Sanitation Workers Protest at Jantar Mantar	Investigative journalism	2026
18	Nearly 500 sewer, septic tank cleaning deaths since 2019, Centre tells	News report	2026

ID	Title	Source Type	Date
19	52 Sanitation Workers Lost Lives Due to Hazardous Cleaning in a Year	News report	2025
20	India's Sewer Death Crisis: Why Fatalities Don't Count as Manual Scave	Investigative journalism	2026
21	Swachh Survekshan 2025: Indore cleanest city for 8th consecutive year	Government release	2025
22	CAG report flags irregularities in Uttar Pradesh's waste management	News report	2025
23	CAG flags procurement deficiencies in Delhi's waste management under S	News report	2020
24	CAG audit flags Assam government's failure to utilize SWM funds	News report	2025
25	Not a single toilet built in Delhi under Swachh Bharat Mission, CAG re	Investigative journalism	2018
26	ED probes BBMP for misusing Swachh Bharat funds in Bengaluru	Investigative journalism	2022
27	Understanding Open Defecation in the Age of SBM: rural Bihar ethnography	Academic paper	2020
28	Toilet usage situation under SBM in rural West Bengal	Academic paper	2019
29	Sanitation change in rural Jharkhand under SBM	Academic paper	2024
30	Toilet usage behavior study, rural Bihar households	Academic paper	2021
31	After 4 years of Swachh Bharat, toilet use does not match construction	Investigative journalism	2019
32	Information is key to making India open-defecation-free	Academic paper	2025
33	NFHS-5 Data Can Help India Become a Truly Open Defecation Free Country	News report	2021
34	Impact assessment of SBM on acute diarrheal disease outbreaks	Academic paper	2019
35	Menstrual Hygiene Management in India	Academic paper	2018
36	Household sanitation and menstrual hygiene among women in rural Odisha	Academic paper	2022
37	Menstrual health and hygiene practices among female students in Assam	Academic paper	2026
38	Menstrual Hygiene Management Programming in Maharashtra	NGO / dev. bank report	2020
39	India Menstrual Health Literature Review	NGO / dev. bank report	2023

ID	Title	Source Type	Date
40	Health, Gender, and Economic Benefits of India's Swachh Bharat Mission	Policy briefing	2024
41	Swachh Bharat Abhiyan (Clean India Mission)	Government release	2020
42	Swachh Bharat Mission 2.0 - Comprehensive Municipal Waste Management i	NGO / dev. bank report	2023
43	Waste-Wise Cities: Best Practices in Municipal Solid Waste Management	Policy briefing	2021
44	Global South Nears Waste Tipping Point, World Bank Warns	News report	2026
45	Swachh Bharat Mission 2.0: Garbage-Free Urban India	Policy briefing	2024
46	Government reports urban SBM achievements: star-rated garbage-free cit	Government release	2020
47	Solid Waste Management and Contribution of Swachh Bharat Mission in Sh	Academic paper	2023
48	Cabinet approves continuation of Swachh Bharat Mission (Urban) till 20	Government release	2021
49	Jal Shakti ministry targets 2025 for Swachh Bharat Mission-Grameen aft	News report	2026
50	Swachh Bharat Mission - Progress and Problems	Policy briefing	2017
51	Centre releases Rs 1,869 crore for Swachh Bharat Mission-Grameen in 20	Government release	2024-12-10
52	Toward a Menstrual-friendly School Environment	Academic paper	2025
53	Guidelines on Gender Issues in Sanitation for Swachh Bharat Mission (G	Policy briefing	2017
54	288 people died cleaning sewers, septic tanks in last 3 years: Governm	News report	2020
55	Two sanitation workers die cleaning sewage shaft at Noida house	News report	2024-05-04

Appendix B: Note on AI Assistance

AI tools were used as a supporting aid during the research and writing process. They were mainly used to help identify and organize relevant pieces of literature, briefly summarize some sources during the initial literature-collection stage, assist with minor coding and troubleshooting, and improve the wording of some sections of the report. The selection of literature, research question, methodology, analysis, interpretation of results, and final conclusions were my own. AI-generated suggestions and summaries were reviewed and edited before being used, and I take responsibility for the final content and conclusions of this project.

Appendix C: Stakeholder Tagging Codebook

Each document was tagged with the stakeholder group or groups it primarily concerns, based on the following fixed criteria, applied consistently across the corpus:

- A stakeholder tag was assigned when a document's central subject, not merely a passing mention, concerned that group's experience, treatment, or outcomes under SBM.
- A document could receive multiple tags where it substantively addressed more than one group, for example, a study on female sanitation workers is tagged both women and sanitation_workers.
- Tags were drawn from a fixed vocabulary of fifteen categories, listed below, rather than created individually per document, to keep counts comparable across the corpus.
- government — documents primarily concerning central or state government policy, funding, or communications.
- sanitation_workers — documents primarily concerning the people who physically carry out sanitation and waste-management labor, including sewer and septic-tank workers.
- rural_households — documents primarily concerning household-level sanitation access and behavior in rural areas.
- urban_local_bodies — documents primarily concerning municipal corporations, city governments, or urban infrastructure management.
- women — documents primarily concerning women's experience of sanitation, including menstrual hygiene management.
- students — documents primarily concerning school-age or university-age populations, including school WASH facilities.
- marginalized_castes — documents primarily concerning caste-based dimensions of sanitation labor or access, including manual scavenging.
- marginalized_communities — documents primarily concerning other socially or economically marginalized groups not specifically defined by caste.
- public_health — documents primarily concerning disease outcomes, mortality, or other health impacts attributable to sanitation conditions.
- policy_experts — documents primarily framed around policy analysis or expert commentary rather than a specific affected population.
- ngo_sector — documents primarily concerning the role of non-governmental or civil-society organizations in SBM implementation.
- general_public — documents primarily addressing the Indian public broadly, without a more specific stakeholder focus.
- citizens — documents primarily framed around citizen participation or civic duty under SBM.
- infrastructure — documents primarily concerning physical infrastructure, such as toilets, sewage systems, or waste facilities, as a subject in itself.

- `local_administration` — documents primarily concerning district- or panchayat-level administrative implementation, distinct from state or national government.

Limitation: as stated in Sections 3.2 and 6, tagging was performed by a single coder without a second, independent coder cross-checking assignments, so inter-rater reliability has not been established for this codebook's application. This should be addressed before these tags are used to support any claim stronger than a directional or descriptive one.

Appendix D: My Reflection on Planning and Running This Project

This appendix is written in my own voice, reflecting on how I actually went about this project, not just what I found.

D.1 Why I Chose Text Analysis (and What Else I Considered)

I wanted to actually test the common but rarely examined claim that SBM “looks different” depending on who you ask, rather than just repeating that claim based on a handful of articles I happened to read. As described in Section 3.1, I considered interviews and a traditional literature review before settling on computational text analysis across a broad corpus, because it was the only approach that let me systematically compare sentiment and word choice across dozens of sources at once rather than relying on my own impressionistic read of a small sample.

D.2 How I Planned My Time

I broke this project into six stages: background reading and identifying the four thematic strands described in Section 2; designing the stakeholder-tagging codebook and corpus structure; collecting and coding sources; building and testing the preprocessing and analysis pipeline in Python; running the full pipeline and interpreting the output; and writing this report. I underestimated how long corpus collection would take relative to the analysis itself. Finding and reading enough sources to tag stakeholders sensibly took considerably longer than writing and running the actual analysis code, which is worth knowing going in if I, or anyone else, were to repeat a project like this.

D.3 People and Tools I Used

I built and ran the analysis pipeline in Python, using Google Colab so that I would not need to install anything locally, and used Jupyter notebooks throughout so that I could see the output of each processing step immediately rather than only seeing a final result. I used Microsoft Excel to manage and review the corpus as it grew, since a spreadsheet made it far easier to check for duplicate or near-duplicate sources and to see at a glance whether my source-type and stakeholder-tag categories were reasonably balanced, rather than accidentally over-representing one type of source.

D.4 Problems I Ran Into, and How I Fixed Them

- Early in corpus construction, I ended up with two documents covering essentially the same underlying story, a UN human-rights dispute over manual scavenging, pulled from

overlapping searches without my noticing at the time. I caught this by reviewing summaries side by side before finalizing the corpus, and replaced the duplicate with a genuinely distinct source, NFHS-5 survey data, instead.

- My first attempt at a sentiment-analysis approach used a small, hand-built list of positive and negative words rather than a validated tool. This produced results that were directionally plausible but not something I could defend methodologically if asked why I chose those specific words. I replaced it with VADER, a peer-reviewed, validated sentiment model, once I understood that hand-building a lexicon was a reasonable way to prototype quickly but not a reasonable way to report a final result.
- Corpus text is composed of summaries rather than full original documents, for the copyright reasons discussed in Section 3.2. This was not something I could fully solve within the time I had, so instead I built a separate tool, using pdfplumber for PDF extraction and manual transcription for sites that block automated scraping, to progressively replace summaries with real source text as an explicit next step, rather than either ignoring the limitation or claiming it was fully resolved when it was not.
- My first run of topic modeling, before I had settled on lemmatization as the primary token representation, produced topics that were noticeably harder to interpret, mixing grammatical variants of the same underlying word as if they were unrelated. Switching from raw tokens to lemmatized tokens before vectorization fixed this.

D.5 What I Learned

The biggest thing I learned was that a finding can be genuinely interesting precisely because it does not match what I expected going in. I expected sanitation workers to score as the most negatively framed stakeholder group in this corpus, given what I already knew about worker deaths and manual scavenging before I ran any analysis. When the actual sentiment score came back close to neutral, my first instinct was to assume something had gone wrong with the code. It took re-reading the underlying documents to realize the score was not wrong; it was accurately reflecting that this material is written in flat, procedural language, which is itself a finding worth reporting rather than a bug worth silently fixing. I also learned, more practically, that building a good, balanced corpus by hand takes real, deliberate effort and does not happen automatically just because a topic has a lot of coverage available online. Without deliberately searching for underrepresented angles, such as worker safety, fund audits, and menstrual hygiene, my corpus would have skewed heavily toward whichever storyline was easiest to find first.