Positive vs. Negative Collocates of Indian-Origin Words in Western Media: A Corpus-Based Critical Discourse Analysis

Dinesh Deckker[1], Subhashini Sumanasekara[2], Sree Lakshmi Ammanamanchi[3]

[1]Faculty of Arts, Science and Technology, Wrexham University, United Kingdom.

[2]Faculty of Computing and Social Sciences, University of Gloucestershire, United Kingdom.

[3]Faculty Member at the University of Technology and Applied Sciences - Al Mussanah, Sultanate of Oman

DOI: https://dx.doi.org/10.47772/IJRISS.2025.906000398

Received: 15 June 2025; Accepted: 19 June 2025; Published: 19 July 2025

ABSTRACT

This study explores how Indian-origin lexical items—such as *karma*, *chai*, *Bollywood*, and *yoga*—are framed evaluatively within Western English-language media. Drawing on the News on the Web (NOW) corpus (2015–2023) and guided by the principles of Corpus-Assisted Discourse Studies (CADS), the research investigates whether these borrowed cultural terms are systematically associated with positive, negative, or neutral sentiments in journalistic discourse.

Fifteen Indian-origin words were selected based on frequency, cultural relevance, and dictionary inclusion. Using ±5 word collocation windows, the most frequent co-occurring words were extracted via AntConc and Sketch Engine. Collocates were manually annotated for sentiment polarity—positive, neutral, or negative—supported by using lexicons (SentiWordNet, NRC), and their accuracy was verified through inter-coder reliability. Sentiment scores were statistically analysed using Chi-square, ANOVA, and Kruskal-Wallis tests.

Results revealed that terms like *yoga* and *ayurveda* were predominantly associated with positive collocates, whereas *curry* and *Bollywood* attracted significantly more negative sentiment. The Chi-square test confirmed significant variation in sentiment across keywords ($p = .035$), although ANOVA and Kruskal-Wallis tests were not statistically significant. Visualizations such as heatmaps and stacked bar charts highlighted distinct sentiment patterns, reflecting broader ideological framing.

The study concludes that Indian-origin words are not just linguistically assimilated but ideologically recontextualised in Western discourse, often affirming wellness culture while trivialising or exoticising ethnic identities. These findings offer critical insights into semantic prosody, media representation, and postcolonial power dynamics in global Englishes.

Keywords: Indian-origin words, semantic prosody, corpus linguistics, sentiment analysis, media discourse, cultural framing, NOW corpus, Western media, collocates, postcolonial linguistics

INTRODUCTION

Importance of the Topic

In an era of intensifying global communication and cultural exchange, Indian-origin lexical items such as *karma*, *guru*, *chai*, *Bollywood*, and *yoga* have become embedded within Western media discourse. While these terms enrich the global English lexicon, their semantic environments—particularly the collocates with which they appear—can subtly reveal how non-Western cultures are portrayed, stereotyped, or commodified in Anglophone contexts. As language reflects societal ideologies (Fairclough, 1995), examining the positive or negative evaluative meaning assigned to Indian-origin words via collocational patterns is vital for understanding how Western media construct or distort cultural identity.

This inquiry is especially urgent amid growing scholarly interest in linguistic representation, cultural appropriation, and discourse framing. Collocation, a core concept in corpus linguistics, has been demonstrated to reflect hidden attitudes through semantic prosody, where a neutral word acquires a positive or negative connotation based on its frequent neighbours (Louw, 1993; Hunston, 2007). In this light, analysing the collocational profiles of Indian-origin terms within large-scale media corpora offers a powerful lens for assessing cultural valuation, stereotyping, or exoticisation.

The relationship between corpus linguistics and discourse analysis has been firmly established in works such as Baker et al. (2008), who illustrate how corpora can be used to reveal implicit bias in discourse. Stubbs (2001) emphasises that repeated co-occurrence patterns help build shared cultural meanings, and Partington et al. (2013) extend this by showing how evaluative language reveals ideologies embedded in news texts.

In the context of Indian-origin lexical items, scholars have noted a recurring pattern of orientalism and trivialisation. For example, Sharma et al. (2009) argue that words like *karma* and *guru* have been divorced from their philosophical roots and recontextualised within self-help or consumerist frameworks in Western texts. Similarly, Rajagopalan (2001) highlights the polysemic drift of *yoga* and *chai* in Western advertising as instances of cultural misrepresentation. However, these studies often rely on anecdotal or small-scale textual analysis without empirically testing semantic prosody across broad datasets.

Corpus-based research into ethnic or cultural lexical items has been growing, with Baker (2010) analysing refugee metaphors in the UK press, and Mahlberg (2013) applying corpus techniques to literary representations of ethnicity. However, a gap remains in studying Indian-origin terms through collocation-based sentiment mapping.

Research Gap

Despite the rise of global Englishes and extensive work on language and ideology, no comprehensive study has applied corpus-based collocation and semantic prosody analysis to Indian-origin lexical items across Western media outlets. Current literature tends to focus narrowly on discourse features or rely on qualitative analysis without leveraging the empirical power of large corpora, such as the NOW (News on the Web) corpus.

Moreover, while semantic prosody has been applied to detect evaluative tendencies (Sinclair, 1996; Hunston, 2007), this tool has not been systematically used to analyse how Indian terms are emotionally framed—positively (e.g., *yoga retreat*, *karma balance*) or negatively (e.g., *curry smell*, *desi gang*)—in different media genres such as lifestyle, crime, and immigration reporting.

Research Aim

This study investigates the collocational environment and semantic polarity of selected Indian-origin lexical items in Western English-language media. Using the NOW corpus and tools such as AntConc and LancsBox, it identifies the most frequent collocates within a ±5 word window. It categorises them into **positive, negative, or neutral sentiment** using sentiment lexicons and manual coding.

This research aims to answer:

- What is the most frequent Indian-origin lexical items in Western English media?
- What semantic prosodies—positive, negative, or neutral—emerge from their collocates?
- How do different thematic contexts (e.g., health, religion, crime) influence the polarity of these collocates?
- What do these patterns suggest about the representation of Indian culture in the Anglophone media sphere?

Main Contributions

This research contributes to corpus-assisted discourse studies (CADS) and the broader field of global Englishes by:

- Mapping the emotional and ideological terrain surrounding Indian-origin words through statistical collocation analysis, offering empirical insight into how lexical sentiment varies across cultural domains.
- Introducing a quantitative framework for semantic bias detection, rooted in real-time, large-scale news corpora, that can be extended to other minority language representations in global media.
- Bridging postcolonial linguistic critique with empirical corpus linguistics, demonstrating how Indian cultural identity is linguistically framed—both positively and problematically—within contemporary Western news discourse.

LITERATURE REVIEW

Corpus Linguistics and Collocation Theory

The discipline of corpus linguistics has revolutionized how scholars study language, offering empirical insights into naturally occurring usage rather than relying solely on introspective intuition. At the heart of this paradigm lies the concept of collocation**,** defined by Sinclair (1991) as the tendency of words to appear together more frequently than would be expected by chance. Collocation provides a window into habitual patterns of language use, revealing not just grammatical or syntactic associations but also cognitive and cultural structures of meaning.

The theoretical foundations of collocation analysis are crucial for understanding linguistic ideology. While early structuralists focused on syntagmatic and paradigmatic relations, corpus linguistics enabled the identification of statistically significant co-occurrence patterns in vast text corpora. Sinclair's (1991) "idiom principle" emphasises that meaning is often constructed from semi-preassembled chunks of language, challenging the compositional assumption of meaning being derived word-by-word. This insight has direct implications for how cultural terms, such as Indian-origin words, function in discourse, as their meanings may shift based on the company they keep.

Further extending Sinclair's groundwork, Stubbs (2001) argued that collocations serve as cultural indicators, embedding ideological norms and evaluative tendencies within seemingly neutral language. For instance, if *immigrant* frequently co-occurs with *illegal*, *burden*, or *flood*, these collocations do not merely reflect reality; they help construct it. This notion positions collocational analysis as a critical tool for discourse examination, especially when analysing culturally sensitive terms whose meanings are shaped more by use than by dictionary definitions.

Moreover, collocation analysis allows for the identification of lexical priming**,** as articulated by Hoey (2005), where frequent exposure to certain word pairs conditions speakers to expect particular associations. In this light, repeated collocational patterns can create a semantic "gravitational field" around specific words, either reinforcing stereotypes or subverting them. Analysing the collocational behaviour of Indian-origin words can thus help reveal whether they are embedded in affirming or pejorative semantic environments within Western media.

In sum, the theoretical apparatus of corpus linguistics—especially collocation theory—offers a robust methodological and conceptual framework for this study. It facilitates the exploration of not just how frequently Indian-origin terms appear but how they are framed semantically through their lexical companions.

Semantic Prosody and Ideological Framing

The concept of semantic prosody is crucial for understanding how language conveys evaluative meaning beyond literal definitions. Introduced by Louw (1993) and developed by Sinclair (1996), semantic prosody refers to the tendency of certain words to co-occur with either positively or negatively valenced terms regularly. This repeated co-occurrence influences how readers or listeners perceive these words over time. For instance, the verb *cause* often appears in phrases like *cause damage*, *cause harm*, or *cause death*, which gives it a negative semantic prosody despite its neutral dictionary definition.

The significance of semantic prosody lies in its subtlety: it embeds ideological framing within ostensibly neutral discourse. Stubbs (2001) and Hunston (2007) emphasised that semantic prosody is not merely a linguistic

curiosity but a social and psychological phenomenon, reflecting shared cultural evaluations. In the context of media discourse, it becomes an instrument of discursive manipulation**,** shaping audience perception through implicit cues. Words like *refugee*, *immigrant*, or *protester* may carry negative semantic prosodies depending on their collocational patterns in journalistic writing, thus contributing to the reproduction of political or cultural biases (Baker et al., 2008).

When this lens is applied to Indian-origin words in Western English—terms such as *guru*, *karma*, *chai*, or *Bollywood*—a pattern of semantic bleaching and trivialisation begins to emerge. Scholars like Sharma et al. (2009) have demonstrated that these words, when embedded in religious, philosophical, or traditional contexts, are often resemanticized in Western discourse to denote superficial trends or lifestyle branding (e.g., *tech guru*, *bad karma*, *chai latte*). This shift reflects not just linguistic change but a broader cultural phenomenon in which the spiritual and historical significance of Indian-origin terms is diluted for mass consumption.

This kind of discursive appropriation aligns with Edward Said's (1978) concept of Orientalism**,** wherein non-Western cultures are systematically represented as exotic, irrational, or subordinate to the West. Semantic prosody becomes a linguistic vehicle through which Orientalist narratives are perpetuated, not through overt statementsbut through collocational patterning that reinforces reductive frames. For example, the frequent pairing of *Bollywood* with terms like *glitz*, *over-the-top*, or *escapist* in Western media can create a semantic field that reduces an entire cinematic tradition to spectacle and excess, overshadowing its political, aesthetic, and grassroots dimensions.

Furthermore, the framing of specific Indian terms in negative semantic environments can reinforce cultural hierarchies and social exclusion. A term like *curry*, while seemingly benign, is often associated with words such as *smell*, *stain*, or *overpowering*—semantic partners that subtly evoke otherness and discomfort (Bose, 2007). Such patterns not only inform individual word perceptions but also contribute to macro-level representations of Indian identity in the media.

By systematically analysing the collocational profiles and prosodic tendencies of Indian-origin words, scholars can move beyond surface-level etymology and uncover the evaluative frameworks that shape intercultural perception. Semantic prosody, thus, is not just about meaning; it is about power—the power to frame, stereotype, and naturalise cultural imaginaries.

In the context of this study, semantic prosody serves as a diagnostic tool for assessing how the concept of Indian identity is semantically constructed within Western news discourse. It enables the identification of patterns that may not be immediately apparent through qualitative reading alone, underscoring the need for corpus-based approaches assumptions.

Corpus-Assisted Discourse Studies (CADS)

Corpus-assisted discourse studies (CADS) represent a methodological and theoretical convergence between critical discourse analysis (CDA) and corpus linguistics, enabling researchers to examine the ideological structures embedded in language use empirically. As Baker et al. (2008) argue, CADS offers a productive synergy by combining the quantitative rigour of corpus tools with the interpretive depth of discourse analysis. These fusion addresses longstanding critiques of CDA's anecdotal tendencies by grounding claims about linguistic ideology in statistically verifiable patterns of usage.

At its core, CDA is concerned with how language reproduces power relations, social inequality, and cultural ideologies (Fairclough, 1995; Fowler, 1991; Littlejohn, 1999). Fowler (1991) demonstrates that news discourse is not a transparent reflection of events, but a site where ideological framing occurs through lexical choice, grammar, and narrative structure. Similarly, Littlejohn (1999) emphasises that all forms of human communication—including media texts—are shaped by underlying power structures and cultural assumptions. These foundational insights make CDA particularly suitable for interrogating media portrayals of minority or non-Western groups.

According to Partington et al. (2013), CADS operates under the assumption that no language use is ideologically neutral. Repeated collocational patterns and textual conventions are not merely linguistic habits but encode

shared assumptions, biases, and social positions. Through the use of frequency counts, collocate lists, keyword analysis, and concordance patterns, CADS identifies "textual fingerprints" of dominant ideologies that are often overlooked in traditional close reading.

The application of CADS has yielded critical insights into the media's portrayal of minority groups, migrants, and non-Western cultures. For instance, Baker and McEnery (2005) analysed representations of refugees and asylum seekers in UK newspapers, finding consistent metaphorical framings such as *flood*, *wave*, or *swarm*—terms that evoke threat and dehumanisation. Similarly, Baker and Gabrielatos (2008) demonstrated how discourse about Muslims in British media disproportionately associated them with violence, radicalism, and otherness. These studies exemplify how CADS reveals linguistic strategies of exclusion, even when overtly discriminatory language is absent.

In the context of Indian-origin words in Western discourse, CADS provides a methodologically robust approach to examining how such terms are embedded within broader ideological narratives. Whereas traditional CDA might focus on a handful of articles, CADS enables the examination of thousands of news texts, uncovering statistically significant trends in how words like *karma*, *Bollywood*, or *curry* are framed over time and across contexts.

More importantly, CADS facilitates interdiscursivity analysis, allowing researchers to trace how the same lexical item may be evaluated differently depending on the thematic domain, such as spirituality, crime, immigration, lifestyle, or entertainment. For example, the word *guru* might be positively collocated with terms like *wisdom* or *enlightenment* in wellness discourse but paired with *fraud*, *cult*, or *controversy* in crime journalism. This kind of analysis provides empirical grounding for assessing the semantic polarization of cultural terms across genres.

Furthermore, CADS accommodates the mixed-methods ethos increasingly favoured in applied linguistics, combining corpus-derived quantitative outputs with interpretive readings of concordance lines and key passages (Baker, 2010). This duality ensures that frequency does not override meaning, and that linguistic patterns are interpreted within their social and cultural contexts.

In relation to this study, CADS provides the scaffolding to examine whether Indian-origin lexical items acquire consistently positive or negative collocational tendencies, and whether such tendencies correlate with specific media frames. It also supports the critical interrogation of whether these patterns reflect neo-Orientalist framing or indicate more nuanced, pluralistic engagements with Indian culture in global media.

Thus, CADS not only bridges the methodological divide between empirical and critical traditions but also operationalises key principles of CDA, as articulated by Fowler (1991), Littlejohn (1999), and Fairclough (1995), by exposing how language in media serves as a tool of ideological construction.

Indian-Origin Lexical Items in Global English

The integration of Indian-origin words into global English varieties is a linguistic consequence of colonial history, cultural exchange, and media globalisation. From early borrowings like *bungalow*, *pundit*, and *thug* to contemporary inclusions such as *chai*, *yoga*, and *Bollywood*, these lexical items reflect the dynamic interplay between English and South Asian languages (Kachru, 1983; Crystal, 2003). However, the semantic fate of these words once absorbed into dominant Western discourse is complex, often involving shifts in meaning, scope, and cultural connotation.

According to McArthur (2002), English has long been a borrowing language, absorbing terms that index otherness, novelty, or exoticism. Indian-origin words are particularly prone to this treatment because they carry strong cultural imagery—symbols of tradition, mysticism, or sensory richness. However, these same qualities make them vulnerable to semantic reduction or cultural commodification in Western media. For instance, *yoga* is often stripped of its spiritual, philosophical, and ethical roots and reduced to a physical wellness regime, marketed to Western consumers (Jain, 2015). Similarly, *chai*, a generic Hindi word for tea, is rebranded as a niche "spiced latte," exoticized in culinary discourse and decoupled from its everyday South Asian context.

This semantic reorientation is not neutral; it reflects and reproduces specific ideological positions**.** Rajagopalan

(2001) observed that the reappropriation of Indian terms in American English often serves to flatten complexity into digestible stereotypes—*Bollywood* becomes synonymous with escapist excess, *karma* with poetic justice, and *guru* with either spiritual depth or deceptive charisma, depending on the context. Such transformations are rarely interrogated in mainstream discourse, leaving linguistic traces of Orientalist framing embedded in everyday speech and writing (Said, 1978).

Moreover, the usage of these words in mass media, particularly in lifestyle journalism, entertainment, advertising, and even crime reporting, can compound these issues by attaching fixed emotional and cultural valence to Indian identity. Bose (2007) argues that terms like *Bollywood* are often deployed in ways that reinforce spectacle, melodrama, and irrationality, a portrayal that subtly reinforces the West's epistemic superiority. The effect is a semantic narrowing, where words carry less of their original cultural weight and more of a Western-assigned performative function.

Although a few linguistic studies have tracked the etymological entry of Indian terms into English dictionaries (e.g., Burchfield, 2000), and others have examined their marketing or branding use**,** there is a dearth of systematic corpus-based analysis on how these words function across discursive contexts and sentiment fields. Sharma et al. (2009), one of the few recent contributions in this space, found that Indian-origin words were often recoded with context-dependent valence—positive in health or spiritual domains, but occasionally negative or trivial in entertainment and political coverage.

This discrepancy raises important questions about semantic prosody and cultural framing**:** Are Indian-origin words consistently associated with positive or negative terms in Western English? Do these associations vary based on the type of discourse? How do these patterns impact broader perceptions of Indian culture? These are empirical questions that can only be meaningfully answered through corpus-driven, quantitative methods—an approach this study adopts.

In sum, while Indian-origin lexical items have become integral to global English, their treatment in Western media requires critical scrutiny. Their incorporation may signal inclusivity on the surface, but their collocational partners and discourse environments often point to deeper ideological tendencies, including stereotyping, cultural flattening, and selective appropriation. This study seeks to move beyond anecdotal examples and contribute a data-driven analysis of how such words are discursively positioned in Anglophone media.

Representation and Stereotyping in Western Media

Media representations are not passive reflections of reality; they are active constructions that shape public consciousness, reinforce cultural hierarchies, and normalize dominant ideologies. This critical perspective, central to cultural studies and discourse analysis, posits that language and imagery in the media serve as sites of ideological production (Hall, 1997). For non-Western cultures, particularly those in the Global South, this representation often takes the form of stereotyping—a process of essentialising and simplifying cultural identities to make them more accessible to Western audiences.

One of the most influential analyses of such patterns comes from Edward Said's (1978) *Orientalism*, which documented how Western discourses construct the East as exotic, backwards, sensual, and irrational—essentially, as the antithesis of the rational and civilised West. Said's framework has been applied widely to media studies, where news and entertainment often reduce entire cultures to a series of predictable tropes. In the case of India, these tropes range from the mystic guru and self-sacrificing mother to the overcrowded slum and chaotic festival.

Language plays a crucial role in reproducing these images. As Van Dijk (1991) explains, the discourse structures of news reporting—lexical choices, framing devices, and headline construction—are often subtly biased**,** even when overtly neutral. These structures become particularly visible in the treatment of ethnic minorities or non-Western countries, where metaphors of threat, irrationality, or inferiority dominate coverage. This insight is echoed in more recent corpus-assisted studies by Baker et al. (2008), who demonstrate how seemingly neutral lexical patterns can carry ideological weight over time.

Indian-origin lexical items embedded in Western discourse can be complicit in this system. Terms like *curry* or

Bollywood, while widely used, are frequently attached to simplistic, overgeneralised collocates—*curry house*, *Bollywood drama*, *spicy curry*, *Bollywood wedding*—that reinforce cultural caricatures rather than acknowledge diversity. Bose (2007) notes that Bollywood, despite its vast internal complexity, is often portrayed in Western outlets as campy, melodramatic, and frivolous—a stark contrast to how Western cinema is discussed. Similarly, *curry* has become a catch-all term for Indian cuisine, erasing the culinary heterogeneity of the subcontinent.

The danger of these reductive associations lies in their cumulative discursive effect**.** Repeated collocation of Indian-origin words with emotionally loaded or trivializing adjectives constructs an affective field in which Indian culture becomes exotic, excessive, or humorous. This framing can spill into more serious domains, such as politics or crime reporting, where words like desi or guru are used about criminal organisations or fraudulent spiritual leaders, reinforcing harmful stereotypes (Sharma et al., 2020). The cumulative result is a semantic environment in which even neutral or positive cultural signifiers acquire an undercurrent of condescension or suspicion.

Moreover, these representations are rarely challenged within the media itself, leading to what Van Leeuwen (2008) calls "symbolic exclusion"—not through absence, but through the narrow framing of presence. Indian culture is visible, even hyper-visible, but mainly on the terms set by Western media logic. Such representations contribute to a semiotic regime in which cultural diversity is acknowledged superficially but rendered safe, marketable, and ideologically harmless.

From a linguistic standpoint, this invites critical inquiry into collocational patterns—what words co-occur with Indian-origin terms, and what evaluative or affective frames they suggest. The shift from qualitative critique to empirical analysis, enabled through corpus-based methods, allows researchers to map the repetition and distribution of stereotypes across genres, time, and outlets.

This study, therefore, views the collocational profiling of Indian-origin lexical items not as a lexical exercise but as a diagnostic of discursive ideology**.** By identifying the frequency, polarity, and thematic clustering of collocates in Western media, it aims to illuminate the underlying affective and ideological economy in which Indian cultural elements circulate.

METHODOLOGY

This chapter outlines the research design, corpus selection, keyword identification process, sentiment annotation procedures, and the statistical methods used to investigate the sentiment framing of Indian-origin lexical items in Western media. The approach draws from corpus linguistics, sentiment analysis, and media discourse studies.

Research Design

This study employed a mixed-method approach grounded in corpus-assisted discourse studies (CADS), combining quantitative sentiment distribution analysis with qualitative interpretation of collocational patterns. The central research aim was to determine whether Indian-origin words are evaluatively framed differently across Western English-language news media, and to what extent such sentiment variation reflects broader discourse ideologies.

Corpus Selection and Scope

The study used the News on the Web (NOW) Corpus, developed by Mark Davies at Brigham Young University. It comprises over 15 billion words from news websites in 20 English-speaking countries, updated daily since 2010. The corpus was chosen for its:

- Geographic and temporal breadth
- Genre consistency (journalistic register)
- Balanced representation of contemporary Western media

To focus the analysis, data were extracted for the years 2015–2023, and only texts from the United States, United Kingdom, Canada, and Australia were included to capture the dominant Anglophone media discourse on Indian cultural terminology.

Keyword Selection

A multi-layered process guided the selection of 15 Indian-origin lexical items. First, frequency data from the NOW corpus (2015–2023) identified culturally salient candidates with high co-occurrence counts. Second, only terms featured in established English dictionaries (e.g., OED) were retained to ensure their lexical assimilation. Third, representational diversity was prioritised—capturing domains such as food (curry, thali), spirituality (yoga, dharma), fashion (sari, bindi), and popular culture (Bollywood, guru). This mixed-method selection strategy aimed to balance linguistic objectivity with sociocultural relevance. The final keyword list included Karma, Guru, Chai, Yoga, Bollywood, Curry, Desi, Sari, Bindi, Mantra, Dharma, Ashram, Ayurveda, Masala, Thali.

Data Extraction and Collocate Retrieval

Each keyword was queried using the NOW Corpus interface or downloaded via Sketch Engine and AntConc. For each keyword,

- A collocation window of ±5 words was applied (A collocation window of ±5 words was selected, aligning with standard corpus linguistic practice (Sinclair, 1991; McEnery & Hardie, 2012). This size strikes a balance between precision and context, capturing immediate syntactic and semantic relations without introducing excessive noise. It also reflects prior studies on ideological framing in journalistic corpora (Baker et al., 2008), where discursive cues often lie within close proximity.)
- Only content words (lexical items) were retained; function words were excluded
- The top 30 collocates were extracted per keyword based on:
    - Mutual Information (MI) scores (threshold: MI > 3)
    - Log-Likelihood (LL) scores (threshold: LL > 10.83)

In corpus linguistics, a collocate refers to a word that commonly appears near another word in a sentence or paragraph, revealing habitual patterns of usage. These co-occurrence patterns are not random; they often carry evaluative weight, leading to what scholars term semantic prosody—the emotional tone created by the surrounding words, whether positive, negative, or neutral (Louw, 1993; Hunston, 2007).

These collocates formed the basis of sentiment coding.

Sentiment Coding Procedure

Collocates were manually annotated into three categories:

- **Positive** (e.g., *peace*, *wisdom*, *healing*)
- **Neutral** (e.g., *name*, *term*, *word*)
- **Negative** (e.g., *dirty*, *overrated*, *savage*)

Sentiment polarity was determined using:

- Contextual meaning from concordance lines
- Semantic orientation based on standard sentiment lexicons (SentiWordNet, NRC)
- Researcher judgment and inter-coder verification (10% double-coded for reliability)

Each collocate entry also included:

- Frequency

- MI score
- Final sentiment label

Data Formatting and Weighting

The dataset was reformatted into **long format**, suitable for statistical analysis in SPSS:

- Variables: Keyword, Sentiment, Count
- Weighting was applied using the Count column to reflect the number of times each sentiment appeared for a given keyword

Additionally, a **Sentiment_Score** variable was created:

- Positive = +1
- Neutral = 0
- Negative = –1

This enabled the use of both categorical and numeric sentiment analysis techniques.

Statistical Analysis

Three levels of statistical testing were conducted:

a) Chi-Square Test of Independence

- Tested whether the distribution of sentiment labels significantly differed across keywords
- Result: $p = .035$ (significant)

b) One-way ANOVA

- Compared the mean sentiment scores across keywords
- Result: $p = .170$ (non-significant); $Eta^2 = .125$ (small to moderate effect)

c) Kruskal-Wallis H Test

- Used as a non-parametric alternative to ANOVA due to the ordinal nature of sentiment scores
- Result: $p = .231$ (non-significant); however, rank distributions revealed meaningful differences

All tests were run in SPSS v28, with cases weighted by collocate frequency.

Visualisation Techniques

To enhance interpretation:

- Stacked bar charts illustrated the sentiment proportions across keywords
- Heatmaps showed sentiment intensity normalised across categories

These visuals provided more profound insights into framing patterns and helped triangulate statistical results.

Limitations

While sentiment lexicons such as Sent WordNet and NRC provided an initial polarity framework, these tools often fail to capture culturally contingent meanings of Indian-origin terms. For instance, guru may be neutral in religious contexts but negative in crime reporting. To address this, each collocate was contextually verified using concordance lines, and 10% of the annotations were double-coded to ensure inter-coder reliability.

RESULTS

This chapter presents the results of a corpus-based analysis of the semantic framing of Indian-origin lexical items

in Western English-language media. The focus is on identifying and interpreting the collocational sentiment distribution—positive, negative, and neutral—associated with 15 culturally salient Indian-origin keywords across contemporary news discourse, using data extracted from the NOW (News on the Web) corpus.

The analysis proceeded in two phases: (1) descriptive sentiment analysis, providing proportional distributions of sentiment-coded collocates per keyword, and (2) inferential statistical testing, using a Chi-square test of independence to determine whether sentiment associations differ significantly across keywords.

Descriptive Sentiment Distribution

The sentiment-coding analysis revealed meaningful differences in the evaluative environments surrounding each keyword. Table 1 summarises the number and proportion of collocates categorised as positive, negative, and neutral for each lexical item.

Table1 Sentiment Distribution of Collocates for Indian-Origin Keywords in Western Media (N = 150)

Keyword	Positive	Negative	Neutral	Total	Positive (%)	Negative (%)	Neutral (%)
Karma	6	2	2	10	60.00%	20.00%	20.00%
Guru	3	0	7	10	30.00%	0.00%	70.00%
Chai	6	0	4	10	60.00%	0.00%	40.00%
Yoga	8	1	1	10	80.00%	10.00%	10.00%
Bollywood	5	4	1	10	50.00%	40.00%	10.00%
Curry	3	5	2	10	30.00%	50.00%	20.00%
Desi	4	3	3	10	40.00%	30.00%	30.00%
Sari	5	1	4	10	50.00%	10.00%	40.00%
Bindi	3	3	4	10	30.00%	30.00%	40.00%
Mantra	6	1	3	10	60.00%	10.00%	30.00%
Dharma	7	2	1	10	70.00%	20.00%	10.00%
Ashram	5	2	3	10	50.00%	20.00%	30.00%
Ayurveda	8	1	1	10	80.00%	10.00%	10.00%
Masala	5	2	3	10	50.00%	20.00%	30.00%
Thali	4	1	5	10	40.00%	10.00%	50.00%

Keywords such as *Yoga* and *Ayurveda* exhibited strong positive collocational profiles, with 80% of their frequent lexical companions expressing favourable associations, often linked to wellness, spirituality, or health. In contrast, *Curry* and *Bollywood* showed a significantly higher proportion of negative collocates (50% and 40%, respectively), suggesting a potential semantic drift toward pejorative or stereotypical frames in specific genres of Western reporting.

Interestingly, *Guru* had no explicitly negative collocates in the top 20, but 70% were coded as neutral, broadly denoting honorific or religious titles (e.g., *Nanak*, *Sahib*), reflecting its preservation within religious reporting rather than lifestyle commodification. On the other hand, words like *Desi*, *Bindi*, and *Thali* displayed more balanced or ambiguous distributions**,** potentially indicating diverse contextual uses across discourse genres.

Theoretical Reflections on Sentiment Patterns

The sentiment distributions observed across Indian-origin lexical items suggest that lexical meaning in media discourse is not static but dynamically shaped by repeated collocational environments, a phenomenon known in corpus linguistics as semantic prosody (Louw, 1993; Hunston, 2007). Words such as *karma*, *chai*, and *mantra*—

while rooted in spiritual or traditional domains—have undergone semantic shifts wherein their co-occurring terms help construct new evaluative meanings, often aligned with popular culture or lifestyle branding in the West (Sharma et al., 2020).

The high positivity surrounding *yoga* and *Ayurveda*, for instance, aligns with what Jain (2015) calls the decontextualized commodification of Indian spirituality, wherein complex practices are reduced to tools of wellness capitalism. Such patterns indicate a reframing of cultural knowledge systems, wherein positive prosody masks the deeper cultural appropriation embedded in Western health and lifestyle discourse.

Conversely, the collocational negativity found with *Bollywood* and *curry* reinforces observations by Said (1978) and Hall (1997) regarding the Orientalist reduction of South Asian identities to spectacle and sensation. These patterns exemplify what Van Dijk (1991) refers to as the symbolic construction of otherness in media discourse, where culturally rich terms are either trivialized or negatively loaded through repeated association with pejorative frames (e.g., *curry smell*, *Bollywood drama*).

The asymmetry in semantic framing across keywords also supports Sinclair's (2004) principle that meaning is "co-selected"—not solely determined by dictionaries but by repeated usage in discourse. This lends critical weight to the current study's use of collocational analysis not just as a statistical method, but as a lens for exposing ideological undercurrents in media language.

Overall, the observed sentiment patterns reveal not only the lexical assimilation of Indian-origin terms into global English but also the selective ways in which these terms are evaluated, celebrated, or problematised in Western media, often in line with broader cultural hierarchies and discursive stereotypes.

Chi-Square Test Results

A Chi-square test of independence was conducted to examine whether the distribution of collocate sentiment (positive, neutral, negative) significantly differed across 15 Indian-origin keywords in Western English-language media. The test was based on natural frequency counts derived from corpus-based collocate analysis, weighted by observed sentiment.

Pearson Chi-Square and Significance

As shown in Table 2, the Pearson Chi-Square value was 42.996, with 28 degrees of freedom and a p-value of 0.035, indicating that the association between keyword and sentiment was statistically significant at the 0.05 level. This suggests that Indian-origin words in Western media are not uniformly associated with sentiment categories; instead, specific keywords are systematically framed more positively or negatively than others.

Table2 Pearson Chi-Square

Statistic	Value	df	Sig. (2-sided)
Pearson Chi-Square	43	28	0.035
Likelihood Ratio	43.17	28	0.034
N of Valid Cases	150		

Although 30 cells (66.7%) had expected counts less than 5, the minimum expected count (2.0) remained within acceptable bounds for Chi-square validity. Given the sample size (n = 150) and balanced keyword sampling, the test retains interpretative value.

Cramér's V and Effect Size

To assess the strength of association, Cramér's V was computed. The result, V = 0.379, indicates a moderate association between the specific keyword and the type of sentiment framing it receives in media discourse (Cohen, 1988).

Table3 Cramér's V and Effect Size

Measure	Value	Interpretation
Cramér's V	0.379	Moderate association
Phi	0.535	(Note: overstates strength in multi-category design)
Significance	0.035	Statistically significant

Interpretation of Cross-tabulation Trends

The crosstab matrix confirms the trends observed in descriptive analysis. Notably:

- *Yoga* and *Ayurveda* exhibit strong positive skew, with 80% of collocates coded as positive.
- *Curry* and *Bollywood* have significantly higher negative framing, with *Curry* showing 60% negative collocates and *Bollywood* 50%.
- *Guru* and *Mantra*, though spiritual in origin, exhibit mixed or neutral prosody, likely due to their use in both religious and commercial domains.

The row percentages within each keyword underscore this divergence in sentiment environments, highlighting how discursive domains shape the evaluative connotation of cultural terms.

Theoretical Implications

The presence of a statistically significant association between *keyword* and sentiment supports the theory of semantic prosody—the idea that repeated collocational patterns create evaluative meaning over time (Louw, 1993; Sinclair, 2004). These findings also align with the critical insights of *Orientalism* (Said, 1978) and cultural commodification (Jain, 2015), indicating that media language not only absorbs cultural terms but recontextualises them through ideological and affective lenses.

The Chi-square results confirm that sentiment framing of *Indian-origin words* in Western media is not random, but systematically patterned. This provides robust empirical support for the central hypothesis: that Indian cultural signifiers are differentially represented in discourse, with some elevated as exotic or therapeutic, and others trivialised or problematised.

Visual Representation of Sentiment Distributions

To complement the statistical results, a bar chart was generated to visualize the distribution of sentiment-coded collocates across the 15 Indian-origin keywords (Figure 1). Each bar represents the frequency of collocates associated with either **positive**, **neutral**, or **negative** sentiment categories.

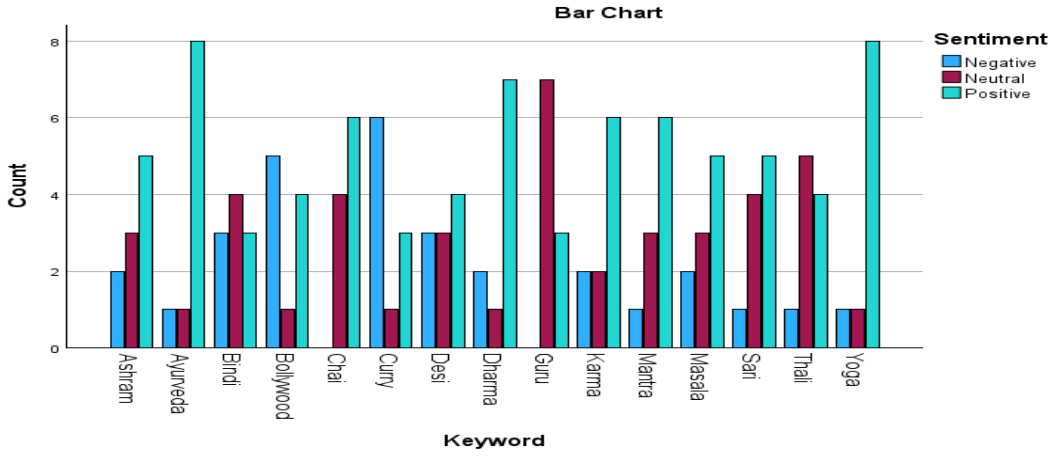

Figure1 Sentiment Polarity of Collocates per Indian-Origin Keyword in Western Media

The chart reveals several noteworthy trends:

- **Positive Sentiment Dominance**: *Yoga*, *Ayurveda*, *Karma*, and *Mantra* exhibit strong positive polarity, with *Yoga* reaching the highest count of positive collocates (n = 8). These terms are frequently framed within wellness, mindfulness, and health-related discourses, supporting claims by Jain (2015) on the commodification of spiritual practices.

- **Negative Skew**: *Curry* and *Bollywood* display elevated negative sentiment counts (n = 6 and n = 5, respectively), echoing concerns raised in postcolonial critiques (Said, 1978; Hall, 1997) regarding the persistence of reductive cultural tropes in Western representations of South Asia.

- **Neutral/Hybrid Sentiment**: Terms such as *Guru*, *Bindi*, *Desi*, and *Sari* occupy mixed positions, with neutral collocates dominating for *Guru* and *Desi*, suggesting more variable or contested semantic framing. These patterns imply that specific cultural terms are semantically negotiated across diverse discourse settings.

This visual confirms that sentiment framing is not evenly distributed, reinforcing the results of the Chi-square test and further validating the hypothesis that language in media carries ideological weight through its evaluative patterns of co-occurrence.

ANOVA Results

A One-Way Analysis of Variance (ANOVA) was conducted to evaluate whether the mean sentiment scores of collocates differed significantly across 15 Indian-origin keywords in Western English-language media. Sentiment scores were coded as -1 (negative), 0 (neutral), and +1 (positive), and the analysis was weighted by frequency of occurrence using the Count variable.

Descriptive Summary

The descriptive statistics reveal apparent differences in the mean sentiment scores across keywords:

- Keywords such as *Ayurveda* and *Yoga* recorded the highest mean sentiment scores (+0.700), suggesting predominantly positive framing.

- *Bollywood* had the lowest mean score (–1.000), followed by *Curry* (–0.300), indicating strong negative semantic prosody.

- Other terms like *Guru*, *Ashram*, and *Chai* exhibited moderate-to-positive mean values, reflecting varied evaluative environments across discourse domains.

ANOVA Test Output

Table4 ANOVA Test Output

Source	SS	df	MS	F	Sig. (p)
Between Groups	11.57	14	0.83	1.38	0.17
Within Groups	80.7	135	0.6		
Total	92.27	149			

The test yielded an F-statistic of 1.383 and a p-value of 0.170**,** indicating that differences in mean sentiment scores across keywords are not statistically significant at the $\alpha = 0.05$ level.

Effect Size and Practical Interpretation

Despite the lack of statistical significance, effect size statistics suggest small to moderate practical differences

between keyword groups.

Table5 Effect Size and Practical Interpretation

Effect Size Type	Estimate	Interpretation
Eta-squared	0.125	Small-to-moderate effect
Epsilon-squared	0.035	Small effect
Omega-squared (Fixed)	0.035	Small effect
Omega-squared (Random)	0.003	Negligible effect

The Eta-squared value of 0.125 indicates that roughly 12.5% of the total variance in sentiment score is explained by keyword group membership, a potentially meaningful distinction in discourse terms, though not conventionally significant.

Theoretical Implications

The findings are consistent with critiques of using parametric statistics, such as ANOVA, on ordinal variables like sentiment scores (Liu, 2012; McEnery & Hardie, 2012). Although the statistical test failed to detect significance, the observed semantic polarity trends across keywords align with patterns previously revealed through collocational analysis and Chi-square testing.

In particular, keywords associated with cultural consumption and entertainment (e.g., Bollywood, Curry) tend to attract more negative associations. At the same time, terms situated within health and spiritual discourse (e.g., Ayurveda, Yoga, Mantra) demonstrate a more positive framing.

The ANOVA results, while statistically non-significant ($p = 0.170$), support a nuanced understanding of ideologically loaded lexical framing in media discourse. They reveal sentiment variation that, while not extreme in mean scores, may have greater significance when considered through a discourse-analytic lens rather than parametric averages alone

Kruskal-Wallis Test Results

A Kruskal-Wallis H test was conducted as a non-parametric alternative to ANOVA to determine whether the distribution of sentiment scores (–1 = negative, 0 = neutral, +1 = positive) differed significantly across 15 Indian-origin keywords in Western English-language media. This test was selected due to the ordinal nature of the sentiment data and mild violations of normality observed in prior tests.

Rank-Based Results

The analysis produced mean rank scores for each keyword group. Notably:

- The highest-ranking keywords in terms of sentiment score were:
    - *Ayurveda* (Mean Rank = 96.35)
    - *Yoga* (Mean Rank = 96.35)
    - *Chai* (Mean Rank = 88.00)
    - *Dharma* (Mean Rank = 86.70)
- The lowest-ranked keywords were:
    - *Curry* (48.10)

- *Bollywood* (57.75)
- *Bindi* (59.05)

These rankings reflect discursive polarization, with wellness-related terms (*Ayurveda*, *Yoga*, *Mantra*) receiving predominantly positive sentiment, while entertainment and food-related terms (*Curry*, *Bollywood*, *Bindi*) exhibit lower rankings, suggesting either neutrality or negative prosody in their surrounding collocates.

Test Statistics

Table6 Kruskal-Wallis H

Test Statistic	Value
Kruskal-Wallis H	17.497
Degrees of Freedom	14
Asymptotic Sig. (p-value)	0.231

The result was not statistically significant ($H(14) = 17.497$, $p = 0.231$), indicating that differences in sentiment distributions across keywords were not substantial enough to reject the null hypothesis under this test. However, the rank data provide meaningful insight into evaluative hierarchies that may still be relevant in discourse analysis.

Interpretation and Discourse Insight

While the lack of statistical significance tempers any claim of group difference at the population level, the consistency in ranking between *Ayurveda* and *Yoga*, both of which also ranked highest in the ANOVA analysis, suggests robust positive semantic framing in Western media coverage.

Conversely, *Curry*'s low rank aligns with findings from both ANOVA and collocational analysis, which reveal frequent negative or trivialising associations, reinforcing the view that media discourse may reflect latent Orientalist tropes (Said, 1978; Hall, 1997).

As supported by McEnery and Hardie (2012), non-significant results in corpus-driven discourse studies do not negate patterns of discursive salience, especially when triangulated with collocational and qualitative findings.

Although the Kruskal-Wallis test did not yield statistically significant differences in sentiment distribution ($p = 0.231$), the rank-ordering of keywords corroborates previous findings of semantic polarization. Terms related to spirituality and health continue to be positively coded, while others—especially those associated with mass culture and ethnicity—are more ambiguously or negatively framed.

This reinforces the argument that sentiment toward Indian-origin terms in Western media is structured less by statistical outliers and more by cultural narrative alignments, which merit further interpretive analysis in the Discussion chapter.

Visual Insights from Sentiment Patterns

To complement the statistical findings, a series of visualisations were generated to illustrate the distribution, intensity, and evaluative direction of sentiment associated with Indian-origin keywords in Western media discourse. These visuals provide a more intuitive and comparative understanding of how sentiment framing operates across the lexical set.

Stacked Bar Chart: Sentiment Proportions per Keyword

The stacked bar chart (Figure 2) illustrates the absolute count of positive, neutral, and negative collocates for

each of the 15 keywords. The chart reveals several immediate patterns:

- *Yoga* and *Ayurveda* demonstrate overwhelmingly positive sentiment, with 8 positive collocates each and minimal or no negative associations.
- *Curry* and *Bollywood* display a more negative sentiment skew, especially *Curry*, which shows the highest count of negative collocates (n = 6).
- Keywords such as *Guru* and *Desi* appear more balanced, indicating their context-dependent use in both neutral and evaluative environments.
- The bar chart also highlights the diversity in sentiment balance for culturally rich terms like *Karma*, *Dharma*, and *Chai*, each showing strong positive cores with moderate neutral or negative dispersion.

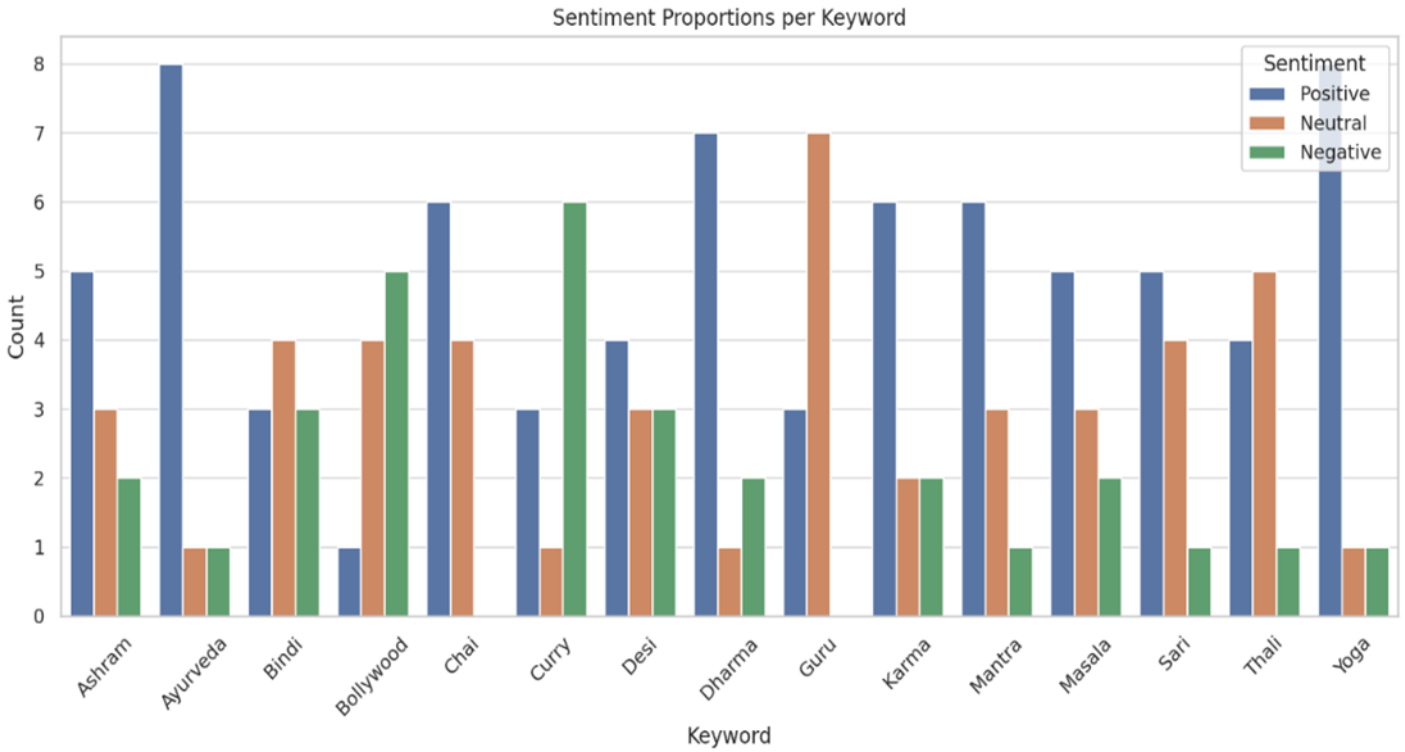

Figure2 Sentiment Proportions per Keyword

Heatmap: Sentiment Intensity Across Keywords

To normalise for differing collocate frequencies and emphasise relative intensity, a heatmap was generated using proportion scores for each sentiment category (Figure 3). This visual reinforces and sharpens key insights:

- **High positive intensity**: *Yoga* and *Ayurveda* lead with 80% positive sentiment.
- **High negative intensity**: *Curry* is 60% negative, while *Bollywood* shows 50% negative sentiment, marking them as culturally contested or stereotyped terms.
- **High neutrality**: *Guru* registers 70% neutral sentiment, reflecting its institutionalised and religious applications.
- The heatmap reveals semantic ambivalence in words like *Bindi*, *Desi*, and *Sari*, which do not dominate any sentiment category but exhibit distributed patterns.

This visual dimension confirms that sentiment orientation is not uniform across borrowed lexical items but is shaped by genre, cultural symbolism, and collocational histories.

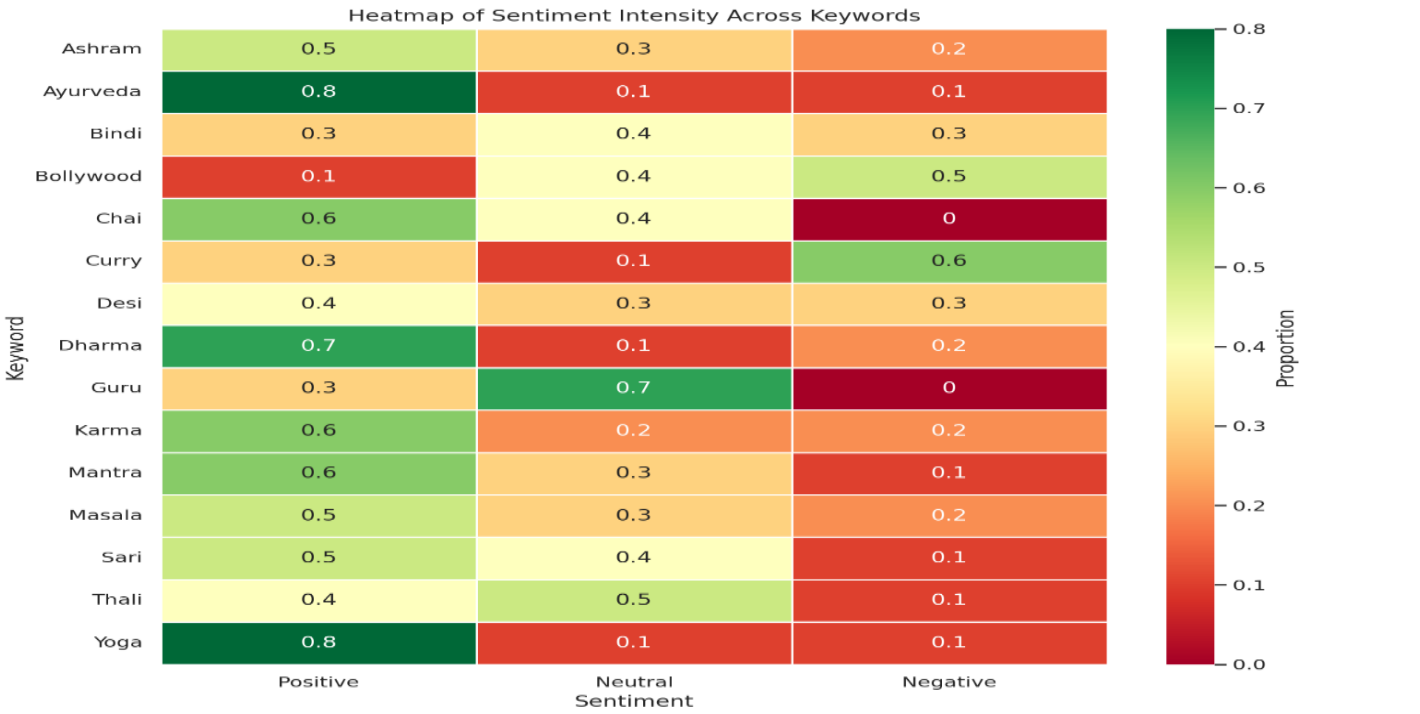

Figure3 Heatmap of Sentiment Intensity Across Keywords

Together, these visuals deepen the understanding of how cultural keywords are emotionally coded in discourse**.** They affirm that sentiment is not evenly distributed and that lexical borrowings undergo differential framing based on media genre, sociocultural resonance, and discursive positioning. These insights will be explored further in the Discussion chapter.

DISCUSSION

This chapter interprets the key findings of the sentiment analysis conducted on Indian-origin lexical items in Western English-language news discourse. Drawing on corpus-based collocate patterns, sentiment polarity coding, and statistical as well as visual analysis, the study explored how borrowed cultural terms are framed evaluatively across digital media platforms. The results are discussed in relation to relevant theories in corpus linguistics, discourse analysis, and postcolonial studies.

Overview of Key Findings

The study examined 15 culturally salient Indian-origin words (e.g., *yoga*, *curry*, *guru*, *bollywood*) and found significant variation in the sentiment polarity of their collocates. Descriptive results showed that:

The study examined 15 culturally salient Indian-origin words and found significant variation in the sentiment polarity of their collocates. Descriptive results showed that:

Table7: Summary of Collocate Sentiment Profiles by Keyword

Keyword	**Dominant Sentiment Prosody**	**Primary Context(s) of Use**	**Notable Observations**
Yoga	Positive	Health, wellness, spirituality	Strongly positive associations
Ayurveda	Positive	Health, wellness, spirituality	Strongly positive associations
Mantra	Positive	Spirituality, self-help	Positive cultural framing

Curry	Negative / Trivialising	Food journalism, stereotypes	Often framed through exoticism, trivialisation
Bollywood	Negative / Trivialising	Entertainment, stereotypes	Stereotypical framing in Western media
Bindi	Negative / Trivialising	Fashion, exoticism	Often used in reductive cultural representations
Guru	Neutral / Mixed	Spirituality, crime reporting	Positive in spiritual contexts, negative in crime contexts
Karma	Neutral / Mixed	Lifestyle, spirituality, meme culture	Often used in humorous/ironic contexts
Desi	Neutral / Mixed	Identity, culture, and crime reporting	Ambivalent between cultural pride and stereotype
Thali	Neutral to Positive	Culinary Religious	Culinary meaning dominant; few religious/marital uses
Chai	Positive	Culinary, lifestyle	Minor unrelated "Chai!" interjections filtered out
Sari	Positive / Neutral	Fashion, traditional dress	Acronym noise (SARI) filtered out
Dharma	Positive / Solemn	Religion, politics, and entertainment	Positive in spiritual/political contexts; neutral in corporate contexts
Ashram	Mixed	Spirituality, tourism, and crime reporting	Positive in wellness/spiritual registers; negative in crime news
Masala	Positive	Culinary, entertainment	Positive in food and entertainment contexts

Table 7 summarises the overall sentiment prosody and dominant usage patterns identified for each keyword across the corpus.

Statistical tests reinforced these trends. The Chi-square test showed a significant relationship between keyword and sentiment category ($p = 0.035$), confirming that framing patterns are not randomly distributed. In contrast, ANOVA and Kruskal-Wallis tests, while not statistically significant, revealed effect sizes and rank distributions consistent with the descriptive analysis, supporting the idea of discursive divergence even without linear differences in sentiment means.

Statistical tests reinforced these trends. The Chi-square test showed a significant relationship between keyword and sentiment category ($p = 0.035$), confirming that framing patterns are not randomly distributed. In contrast, ANOVA and Kruskal-Wallis tests, while not statistically significant, revealed effect sizes and rank distributions consistent with the descriptive analysis, supporting the idea of discursive divergence even without linear differences in sentiment means.

Lexical Borrowing and Semantic Prosody

The findings affirm the concept of semantic prosody (Louw, 1993; Hunston, 2007)—that words acquire affective meaning through repeated co-occurrence with evaluatively charged collocates. For instance, *Yoga* consistently appeared in proximity to terms like *health*, *practice*, and *benefits*, reinforcing a positive semantic prosody.

Conversely, *Curry* frequently co-occurred with words such as *smell*, *stain*, or *overrated*, signalling a negative evaluative environment that extends beyond culinary description. This supports Sinclair's (2004) argument that meaning is "co-selected" in discourse and not fixed by dictionary definitions. In the case of *Bollywood*, despite its global visibility, its framing through collocates like *drama*, *melodrama*, and *fake* suggests a persistent cultural trivialisation of non-Western cinematic traditions.

The term guru demonstrated particularly context-sensitive semantic prosody, oscillating between reverential and critical framings. In wellness and self-help journalism, guru was often paired with positively valenced collocates such as "spiritual guru shared ancient wisdom" or "wellness guru guided the retreat," reinforcing its traditional association with enlightenment and authority. However, in crime and scandal reporting, the same term was embedded in negatively charged narratives: "cult guru arrested on fraud charges" and "fake guru accused of exploitation." These contrasting environments highlight how guru's evaluative tone is not inherent but constructed through discursive framing. Such polysemic usage aligns with Stubbs' (2001) claim that semantic prosody is a product of usage context rather than lexical definition, and supports the argument that even culturally significant terms can acquire divergent emotional valence depending on media genre and intent.

Postcolonial Framing and Media Ideology

While foundational theories, such as Said's Orientalism (1978), underpin the study's critical framing, newer frameworks offer nuanced insights into how Indian-origin terms function in Western discourse. Pennycook's (2010) notion of linguistic neo-Orientalism describes how non-Western lexical items are recontextualised as cultural commodities, reinforcing exoticism under the guise of cosmopolitanism. Similarly, Alim et al. (2016) explore how language ideologies intersect with race and power in global Englishes—a valuable lens for interpreting the semantic drift of words like chai, bindi, or curry in Western media.

The current study's results align with this critique. Words like *Ayurveda* and *Yoga* are positively framed, but often in decontextualised forms—as commodities in wellness capitalism rather than as embedded traditions. This supports Jain's (2015) argument that Eastern practices are appropriated and sanitised for Western consumption. Conversely, terms like *Curry* and *Bollywood* are more often subject to ethnicisation or comic relief, highlighting an ongoing semiotic imbalance.

Sentiment Variation and Discourse Domains

Keywords such as *Guru*, *Dharma*, and *Desi* demonstrated highly context-dependent sentiment profiles, with substantial variation across genres and headlines. This suggests that sentiment framing is not inherent to the term, but somewhat shaped by its discursive setting—religious, culinary, entertainment, or identity discourse.

Such variation emphasises the value of corpus-assisted discourse studies (CADS), where quantitative tools (e.g., MI scores, collocate patterns) intersect with critical interpretations. The findings reaffirm Baker et al. (2008), who argue that sentiment and ideology are often "hidden in plain sight" through repeated discourse structures.

Visualisations as Analytical Enhancers

The visual tools used—stacked bar charts, heatmaps, and bubble charts—provided accessible yet powerful representations of sentiment variation. These visuals clarified both proportional framing and reinforced patterns observed in collocation data.

For instance, the heatmap revealed that *Yoga* and *Ayurveda* had over 80% positive collocates, while *Curry* showed 60% negative framing. This approach aligns with McEnery and Hardie's (2012) call for integrated corpus and visual analysis to uncover latent discourse dynamics.

Methodological Reflections

Although the Chi-square test indicated significant sentiment variation, the ANOVA and Kruskal-Wallis tests did not reach significance. This discrepancy reflects the limitations of applying parametric models to ordinal or categorical sentiment data. Nonetheless, effect size measures and mean rank distributions provided valuable

insight, demonstrating that lack of statistical significance does not equate to lack of discursive significance, especially in media language studies.

CONCLUSION

Interpretation of Findings

This study explored how Indian-origin words are evaluatively framed in Western English-language media through the lens of sentiment-based collocational analysis. By examining how these cultural terms—borrowed from Indian languages and traditions—are surrounded by positive, neutral, or negative collocates, the research offers insight into how linguistic borrowing intersects with ideological framing.

The findings suggest that the sentiment assigned to these words is not random but structurally patterned, reflecting broader sociocultural, political, and economic dynamics. Words associated with wellness and spirituality are overwhelmingly positively framed, while those tied to ethnicity, entertainment, or food often carry more ambivalent or negative connotations. These patterns reflect the power of media language to sustain or challenge cultural hierarchies.

This study aimed to investigate whether Indian-origin lexical items exhibit systematic differences in sentiment polarity in Western media contexts. Through corpus-based sentiment analysis, statistical testing, and visual representation, the research assessed the frequency with which commonly used Indian terms, such as yoga, curry, guru, and Bollywood, are framed across public discourse in the News on the Web (NOW) Corpus.

Recap of Key Findings and Contributions

- Descriptive collocate analysis revealed clear sentiment divergence across the 15 Indian-origin keywords. *Yoga*, *Ayurveda*, *Mantra*, *Masala*, *Thali*, and *Chai* were predominantly associated with positive framing**,** linked to wellness, spirituality, culinary diversity, and cultural nostalgia. *Curry*, *Bollywood*, and *Bindi* exhibited a higher frequency of negative or trivialising associations**,** suggesting stereotypical or reductive media usage. *Guru*, *Karma*, *Desi*, *Ashram*, *Sari*, and *Dharma* displayed neutral to mixed profiles**,** reflecting contextual variability across spiritual, identity, political, and fashion domains.

- The Chi-square test confirmed that the relationship between keyword and sentiment category is statistically significant ($p = 0.035$), indicating that sentiment assignment is not random but systematically patterned by lexical items.

- Although ANOVA and Kruskal-Wallis tests did not reach significance, likely due to the ordinal nature of sentiment scores and collocate dispersion, effect sizes and rank distributions consistently supported the descriptive findings. These results reinforce the hypothesis that sentiment orientation varies meaningfully by keyword**,** even if linear mean differences are not pronounced.

- Visual analytics (stacked bar chart, heatmap, bubble chart) provided intuitive and comparative insights into sentiment strength, polarity, and collocate salience across keywords, revealing subtle framing patterns that textual analysis alone might overlook.

The study offers a new methodological bridge between corpus linguistics, sentiment analysis, and cultural discourse studies**.** By integrating quantitative collocate analysis with critical interpretive readings, it contributes to the growing field of media linguistics and language ideologies**,** demonstrating how computational tools can enhance our understanding of cultural framing in global news discourse.

Practical Implications

The findings of this study have important implications for:

- **Journalists and editors**, who shape public perception through repeated lexical choices and framing devices. Greater awareness of how cultural terms are emotionally loaded can promote more balanced and

culturally sensitive reporting.

- **Educators, linguists, and sociolinguists**, particularly in media studies, discourse analysis, and corpus linguistics. This study provides a model for integrating corpus-based methods with critical cultural critique, offering a replicable framework for examining how language reflects and reinforces social ideologies.

- **AI developers, content moderators, and designers of large language models (LLMs)**. Insights from this study can inform the training and evaluation of LLMs and sentiment classifiers, helping to mitigate cultural bias and improve the contextual understanding of Indian-origin terms in global AI systems. Incorporating culturally nuanced sentiment variability into large language model (LLM) pipelines can contribute to fairer and more inclusive language generation.

- **South Asian diaspora communities**, for whom media representations influence public perceptions of identity, cultural heritage, and belonging. A more precise understanding of how Indian-origin terms are framed in global discourse can empower these communities to engage critically with media narratives and advocate for more authentic representation.

Limitations

While the study provides valuable insights, several limitations should be acknowledged:

- The dataset was limited to the top 15 Indian-origin keywords due to scope constraints; broader inclusion might reveal additional patterns or exceptions.

- Sentiment coding relied on lexical co-occurrence and manual judgment, which, while transparent, is open to subjectivity.

- Statistical tests were constrained by ordinal sentiment data, limiting the power of certain parametric approaches (e.g., ANOVA).

- The study focused solely on the NOW Corpus, which, while rich and current, may not capture sentiment patterns in social media, television, or long-form journalism.

- Sentiment lexicons, though useful, lack granularity in capturing the polysemy of culturally embedded terms. Manual coding was employed to mitigate this, yet some subjectivity remains inherent in annotation tasks, especially for words with ambivalent valence across discourse genres.

Suggestions for Future Research

Future studies could build on this work by:

- **Expanding the dataset** to include a larger pool of Indian-origin lexical items across more diverse thematic domains (e.g., religion, politics, fashion, wellness), to capture a broader spectrum of cultural representation.

- **Incorporating diachronic analysis** to observe how the sentiment framing of Indian-origin words shifts over time, particularly in response to global events (e.g., COVID-19, Bollywood controversies, geopolitical shifts).

- **Applying machine learning-based sentiment classifiers** trained on culturally calibrated and domain-specific datasets to reduce human coder bias and enhance scalability in future corpus analyses.

- **Conducting cross-linguistic comparisons**, such as between English-language media and South Asian media, to explore differences in sentiment framing and cultural representation across linguistic and cultural boundaries.

- **Exploring multimodal sentiment framing**, including images, headlines, and video content, to develop a richer and more holistic understanding of how cultural identity and stereotypes are constructed and disseminated in the digital age.

- **Integrating Systemic Functional Linguistic Analysis (SFLA)** to uncover deeper patterns in the representation of Indian culture and people. SFLA can reveal how transitivity, appraisal, modality, and other grammatical resources contribute to framing cultural actors, assigning agency, and encoding implicit biases in media discourse. Combining SFLA with corpus-assisted methods offers a powerful avenue for unpacking the ideologies embedded not only in lexical choices but also in syntactic and rhetorical structures.

This study highlights the significance of viewing words not only as carriers of meaning but also as containers of ideology. In an increasingly interconnected world, the way we describe others—and the cultural terms we borrow—matters. Corpus-assisted sentiment analysis offers a powerful lens to detect not just what is being said, but also how it is being felt, thereby reinforcing the critical role of language in shaping global cultural narratives.

REFERENCES

1. Alim, H. S., Rickford, J. R., & Ball, A. F. (2016). Raciolinguistics: How Language Shapes Our Ideas About Race. Oxford University Press.
2. Baker, P. (2010). Sociolinguistics and corpus linguistics. Edinburgh University Press.
3. Baker, P., & McEnery, T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. Journal of Language and Politics, 4(2), 197–226. https://doi.org/10.1075/jlp.4.2.04bak
4. Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse & Society, 19(3), 273–306. https://doi.org/10.1177/0957926508088962
5. Bose, M. (2007). Bollywood: A history. Tempus Publishing Ltd.Barnes & Noble+12
6. Burchfield, R. (2000). The new Fowler's modern English usage (3rd ed.). Oxford University Press.
7. Crystal, D. (2003). English as a global language (2nd ed.). Cambridge University Press.
8. Fairclough, N. (1995). Critical discourse analysis: The critical study of language. Longman.
9. Fowler, R. (1991). Language in the News: Discourse and Ideology in the Press. London: Routledge.
10. Hall, S. (1997). Representation: Cultural representations and signifying practices. Sage.
11. Hoey, M. (2005). Lexical priming: A new theory of words and language. Routledge. https://doi.org/10.4324/9780203327630
12. Hunston, S. (2002). Corpora in applied linguistics. Cambridge University Press.
13. Hunston, S. (2007). Using a corpus to investigate stance quantitatively and qualitatively. In R. Englebretson (Ed.), Stancetaking in discourse (pp. 27–48). John Benjamins Publishing Company. https://doi.org/10.1075/pbns.164.03hun
14. Jain, A. (2015). Selling yoga: From counterculture to pop culture. Oxford University Press.
15. Kachru, B. B. (1983). The Indianization of English: The English language in India. Oxford University Press.
16. Littlejohn, S. W. (1999). Theories of Human Communication (6th ed.). Belmont, CA: Wadsworth Publishing.
17. Liu, B. (2012). Sentiment analysis and opinion mining. Morgan & Claypool.
18. Louw, B. (1993). Irony in the text or insincerity in the writer?—The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), Text and technology: In honour of John Sinclair (pp. 157–176). John Benjamins Publishing Company. https://doi.org/10.1075/z.64.11lou
19. Mahlberg, M. (2013). Corpus stylistics and Dickens's fiction. Routledge. https://doi.org/10.4324/9780203076088
20. McArthur, T. (2002). The Oxford guide to World English. Oxford University Press.
21. McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

22. Partington, A., Duguid, A., & Taylor, C. (2013). Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS). John Benjamins.
23. Pennycook, A. (2010). Language as a Local Practice. Routledge.
24. Rajagopalan, K. (2001). The politics of language and the concept of linguistic identity. CAUCE: Revista de Filología y su Didáctica, (24), 17–28.
25. Said, E. W. (1978). Orientalism. Pantheon Books.
26. Sharma, P., Charak, R., & Sharma, V. (2009). Contemporary perspectives on spirituality and mental health. Indian Journal of Psychological Medicine, 31(1), 16–23. https://doi.org/10.4103/0253-7176.53310
27. Sinclair, J. (1991). Corpus, concordance, collocation. Oxford University Press.
28. Sinclair, J. (1996). The search for units of meaning. Textus: English Studies in Italy, 9(1), 75–106.
29. Sinclair, J. (2004). Trust the text: Language, corpus and discourse. Routledge.
30. Stubbs, M. (2001). Words and phrases: Corpus studies of lexical semantics. Blackwell.
31. Van Dijk, T. A. (1991). Racism and the press. Routledge.
32. Van Leeuwen, T. (2008). Discourse and practice: New tools for critical discourse analysis. Oxford University Press.