Sentiment Analysis of Code-Mixed Punjabi–English Social Media Text: A Survey of Approaches, Datasets, and Open Challenges

Rajvinder Kaur
Department of Computer Science and Applications,
Khalsa College, Amritsar, Punjab

ABSTRACT
Sentiment analysis of user-generated text has matured rapidly for high-resource languages, but Punjabi—despite a speaker population exceeding one hundred million—remains poorly served by annotated corpora, lexicons, and evaluation benchmarks. The problem is compounded on social media, where users routinely produce Punjabi–English code-mixed text written variously in Gurmukhi, Shahmukhi, and Romanized Punjabi, with unstandardized spelling, transliteration variants, slang, and informal grammar. This paper surveys research on sentiment analysis of Punjabi and Punjabi–English code-mixed text. Following a documented search protocol, this survey identifies and critically examines the primary studies published between 2014 and 2025, covering lexicon- and rule-based methods, classical machine learning classifiers, deep neural architectures, and multilingual transformer models. This survey catalogues the datasets used in this line of work, including their domains, scripts, sizes, and availability; summarizes the preprocessing pipelines, feature representations, and evaluation metrics reported; and tabulates the published results of each study to enable direct comparison. The survey shows that the genuinely code-mixed Punjabi–English literature is still small and fragmented: datasets are largely private and domain-specific, no shared benchmark exists, and transformer-based methods have so far been demonstrated mainly on the related Urdu–Punjabi (Shahmukhi) setting rather than on Romanized Punjabi–English text. This survey closes by identifying concrete research needs—public benchmark corpora, script normalization and transliteration resources, and Punjabi-adapted pretrained models—that would allow this area to progress beyond isolated proof-of-concept studies.

Keywords
Sentiment analysis, code-mixed text, Punjabi–English, low-resource languages, natural language processing, transformer models, social media

1. INTRODUCTION
Sentiment analysis, or opinion mining, seeks to identify and classify the opinions, emotions, and attitudes expressed in text. The growth of social media platforms, review sites, and online forums has made opinion-rich text abundant, and sentiment analysis now supports applications ranging from customer-feedback monitoring and recommendation systems to political and public-health analysis. The methodological trajectory of the field—from lexicons and handcrafted rules through classical machine learning to deep learning and pretrained transformers—has been documented in recent comprehensive reviews [1]. Most of this progress, however, has accrued to high-resource languages, English above all.

Punjabi occupies an awkward position in this landscape. It is among the most widely spoken languages in the world, with well over one hundred million speakers concentrated in Indian Punjab and Pakistani Punjab [2], yet it is a low-resource language in computational terms: annotated corpora, sentiment lexicons, pretrained models, and standard preprocessing tools (tokenizers, stemmers, part-of-speech taggers) are scarce. Techniques developed for resource-rich languages therefore cannot be transferred without substantial adaptation, and the absence of shared benchmarks makes it hard even to compare the adaptations that do exist.

The difficulty deepens on social media. India is a multilingual society—the Census of India 2011 records that a substantial share of the population reports proficiency in more than one language [3]—and bilingual users routinely produce code-mixed text, combining two or more languages within a single utterance. A sentence such as “Tuhade kol bahut vadhiya dress hai” follows Punjabi grammar while borrowing the English noun dress; “Oh yaar sarkaara da ki ae, they keep on changing” switches grammatical frame mid-sentence. Punjabi–English code-mixing of this kind is common on Facebook, X (formerly Twitter), Instagram, WhatsApp, and YouTube. It is further complicated by script diversity: Punjabi is written in Gurmukhi in India and Shahmukhi in Pakistan, while informal digital communication overwhelmingly uses unstandardized Romanized Punjabi, so the same word may surface in several spellings and scripts within one thread.

Computational approaches to sentiment analysis have evolved through several generations. Early systems used sentiment lexicons and handcrafted rules; classical supervised learners—Naïve Bayes, Support Vector Machines (SVM), logistic regression, decision trees, and random forests—followed, trained on labeled data with statistical features. Deep architectures such as Long Short-Term Memory networks [4] and convolutional neural networks for sentence classification [5] then removed much of the feature engineering, and the current generation of multilingual pretrained transformers—BERT and its multilingual variant mBERT [6], XLM-RoBERTa [7], IndicBERT [8], and MuRIL [9]—provides contextual representations that transfer across languages. Each generation has been applied, unevenly, to Punjabi.

Code-mixed sentiment analysis for Indian languages has been studied most intensively for Hindi–English, beginning with sociolinguistic analyses of code-mixing behavior in Facebook [10] and extending to sentiment mining of mixed-script Indic sentences [11] and annotated corpora for related tasks such as sarcasm detection in Hindi–English tweets [12]. Shared tasks and benchmarks—most prominently SemEval-2020 Task 9 on sentiment analysis of code-mixed Hindi–English and Spanish–English tweets [13] and the GLUECoS evaluation suite for code-switched NLP [14]—have given Hindi–English research a common footing that Punjabi–English research still lacks.

Work on Punjabi itself falls into three strands. Monolingual

Punjabi sentiment analysis began with lexicon- and scoring-based polarity detection for political reviews [15], continued with SVM-based classification over a Punjabi subjective lexicon [16], and reached deep learning with a morphology-aware neural framework evaluated on Punjabi news documents [17]. A second, smaller strand addresses Punjabi–English code-mixed text directly: a language-identification and normalization pipeline with SVM and Naïve Bayes classifiers for agriculture-domain tweets [18], machine-learning classification of bilingual social media posts [19], and, recently, a logistic-regression emotion classifier trained on a substantially larger corpus of English–Punjabi posts [20]. A third strand comprises reviews of Punjabi sentiment analysis in general [21] and applications of language-specific toolkits such as the Indic NLP library [22], together with closely related work on Urdu–Punjabi code-switched text in Shahmukhi script, where a transformer framework combining XLM-RoBERTa and GPT-2 was trained on the million-review LUPCSA-25 corpus [23], and on foundational Shahmukhi resources such as word embeddings and classification datasets [24].

No existing survey consolidates this literature with a specific focus on the Punjabi–English code-mixed setting. The closest prior review [21] covers Punjabi sentiment analysis broadly, alongside other vernacular languages, and does not address code-mixing, script diversity, or the Romanized text that dominates social media. The present survey fills that gap. The state of the field it describes calls for candor: the strictly Punjabi–English code-mixed literature currently amounts to a handful of primary studies. Precisely because the area is small, a careful consolidation—one that reports what each study actually did and actually achieved, catalogues the datasets with their limitations, and connects the area to transferable resources from Hindi–English and Shahmukhi research—can serve as a reliable starting point for the work the field still needs.

The contributions of this survey are as follows:

- A systematically collected and critically analyzed review of sentiment analysis research on Punjabi and Punjabi–English code-mixed text published between 2014 and 2025, with each study characterized by its verified methods and reported results.
- A consolidated catalogue of the datasets used in this literature, documenting language pair, script, domain, size, and public availability, together with related benchmarks from Hindi–English and Urdu–Punjabi research.
- A comparative analysis of lexicon-based, classical machine learning, deep learning, and transformer-based approaches as they have actually been applied to Punjabi, identifying what has and has not been demonstrated.
- An assessment of open problems—benchmark scarcity, script and transliteration variability, and the absence of Punjabi-adapted pretrained models—with concrete directions for addressing them.

The remainder of the paper is organized as follows. Section 2 describes the survey methodology. Section 3 provides background on the Punjabi language, code-mixing, and the specific challenges they pose. Section 4 reviews the literature. Section 5 examines datasets, preprocessing, feature representation, and evaluation practice. Section 6 presents a comparative analysis, Section 7 discusses challenges and future directions, and Section 8 concludes.

2. SURVEY METHODOLOGY

Studies were identified through keyword searches of IEEE Xplore, the ACM Digital Library, ScienceDirect, SpringerLink, Google Scholar, and arXiv, conducted between 2024 and 2025. Search strings combined the terms “Punjabi” and “sentiment analysis” with the modifiers “code-mixed,” “code-switched,” “bilingual,” “English–Punjabi,” “Gurmukhi,” “Shahmukhi,” and “Romanized.” Reference lists of retrieved papers were traversed backward to locate earlier work, and citing articles were traversed forward to locate more recent work.

A publication was included as a primary study if it (a) performed sentiment or emotion classification on Punjabi or Punjabi–English text, or introduced a dataset or review dedicated to that task, and (b) appeared in a peer-reviewed venue or as a publicly archived preprint between 2014 and 2025. Closely related work on Hindi–English code-mixing and Urdu–Punjabi (Shahmukhi) sentiment analysis was included as context where it supplies transferable methods, benchmarks, or resources, and is identified as such throughout. Publications were excluded if they addressed Punjabi NLP tasks other than sentiment or emotion analysis, or if their content could not be verified against the published record. All bibliographic details and reported results cited in this survey were checked against the original sources.

3. BACKGROUND

3.1 The Punjabi Language and its Scripts

Punjabi is an Indo-Aryan language with a speaker population, across all dialects, exceeding one hundred million, the majority in Pakistani Punjab and a large minority in Indian Punjab and the global diaspora [2]. It is written in two standard scripts—Gurmukhi in India and Shahmukhi (a Perso-Arabic script) in Pakistan—while informal digital communication predominantly uses ad hoc Romanization. No standard Romanization scheme exists, so a single word may be spelled several ways depending on the writer’s pronunciation and typing habits. Despite its size, Punjabi is a low-resource language for NLP: annotated corpora, sentiment lexicons, mature tokenizers, stemmers, lemmatizers, and part-of-speech taggers are limited in number and coverage, a gap documented repeatedly in the studies reviewed below [16], [17], [21].

3.2 Code-Mixing in Punjabi–English Communication

Code-mixing combines words, phrases, or clauses from two or more languages within a single sentence or conversation. In Punjabi–English social media text, English content words are commonly embedded in Punjabi grammatical frames, and switches can also occur at clause level, as the examples in Section 1 illustrate. English items are often preferred where they are more current than their Punjabi equivalents, while Punjabi supplies the morphosyntax. This behavior is natural for bilingual speakers but difficult for automatic systems, which must identify the language of each token before applying language-appropriate processing.

3.3 Why Punjabi–English Sentiment Analysis Is Hard

Five factors distinguish this setting from monolingual sentiment analysis. First, script diversity: the same content circulates in Gurmukhi, Shahmukhi, and Romanized form, and models trained on one script do not transfer to the others. Second, transliteration variability: unstandardized Romanization multiplies spellings of the same word, inflating

vocabulary and fragmenting the evidence available to any statistical model. Third, token-level language identification: distinguishing Punjabi from English tokens is itself nontrivial, since short Romanized strings can be valid words in either language with different meanings. Fourth, informal register: abbreviations, elongated characters, emojis, hashtags, slang, and dialectal forms carry sentiment signal but defeat conventional tokenization and normalization. Fifth, resource scarcity: no large public annotated corpus, no standard sentiment lexicon, and no shared benchmark currently exist for Punjabi–English code-mixed sentiment analysis. Each of these factors appears concretely in the studies reviewed in Section 4.

4. LITERATURE REVIEW

Figure 1 summarizes the families of computational approaches that have been applied to Punjabi and Punjabi–English sentiment analysis. The subsections that follow review the literature in four groups: Hindi–English code-mixed research and shared benchmarks, which established the methods and evaluation practice of Indian code-mixed NLP; monolingual Punjabi sentiment analysis; Punjabi–English code-mixed studies, the core of this survey; and closely related Urdu–Punjabi (Shahmukhi) work.

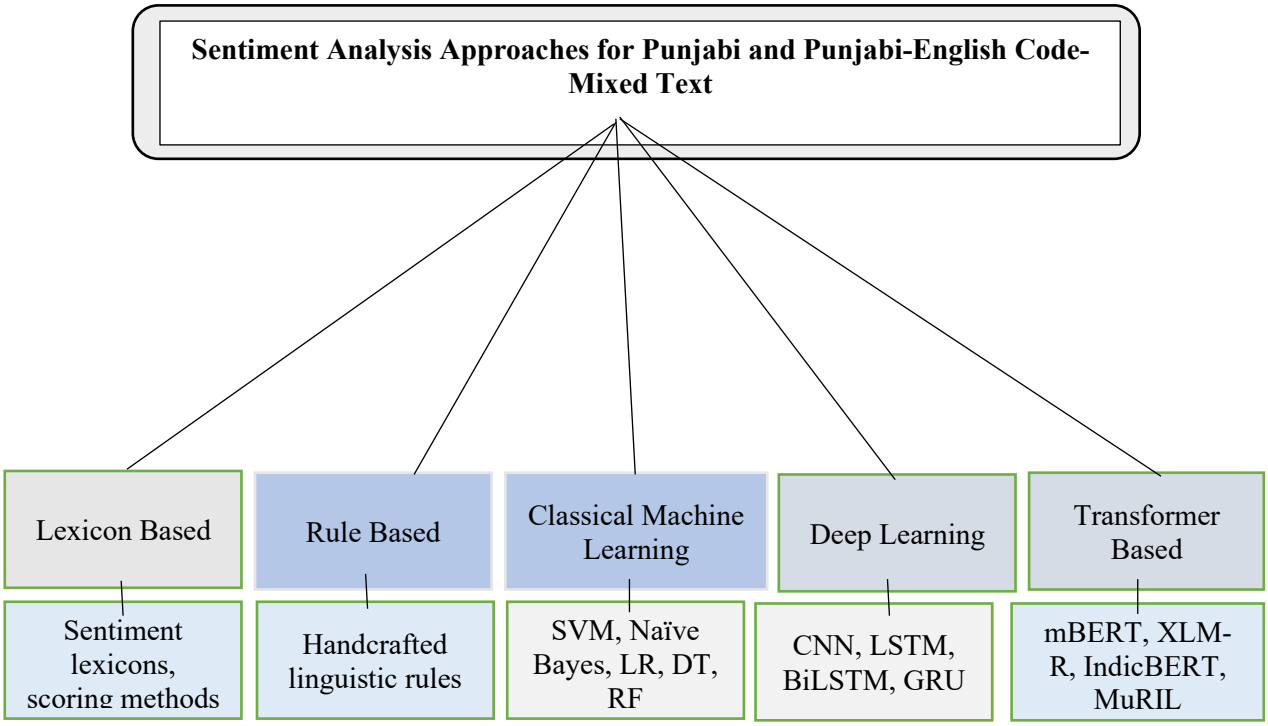

Figure 1. Taxonomy of computational approaches applied to Punjabi and Punjabi–English code-mixed sentiment analysis.

4.1 Hindi–English Code-Mixed Research and Shared Benchmarks

Research on Indian-language code-mixing began with descriptive and sociolinguistic analysis. Bali et al. [10] analyzed English–Hindi code-mixing in Facebook posts, characterizing the structural patterns of mixing and the ambiguity between borrowing and mixing; although the study did not perform sentiment classification, it established the linguistic ground on which later computational work was built. Bhargava et al. [11] contributed an annotated English–Hindi code-mixed tweet corpus for sarcasm detection which, while built for a different task, demonstrated the corpus-construction methodology that Punjabi–English research has yet to replicate at scale.

The consolidation of Hindi–English research came through shared evaluation. SemEval-2020 Task 9 (SentiMix) released a sentiment-annotated Hindi–English tweet corpus and drew dozens of participating systems spanning classical, neural, and transformer methods [13], and the GLUECoS benchmark assembled a suite of code-switched tasks, including sentiment analysis, for standardized comparison [14]. Nothing comparable exists for Punjabi–English; the contrast between the two literatures is instructive, because the availability of a shared benchmark, more than any modeling advance, is what allowed Hindi–English work to accumulate.

4.2 Monolingual Punjabi Sentiment Analysis

Arora and Kaur [15] provided an early demonstration of Punjabi sentiment analysis on political reviews, determining the polarity of Punjabi text through a lexicon-driven scoring method and producing separate positive and negative summaries for decision support. The study's contribution was feasibility rather than benchmarked accuracy: it showed that polarity detection in Gurmukhi Punjabi was practical despite the absence of language resources, using a scoring approach rather than supervised machine learning.

Kaur and Gupta [16] moved to supervised classification, constructing a Punjabi subjective lexicon by combining a Punjabi dictionary with the Hindi subjective lexicon and training an SVM classifier over it, with a reported accuracy of 78.02%. The study established the pattern—resource construction by cross-lingual transfer from Hindi, followed by classical supervised learning—that characterizes much subsequent Punjabi work.

Singh et al. [17] brought deep learning to Punjabi sentiment analysis with a framework that couples morphological evaluation and normalization of Punjabi words with a deep

neural network classifier. The corpus consisted of 275 Punjabi text documents concerning farmer-suicide cases reported in Punjab, and the model classified text into four negatively oriented classes tailored to that domain, achieving per-class accuracies of 93.85%, 88.53%, 83.30%, and 95.45% under 10-fold cross-validation and an overall accuracy of 90.29%. Two qualifications matter for interpreting this result. The task is domain-specific fine-grained negative-sentiment classification rather than generic positive/negative/neutral polarity, and the corpus is small and, to the authors' knowledge, not publicly released. Within those limits, the study is the strongest evidence to date that morphology-aware neural modeling benefits Gurmukhi Punjabi.

4.3 Punjabi–English Code-Mixed Sentiment Analysis

Singh et al. [18] carried out the earliest identified study specifically devoted to Punjabi–English code-mixed sentiment analysis, targeting agriculture-domain social media content. Their pipeline performs token-level language identification, normalizes the noisy Romanized text, and constructs an English–Punjabi bilingual dictionary before applying SVM and Naïve Bayes classifiers. The work demonstrated that preprocessing and normalization are not optional refinements but preconditions for handling code-mixed input, given the spelling and transliteration variation of social media writing. Its limitations are the domain restriction to agriculture and the reliance on manually constructed resources that do not transfer to other domains.

Yadav et al. [19] followed with a machine-learning study of bilingual sentiment classification for Punjabi–English code-mixed social media text, again confronting spelling inconsistency, transliteration, and lexical ambiguity. The corpus was small and is not publicly available, which restricts both the generalization of the reported models and the comparability of the results, but the study broadened the evidence beyond a single domain and remains one of the most directly relevant contributions to this specific problem.

Tiwari et al. [20] represent the current state of the primary literature. Their system classifies English–Punjabi code-mixed posts into positive, negative, and neutral classes using logistic regression over cleaned and vectorized text, trained on 37,805 labeled posts and evaluated on 10,307 test posts, attaining 92% test accuracy, and is deployed through a real-time Streamlit web interface. Two observations follow. First, this is by a wide margin the largest labeled Punjabi–English sentiment corpus reported to date, though the paper does not state that the data are publicly released. Second, the modeling is deliberately simple—a linear classifier over surface features—which sets a strong, reproducible baseline but leaves the benefit of contextual neural models on Punjabi–English text an open empirical question rather than an established fact.

Two review-oriented contributions complete this strand. Sekhri and Bansal [21] reviewed sentiment analysis methods for vernacular Indian languages including Punjabi, summarizing lexicon-based and machine-learning approaches; the review does not address code-mixed text, which is precisely the gap the present survey covers. Sekhri et al. [22] applied the Indic NLP library with N-gram features to Punjabi text classification, illustrating the value of language-specific preprocessing toolkits in resource-constrained settings.

4.4 Related Work: Urdu–Punjabi and Shahmukhi Resources

The most substantial recent resource development has occurred not in Punjabi–English but in the adjacent Urdu–Punjabi code-switched setting. Hussain et al. [23] introduced LUPCSA-25, a corpus of more than one million user reviews in Urdu and Shahmukhi Punjabi spanning food, sports, entertainment, applications, and political commentary, annotated with the involvement of doctoral-level Urdu and Punjabi specialists, and trained a framework integrating XLM-RoBERTa and GPT-2 with head pruning on it. The study demonstrates that multilingual transformers handle code-switched Punjabi-family text effectively at scale. Its relevance to the present survey must nevertheless be stated precisely: LUPCSA-25 is Urdu–Punjabi in Shahmukhi script, not Punjabi–English in Romanized script, so the transfer of these results to the Romanized code-mixed setting remains to be demonstrated. Complementary Shahmukhi resources, including word embeddings and text-classification datasets [24], strengthen the Pakistani-Punjabi side of the resource landscape and could support cross-script transfer experiments.

Taken together, the literature shows a clear methodological progression—from lexicons and scoring [15], through classical supervised learning [16], [18], [19], [20],

to deep and transformer-based models [17], [23]—but an equally clear evidential gap: transformer methods have been validated on monolingual Gurmukhi and on Shahmukhi Urdu–Punjabi, while the Romanized Punjabi–English setting has so far been addressed only with classical classifiers.

5. DATASETS, PREPROCESSING, FEATURE REPRESENTATION, AND EVALUATION

5.1 Datasets

Table 1 catalogues the datasets used in the surveyed studies, together with the SemEval-2020 SentiMix Hindi–English corpus as the nearest public benchmark. The pattern is stark: every Punjabi or Punjabi–English dataset in the primary literature is either small, domain-specific, not publicly released, or all three. The largest resource associated with Punjabi sentiment research, LUPCSA-25, belongs to the Urdu–Punjabi Shahmukhi setting. Annotation schemes also differ across studies—binary polarity, three-way polarity, and four-way fine-grained negative classes all appear—so results are not directly comparable even where they are reported. The single most consequential step available to this field is the construction and public release of a standardized, multi-domain Punjabi–English code-mixed sentiment benchmark.

Table 1. Datasets used in Punjabi and Punjabi–English sentiment analysis research, with related benchmarks.

Study	Language(s)	Script	Domain	Size	Availability
Arora & Kaur, 2015 [16]	Punjabi	Gurmukhi	Political reviews	Small (not stated)	Not public
Kaur & Gupta, 2017 [17]	Punjabi	Gurmukhi	General text	Not stated	Not public
Singh et al., 2019 [19]	Punjabi–English	Romanized	Agriculture tweets	Not stated	Not public
Yadav et al., 2020 [20]	Punjabi–English	Romanized	Social media	Small (not stated)	Not public
Patwa et al., 2020 [14] (related benchmark)	Hindi–English	Romanized	Tweets	≈20,000 tweets	Public
Singh et al., 2021 [18]	Punjabi	Gurmukhi	Farmer-suicide news reports	275 documents	Not public
Hussain et al., 2025 [24] (related resource)	Urdu–Punjabi	Shahmukhi	Multi-genre user reviews	>1,000,000 reviews	Research use
Tiwari et al., 2025 [21]	Punjabi–English	Romanized	Social media posts	48,112 posts (37,805 train / 10,307 test)	Not stated

5.2 Preprocessing

The surveyed studies converge on a common pipeline whose stages are dictated by the properties of code-mixed social media text. Cleaning removes URLs, HTML artifacts, duplicate characters, and extraneous symbols, with emojis either stripped or retained as sentiment cues. Tokenization must contend with hashtags, abbreviations, and transliterated words. Token-level language identification then assigns each token to Punjabi or English so that language-appropriate processing can follow; Singh et al. [18] treat this as the pivotal stage of their pipeline. Normalization maps the many Romanized spellings of a word onto a canonical form, in several studies with the help of bilingual dictionaries [18], [19], while morphological normalization plays the corresponding role for Gurmukhi text [17]. Stop-word removal requires separate lists for the two languages, and stemming and lemmatization remain only partially solved for Punjabi given the immaturity of the available tools. No two studies use identical pipelines, and none has been released as reusable software—a further obstacle to reproducibility.

5.3 Feature Representation

Feature practice tracks the general history of the field. The classical studies use Bag-of-Words and TF–IDF representations with SVM and Naïve Bayes classifiers [16], [18]; word- and character-level N-grams add local context, and character-level features are particularly attractive for code-mixed text because they absorb spelling and transliteration variation [22]. Linguistically informed features appear in the morphology-aware framework of Singh et al. [17]. Dense embeddings—Word2Vec [25], GloVe [26], and FastText [27]—capture distributional semantics, with FastText's subword modeling especially suited to morphologically rich, orthographically unstable text, though pretrained Punjabi embeddings of adequate quality and coverage remain scarce. Contextual representations from multilingual transformers [6]–[9] are the current standard elsewhere in multilingual NLP; within the Punjabi family they have been demonstrated on Shahmukhi Urdu–Punjabi [23] but not yet on Romanized Punjabi–English.

5.4 Evaluation Metrics

Reported evaluation uses the standard classification metrics derived from the confusion matrix, whose entries are true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$Precision = TP / (TP + FP) \quad (2)$$

$$Recall = TP / (TP + FN) \quad (3)$$

$$F1 = (2 \times Precision \times Recall) / (Precision + Recall) \quad (4)$$

Because sentiment classes in social media corpora are typically imbalanced, macro-averaged and weighted F1 are preferable to raw accuracy for multi-class settings, and shared-task practice in the neighboring Hindi–English literature reports them as standard [13]. The Punjabi studies surveyed here mostly report accuracy alone, which should be borne in mind when reading Table 2.

6. COMPARATIVE ANALYSIS

Table 2 compares the surveyed studies on verified methodological detail and published results. Where a study's numerical results could not be confirmed against the published record, the table says so explicitly rather than reproducing unverifiable figures.

Table 2. Comparative summary of Punjabi and Punjabi–English sentiment analysis studies.

Study	Languages	Approach	Features	Reported results	Contribution	Limitations
Arora & Kaur [16]	Punjabi	Lexicon-based scoring	Polarity lexicon scores	Qualitative polarity summaries; no benchmark accuracy reported	Early feasibility of Punjabi polarity detection	Small private data; no supervised evaluation
Kaur & Gupta [17]	Punjabi	SVM over Punjabi subjective lexicon	Lexicon + statistical features	78.02% accuracy	Cross-lingual lexicon construction from Hindi	Limited contextual modeling; data not public
Singh et al., 2019 [19]	Punjabi–English	SVM, Naïve Bayes after language ID + normalization	Statistical features + bilingual dictionary	Classifier comparisons reported in paper	First dedicated Punjabi–English code-mixed pipeline	Agriculture domain only; manual resources
Yadav et al., 2020 [20]	Punjabi–English	Classical ML classification	Text-based features	Results reported on a small bilingual corpus	Extended code-mixed evidence beyond a single domain	Small private corpus; limited generalization
Singh et al., 2021 [18]	Punjabi	Morphology-aware deep neural network	Morphological features	90.29% overall; per-class 93.85 / 88.53 / 83.30 / 95.45% (10-fold CV)	Deep learning for Gurmukhi Punjabi; four fine-grained negative classes	275 documents; single (farmer-suicide) domain
Sekhri & Bansal, 2025 [22]	Punjabi (review)	Survey of ML approaches	N/A	N/A (review)	Consolidated monolingual Punjabi SA literature	Does not cover code-mixed text
Hussain et al., 2025 [24]	Urdu–Punjabi (Shahmukhi)	XLM-RoBERTa + GPT-2 with head pruning	Contextual embeddings	Full experimental results reported on LUPCSA-25	Largest Punjabi-family sentiment corpus (>1M reviews)	Not Punjabi–English; high compute cost
Tiwari et al., 2025 [21]	Punjabi–English	Logistic regression	Cleaned, vectorized text	92% test accuracy (37,805 train / 10,307 test)	Largest labeled Punjabi–English corpus; deployed real-time interface	Linear model; three coarse classes; availability not stated

Three findings emerge. First, classical machine learning remains the only approach actually demonstrated on Romanized Punjabi–English text, and it performs respectably: the strongest reported result in that setting, 92% accuracy [20], comes from logistic regression over a comparatively large corpus, which suggests that data volume has so far mattered more than model sophistication. Second, the demonstrated benefits of deep and transformer-based modeling within the Punjabi family come from adjacent settings—Gurmukhi monolingual text [17] and Shahmukhi Urdu–Punjabi [23]—and their transfer to Romanized code-mixed text is a plausible but untested hypothesis. Third, no two primary studies share a dataset, annotation scheme, or evaluation protocol, so cross-study numerical comparison is not meaningful; the accuracy figures in Table 2 characterize individual studies, not a ranking.

Figure 2 places these eight studies on a single timeline by publication year and approach category, and the pattern it shows is the same one argued qualitatively above: classical machine learning (SVM, Naïve Bayes, logistic regression) is the only approach demonstrated on Romanized Punjabi–English text at any point in the 2015–2025 window, while the two studies that moved beyond it—Singh et al. [17] and Hussain et al. [23]—did so in the monolingual Gurmukhi and Shahmukhi Urdu–Punjabi settings respectively, not in the code-mixed Romanized setting this survey centers on.

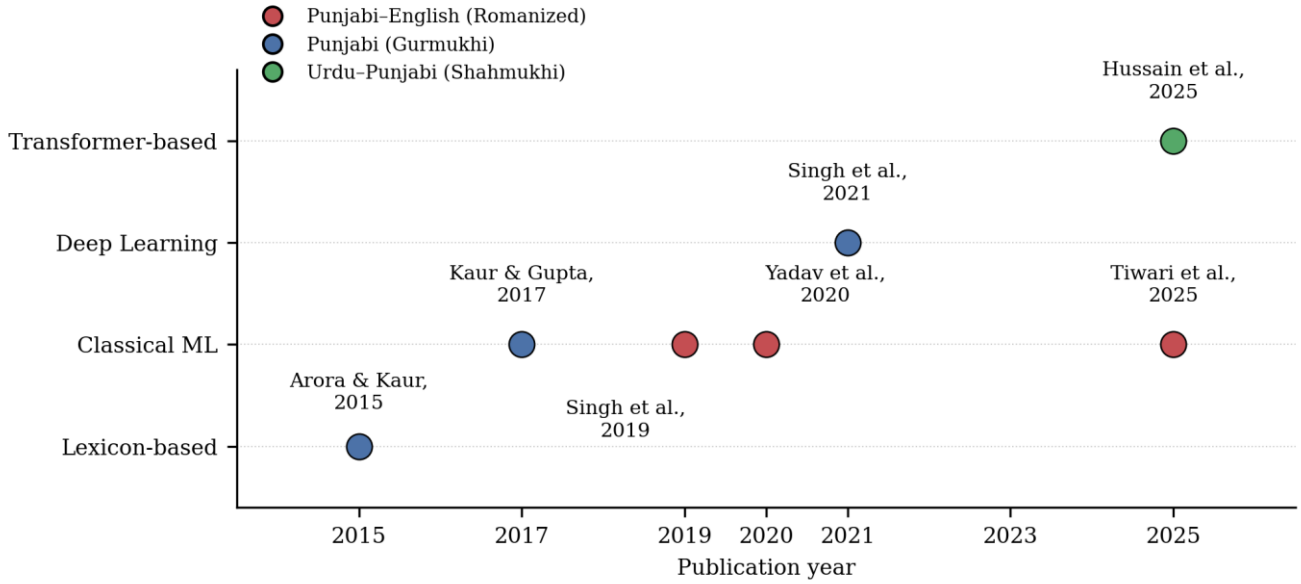

Figure 2. Publication year and computational approach for each surveyed study, grouped by script/language setting.

Figure 3 summarizes dataset availability from Table 1: five of the eight datasets are explicitly not public, one is not stated, and only the two related resources—Patwa et al.'s SemEval-2020 corpus [13] and Hussain et al.'s LUPCSA-25 [23]—are public or available for research use. None of the primary Punjabi or Punjabi–English studies has released its data, consistent with the absence of a shared benchmark noted throughout this survey.

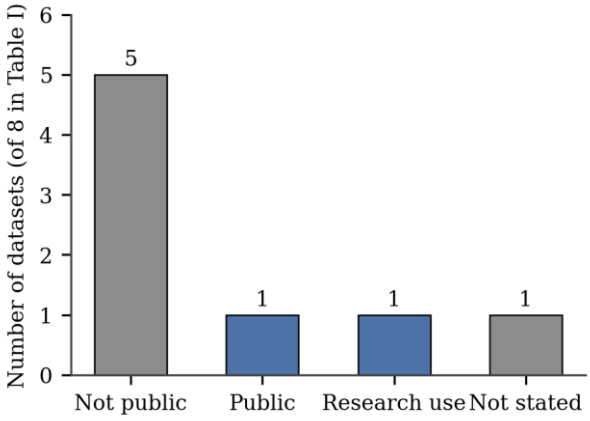

Figure 3. Availability of the datasets catalogued in Table 1.

7. CHALLENGES AND FUTURE DIRECTIONS

7.1 Benchmark and Dataset Development

The absence of a public, multi-domain, consistently annotated Punjabi–English code-mixed benchmark is the field's binding constraint. Every dataset in Table 1 except the related Hindi–English and Shahmukhi resources is private. A SentiMix-style shared corpus [13]—with token-level language tags, standardized three-way polarity labels, published annotation guidelines and agreement statistics, and train/test splits—would do more for this area than any modeling contribution, and the corpus-construction methodology is already well established in the Hindi–English literature [12], [13].

7.2 Script Normalization and Transliteration

Robust handling of Gurmukhi, Shahmukhi, and Romanized Punjabi within one system remains open. Needed components include Romanization normalizers that map variant spellings to canonical forms, transliteration models between the three scripts, and evaluation sets for both. Shahmukhi resources [23], [24] and Gurmukhi tools [22] provide starting points for the two standard scripts; the Romanized form, though the most common online, is the least resourced.

7.3 Punjabi-Adapted Pretrained Models

Multilingual transformers [6]–[9] include Punjabi in their pretraining data only thinly, and Romanized Punjabi essentially not at all. Continued pretraining on Punjabi text across scripts, and evaluation of parameter-efficient adaptation methods (adapters, LoRA-style fine-tuning, prompt-based few-shot use of large language models), are natural next steps. The Urdu–Punjabi results of Hussain et al. [23] indicate that transformer capacity is not the obstacle; Punjabi-specific data is.

7.4 Informal Language, Sarcasm, and Cultural Context

Emojis, slang, elongations, and dialectal forms carry sentiment signal that current pipelines mostly discard, and sarcasm—annotated for Hindi–English [12]—has no Punjabi–English counterpart. Sentiment lexicons that cover Romanized Punjabi, emoji-aware models, and culturally informed annotation would each address a documented weakness of the existing studies [18], [19].

7.5 Evaluation Practice

Future studies should report macro- and weighted F1 alongside accuracy, publish class distributions, and, wherever possible, evaluate on shared data. Until a shared benchmark exists, authors can improve comparability by releasing their splits and preprocessing code.

8. CONCLUSION

This survey consolidated the research on sentiment analysis of Punjabi and Punjabi–English code-mixed text published between 2014 and 2025. The literature progresses from

lexicon-based scoring through classical supervised learning to deep and transformer-based models, but the progression is uneven across settings: classical classifiers are the only methods yet demonstrated on the Romanized Punjabi–English text that dominates social media, while neural and transformer results come from monolingual Gurmukhi and from the related Shahmukhi Urdu–Punjabi setting. The primary code-mixed literature remains small, its datasets private and heterogeneous, and its results mutually incomparable. The field's clearest needs are a public multi-domain benchmark with standardized annotation, normalization and transliteration resources spanning Punjabi's three written forms, and Punjabi-adapted pretrained models. Progress on these fronts would allow Punjabi–English sentiment analysis to move from isolated feasibility studies to the cumulative, comparable research that neighboring language pairs already enjoy, and would extend the reach of language technology to one of the world's largest and least-served speech communities.

9. DECLARATIONS

Funding: The author received no specific funding for this work.

Conflict of Interest: The author declares no conflict of interest.

Data Availability: This is a survey article and does not generate new primary data. The publications reviewed are cited in the reference list; several of the datasets they describe are not publicly available, as documented in Table 1.

10. REFERENCES

[1] M. Kumar, L. Khan, and H.-T. Chang, "Evolving techniques in sentiment analysis: A comprehensive review," PeerJ Computer Science, vol. 11, e2592, 2025, doi: 10.7717/peerj-cs.2592.

[2] D. M. Eberhard, G. F. Simons, and C. D. Fennig, Eds., Ethnologue: Languages of the World, 27th ed. Dallas, TX, USA: SIL International, 2024.

[3] Office of the Registrar General & Census Commissioner, India, Census of India 2011: Language and Mother Tongue Data. New Delhi, India: Ministry of Home Affairs, Government of India, 2011.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] Y. Kim, "Convolutional neural networks for sentence classification," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746–1751.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[7] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in Proc. 58th Annu. Meeting Assoc. Computational Linguistics (ACL), 2020, pp. 8440–8451.

[8] D. Kakwani et al., "IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages," in Findings of EMNLP, 2020, pp. 4948–4961.

[9] S. Khanuja et al., "MuRIL: Multilingual representations for Indian languages," arXiv preprint arXiv:2103.10730, 2021.

[10] K. Bali, J. Sharma, M. Choudhury, and Y. Vyas, "'I am borrowing ya mixing?' An analysis of English-Hindi code mixing in Facebook," in Proc. 1st Workshop on Computational Approaches to Code Switching, 2014, pp. 116–126.

[11] R. Bhargava, Y. Sharma, and S. Sharma, "Sentiment analysis for mixed script Indic sentences," in Proc. Int. Conf. Advances in Computing, Communications and Informatics (ICACCI), 2016, pp. 524–529.

[12] S. Swami, A. Khandelwal, V. Singh, S. S. Akhtar, and M. Shrivastava, "A corpus of English-Hindi code-mixed tweets for sarcasm detection," arXiv preprint arXiv:1805.11869, 2018.

[13] P. Patwa et al., "SemEval-2020 Task 9: Overview of sentiment analysis of code-mixed tweets," in Proc. 14th Workshop on Semantic Evaluation (SemEval-2020), 2020, pp. 774–790.

[14] S. Khanuja, S. Dandapat, A. Srinivasan, S. Sitaram, and M. Choudhury, "GLUECoS: An evaluation benchmark for code-switched NLP," in Proc. 58th Annu. Meeting Assoc. Computational Linguistics (ACL), 2020, pp. 3575–3585.

[15] P. Arora and B. Kaur, "Sentiment analysis of political reviews in Punjabi language," International Journal of Computer Applications, vol. 126, no. 14, pp. 20–23, 2015, doi: 10.5120/ijca2015906297.

[16] A. Kaur and V. Gupta, "A novel approach for sentiment analysis of Punjabi text using SVM," International Arab Journal of Information Technology, vol. 14, no. 5, 2017.

[17] J. Singh, G. Singh, R. Singh, and P. Singh, "Morphological evaluation and sentiment analysis of Punjabi text using deep learning classification," Journal of King Saud University – Computer and Information Sciences, vol. 33, no. 5, pp. 508–517, 2021, doi: 10.1016/j.jksuci.2018.04.003.

[18] M. Singh, V. Goyal, and S. Raj, "Sentiment analysis of English-Punjabi code mixed social media content for agriculture domain," in Proc. 4th Int. Conf. Information Systems and Computer Networks (ISCON), 2019, doi: 10.1109/ISCON47742.2019.9036204.

[19] K. Yadav, A. Lamba, D. Gupta, A. Gupta, P. Karmakar, and S. Saini, "Bilingual sentiment analysis for a code-mixed Punjabi English social media text," in Proc. 2020 5th Int. Conf. Computing, Communication and Security (ICCCS), 2020, pp. 1–5, doi: 10.1109/ICCCS49678.2020.9277309.

[20] A. Tiwari, J. Sehgal, M. Singh, and A. Mishra, "Sentiment analysis in English-Punjabi mixed social media posts," in Proc. 2025 IEEE Int. Conf. Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), vol. 3, 2025, pp. 1–6.

[21] S. Sekhri and K. Bansal, "A review on Punjabi language sentiment analysis using machine learning," Journal of Computer Technology & Applications, vol. 16, no. 2, pp. 116–122, 2025.

[22] S. Sekhri, N. Shrivastav, P. Kaur, K. Bansal, S. Wadhwa, and P. Bhalla, "Punjabi text sentiment analysis using Indic

NLP," OmniScience: A Multi-disciplinary Journal, vol. 15, no. 1, pp. 38–42, 2025.

[23] M. Hussain, S. Ali, H. Sattar, A. Raza, M. H. Akbar, and M. A. Rafiq, "Urdu-Punjabi code switched sentiment analysis empowered by a deep learning framework integrating XLM-R and GPT," VAWKUM Transactions on Computer Sciences, vol. 13, no. 2, pp. 1–20, 2025, doi: 10.21015/vtcs.v13i2.2144.

[24] M. Shabbir et al., "Advancing NLP for Shahmukhi Punjabi: Word embedding and text classification with a novel dataset," VAWKUM Transactions on Computer Sciences, vol. 13, no. 1, pp. 22–43, 2025.

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems (NeurIPS), 2013, pp. 3111–3119.

[26] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[27] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146, 2017.