Dyal Singh Journal of Humanities and Social Sciences (DSJHSS)
Vol. 4, No. 1, 2026, pp. 44-61
ISSN (P): 3006-6522
ISSN (E): 3006-6530

PSYCHO-FORENSIC LINGUISTIC SURVEILLANCE INITIATIVE SERIES PAPER 5

From Lexicon to Prediction: A Protocol for Transformer Fine-Tuning and Multilingual Validation of a Psycho-Forensic Linguistic Framework for Suicide-Related Discourse Detection

Afshan Ishfaq[1]
Nida Sultan[2]

Received: 15-03-26

Reviewed: 16=05-26

Accepted: 28-06-26

Published: 30-06-26

Abstract

Suicide prevention requires attention not only to clinically recognized crises but also to communicative signals that may precede or accompany them. Language offers an observable behavioural trace through which distress can be studied outside formal clinical encounters, although linguistic evidence must not be treated as a diagnosis. The Psycho-Forensic Linguistic Surveillance Initiative (PFLSI) has progressed from culturally situated Pakistani suicide notes to corpus-scale English analysis, theoretical integration, and a proposed multilingual Urdu-Punjabi corpus. Series Paper 5 extends that programme by specifying a protocol for transformer-based modelling and multilingual validation of suicide-related discourse in English, Urdu, Punjabi, Roman Urdu, and Urdu-English code-switched communication. Because the proposed Multilingual Suicide Discourse Corpus (M-SDC) has not yet been completed, this article does not report empirical model results. Instead, it prespecifies the classification task, corpus-matching principles, annotation procedures, baseline and transformer models, culturally informed feature streams, validation regimes, evaluation metrics, ablation strategy, explainability procedures, and ethical boundaries to be implemented after corpus completion. The protocol compares corpus-linguistic and classical machine-learning baselines with multilingual BERT and XLM-RoBERTa, while testing whether structured linguistic and culturally situated features add information beyond contextual representations. Evaluation prioritizes discrimination, calibration, source-held-out generalization, cross-linguistic and cross-script transfer, subgroup stability, and interpretability. The central argument is that multilingual suicide-discourse modelling should be evaluated not by predictive performance alone but by linguistic validity, cultural calibration, source generalization, transparent error analysis, and strict separation between computational risk indication and clinical judgement.

Keywords: suicide discourse; psycho-forensic linguistics; multilingual NLP; XLM-RoBERTa; cultural calibration

This is an open access article distributed under the terms of CC-BY-4.0, which permits unrestricted use, distribution, reproduction, and adaptation in any medium and for any purpose with proper attribution to the original author(s), title, journal, and URL.

[1] Head Academics, Institute of Law, Lahore
[2] Lecturer in English, NAMAL University, Mianwali

ISSN (P): 3006-6522
ISSN (E): 3006-6530

1. INTRODUCTION

1.1 Background of the Study

Suicide is a complex public-health phenomenon shaped by psychological, interpersonal, social, cultural, and communicative processes (World Health Organization, 2023). Before an acute crisis becomes visible through clinical contact or overt behaviour, aspects of distress may already be expressed through language. This makes language a legitimate research object for prevention-oriented inquiry, particularly when lexical choice, self-reference, temporal orientation, narrative structure, interpersonal positioning, and metaphor are studied as patterns rather than as deterministic symptoms. Linguistic evidence cannot establish a clinical diagnosis or predict an individual's behaviour on its own; at most, it can contribute research-level signals that require contextual interpretation and, in any applied setting, human professional judgement.

The PFLSI programme was established to investigate these possibilities through a cumulative research design. Series Paper 1, Decoding Despair, examined a culturally rich corpus of 140 Pakistani suicide notes and identified recurrent themes in apology, guilt, relational obligation, coercion, and self-evaluation (Ishfaq, Ahmad, & Sultan, 2025). Series Paper 2, Language of Stress, expanded the evidence base through a 452,000-token English-language Suicide Discourse Corpus and a matched control corpus (Ishfaq & Sultan, 2025c). Series Paper 3, The Lexicon of Crisis, integrated psychological and linguistic theories into a seven-layer psycho-forensic linguistic architecture (Ishfaq, Sultan, & Bhatti, 2026). Series Paper 4, Voices Across Scripts, identified the multilingual resource gap and proposed a 500,000-token Urdu-Punjabi Suicide Discourse Corpus with a matched control corpus (Ishfaq & Sultan, 2026b). The present paper constitutes Series Paper 5 and addresses the next methodological problem: how those accumulated linguistic and cultural observations can be translated into a reproducible computational validation protocol.

1.2 From Series 1 to Series 5

The five papers form a cumulative research programme. Series Paper 1 established culturally situated qualitative evidence; Series Paper 2 added corpus-scale quantitative evidence; Series Paper 3 developed theoretical and architectural integration; and Series Paper 4 specified the design and governance requirements for multilingual corpus construction. Series Paper 5 does not claim completed transformer validation. Rather, it prespecifies the modelling and validation procedures required to test whether the earlier observations generalize across languages, scripts, sources, and model families.

Series Paper 4 describes the central unresolved issue as a monolingual ceiling. The strongest quantitative evidence in the programme was produced from the 452,000-token English corpus, whereas the most culturally specific evidence came from the smaller Pakistani suicide-note corpus. Cultural calibration therefore remains a hypothesis to be tested rather than an established property of the system until sufficiently large Urdu, Punjabi, Roman Urdu, and code-switched data are collected, annotated, and evaluated.

1.3 Statement of the Problem

There is no adequate basis for assuming that linguistic markers identified in English retain the same statistical meaning in Urdu, Punjabi, Roman Urdu, or code-switched Pakistani discourse. Low-resource languages face additional problems of limited annotated data, orthographic variability, domain mismatch, and uneven pretrained-model coverage (Khattak et al., 2021; Bialer et al., 2022). Series Paper 4 identifies three interconnected gaps: the absence of a native-script Urdu and Punjabi suicide-discourse corpus comparable with the English corpus, the scarcity of mature Urdu computational resources, and the lack of an annotated resource for Urdu-English crisis code-switching (Ishfaq & Sultan, 2026b). The present protocol addresses these gaps while

ISSN (P): 3006-6522
ISSN (E): 3006-6530

also distinguishing its target construct from clinical suicide prediction: the primary task is classification of suicide-related discourse, not diagnosis of suicidal ideation or prediction of an individual's future behaviour.

1.4 Research Objectives

- To develop a reproducible transformer-based protocol for multilingual suicide-discourse classification.
- To compare multilingual transformer models with corpus-linguistic and classical machine-learning baselines.
- To test whether culturally specific features provide incremental value for multilingual suicide-discourse classification.
- To evaluate cross-linguistic, cross-script, subgroup, and source-held-out generalization.

1.5 Research Questions

- Can multilingual transformer models distinguish suicide-related discourse from matched non-suicide-related discourse across English, Urdu, Punjabi, Roman Urdu, and Urdu-English code-switched text?
- Does multilingual transformer fine-tuning outperform lexical, corpus-linguistic, and classical machine-learning baselines under both random and source-held-out validation?
- Do culturally situated features, including the izzat register, relational guilt, honour-shame framing, metaphor, and code-switching, improve classification and calibration beyond contextual representations alone?
- How well do learned representations transfer across languages, scripts, and source domains, and where does performance degrade?

1.6 Hypotheses

H1: Multilingual transformer models will outperform classical lexical baselines on prespecified discrimination metrics.

H2: XLM-RoBERTa will demonstrate stronger cross-linguistic transfer than monolingual English models.

H3: Adding structured linguistic and culturally situated features will improve performance or calibration relative to contextual representations alone.

H4: Culturally situated features will provide their greatest incremental value in Urdu, Punjabi, and Roman Urdu discourse.

H5: Explicit code-switch information will improve robustness on mixed-language material.

H6: Pain and suffering, first-person reference, future orientation, absolutist language, isolation, and relational language will emerge as important explanatory feature groups.

H7: Performance will decline under source-held-out validation relative to random document-level validation if models learn source-specific cues.

H8: The full proposed model will demonstrate greater subgroup stability across language, script, gender where valid metadata exist, and source.

1.7 Significance of the Study

The study contributes to psycho-forensic linguistics by specifying how theoretically grounded linguistic indicators can be represented computationally without abandoning cultural interpretation. It also contributes to Urdu and Punjabi NLP by defining a structured agenda for low-resource mental-health language modelling. Methodologically, the proposed novelty lies in combining contextual multilingual representations with interpretable corpus-linguistic and

ISSN (P): 3006-6522
ISSN (E): 3006-6530

culturally situated features, then evaluating them through source-held-out testing, cross-script transfer, calibration, subgroup analysis, and ablation rather than accuracy alone. This contribution is prospective: its empirical value remains to be demonstrated once the M-SDC is completed and the experiments are run.

1.8 Scope and Delimitations

The study concerns text-based suicide-related discourse and does not claim to diagnose suicidal ideation or predict individual behaviour. The primary positive class will consist of textual units that meet prespecified source and annotation criteria for suicide-related content, such as explicit discussion of suicidal ideation, intent, planning, attempts, or verified suicide-note provenance. The comparison class will consist of register- and source-matched material that does not meet those criteria. Ambiguous cases will be separately flagged during annotation rather than forced into a binary label. The initial modelling task is therefore a research classification problem; finer-grained clinical categories and individual-level risk claims are reserved for later work requiring clinically anchored labels and prospective validation.

2. LITERATURE REVIEW

2.1 Language and Psychological Crisis

Language is one of the few observable traces through which psychological processes can be studied in naturally occurring communication. Psycholinguistic research has examined pronouns, emotion words, temporal orientation, cognitive-processing terms, social references, and narrative structure as indicators of attention, self-focus, and interpersonal orientation (Pennebaker, 2011). In suicide-related research, absolutist language and other lexical patterns have also been investigated as correlates of distress (Al-Mosaiwi & Johnstone, 2018). Such indicators must nevertheless be interpreted cautiously: the same expression can serve different communicative purposes across speakers, genres, languages, and cultures, and no isolated lexical feature should be treated as a diagnostic marker.

2.2 Suicide Notes and Psycho Forensic Linguistics

Series Paper 1 positioned suicide notes as communicative artefacts rather than merely collections of negative words. Its Pakistani corpus contained 140 notes and identified culturally situated patterns involving apology, regret, self-worth, guilt, family obligation, coercion, and relational discourse (Ishfaq, Ahmad, & Sultan, 2025). Series Paper 4 subsequently proposed using the original note corpus as a potential validation subset through which those qualitative findings could be tested quantitatively (Ishfaq & Sultan, 2026b). This progression motivates the present emphasis on retaining interpretable linguistic constructs alongside transformer representations.

2.3 Corpus Linguistics and Suicide Discourse

Series Paper 2 expanded the research from individual notes to corpus-scale comparison. Its 452,000-token Suicide Discourse Corpus was analysed against a matched control corpus using log-likelihood, semantic-domain profiling, collocation, and effect-size analysis (Ishfaq & Sultan, 2025c). The implication for Series Paper 5 is methodological: transformer modelling should extend rather than erase these corpus-linguistic foundations. Earlier indicators should therefore be retained as interpretable baselines and as structured feature groups that can be tested through ablation.

2.4 Computational Mental Health NLP

Computational mental-health research has frequently relied on English-language social-media or online-support data, using lexical features, classical machine learning, deep learning, and increasingly pretrained language models (Coppersmith et al., 2014; Gkotsis et al., 2017). Related work has shown both the promise of language-based risk modelling and the importance of domain-

ISSN (P): 3006-6522
ISSN (E): 3006-6530

specific data, clinically meaningful cues, and careful validation (Bialer et al., 2022; Izmaylov et al., 2023). At the same time, social-media proxies and platform-specific cues can make apparently strong performance difficult to generalize beyond the source on which a model was trained. The present protocol therefore treats source-held-out validation and explicit construct definition as central requirements rather than optional secondary analyses.

2.5 Multilingual and Low Resource NLP

Multilingual transformer models offer a potential response to data scarcity because multilingual pretraining can create shared contextual representations. Multilingual BERT established the feasibility of fine-tuning bidirectional transformer representations across many downstream tasks (Devlin et al., 2019), while XLM-RoBERTa demonstrated strong cross-lingual transfer at substantially larger multilingual scale, including gains for low-resource languages such as Urdu (Conneau et al., 2020). Shared representation, however, does not guarantee cultural or domain equivalence. Low-resource languages may remain disadvantaged by limited pretraining data, orthographic variation, code-switching, uneven tokenization, and source mismatch. Accordingly, multilingual performance must be reported separately by language and script rather than inferred from aggregate scores alone.

2.6 Urdu, Punjabi, Roman Urdu and Code Switching

Urdu presents computational challenges associated with limited annotated resources, spelling variation, script complexity, and code-switching (Khattak et al., 2021). Roman Urdu adds substantial orthographic variability, while Punjabi in Pakistan introduces an additional script issue because Shahmukhi resources cannot be assumed to behave like resources developed for Gurmukhi or other representations. Work on code-switched language identification demonstrates the importance of token-level language labels and shared annotation conventions (Bali et al., 2014; Solorio et al., 2014; Molina et al., 2016). Series Paper 4 therefore proposes token-level code-switch annotation adapted from the CALCS shared-task tradition for Urdu-English and related mixed-language material (Ishfaq & Sultan, 2026b).

2.7 The Izzat Problem

A central contribution of the PFLSI programme is the hypothesis that culturally embedded suicide-related discourse may include honour and shame, family obligation, relational guilt, domestic entrapment, coercion, and socially constrained self-positioning. Series Paper 4 describes this as the izzat problem and proposes an inductively constructed, culturally specific lexicon and annotation layer (Ishfaq & Sultan, 2026b). In the present protocol, the izzat register is treated as a testable feature family rather than as a universal marker: its construct validity, reliability, prevalence, and incremental predictive value must be demonstrated empirically.

2.8 Theoretical Framework

The theoretical framework integrates seven perspectives carried forward from Series Paper 3. Shneidman's psychache theory provides the construct of psychological pain (Shneidman, 1993). Joiner's Interpersonal Theory contributes perceived burdensomeness and thwarted belongingness (Joiner, 2005). Beck's cognitive model and the Hopelessness Scale literature inform negative future orientation and pessimistic cognition (Beck, 1963; Beck et al., 1974). O'Connor's Integrated Motivational-Volitional Model contributes defeat, entrapment, and the distinction between motivational and volitional processes (O'Connor, 2011). Pennebaker's work informs analysis of pronouns and broader language-use patterns (Pennebaker, 2011). Conceptual Metaphor Theory supports analysis of metaphor as structured meaning rather than decorative language (Lakoff & Johnson, 1980; Kövecses, 2015). Menon's Narrative Crisis Model contributes narrative disruption

ISSN (P): 3006-6522
ISSN (E): 3006-6530

and crisis framing (Menon, 2024). These perspectives are used as complementary interpretive anchors, not as interchangeable clinical measures.

2.9 Conceptual Model

The conceptual model proposes that psychological crisis may be represented through interacting linguistic dimensions rather than a single vocabulary. General dimensions include psychological pain, self-reference, temporal orientation, absolutist language, isolation, and negative evaluation. Culturally situated dimensions include relational guilt, family obligation, honour-shame framing, coercion, and entrapment. Transformer representations provide contextual modelling, whereas structured linguistic and cultural features preserve interpretability and permit explicit tests of incremental value. The model therefore treats cultural calibration as an empirical question to be evaluated through reliability, ablation, subgroup analysis, and cross-source generalization.

2.10 Research Gap

The research gap is not simply the absence of a multilingual classifier. Existing work demonstrates that pretrained language models can support suicide-related classification and that low-resource settings create distinct challenges (Bialer et al., 2022; Izmaylov et al., 2023). The unresolved problem addressed here is the absence of a validated bridge between corpus-linguistic evidence, culturally specific Pakistani discourse, multilingual representation, and robust cross-source computational evaluation. Series Paper 4 explicitly states that corpus construction had not been completed and that empirical findings remained to be generated (Ishfaq & Sultan, 2026b). The contribution of Series Paper 5 is therefore a prespecified validation architecture whose novelty must ultimately be judged by completed multilingual experiments rather than by proposal alone.

RESEARCH METHODOLOGY

3.1 Research Design

The study adopts a mixed computational and corpus-linguistic protocol. It combines supervised transformer classification with corpus-based feature analysis, cultural annotation, ablation experiments, source-held-out validation, cross-linguistic testing, calibration, and explainability analysis. The design is sequential: corpus construction precedes annotation; annotation and codebook stabilization precede model training; model development is restricted to the training and validation partitions; and final evaluation is conducted only on held-out test sets. This combination of purposive corpus construction and quantitative comparison is consistent with mixed-method sampling principles in which sampling decisions are tied to the inferential purpose of the study (Teddlie & Yu, 2007).

3.2 Corpus Design

The M-SDC design proposed in Series Paper 4 targets 500,000 tokens of Urdu and Punjabi suicide-related discourse and a matched 500,000-token control corpus. The proposed source composition includes 35% crisis-helpline transcripts, 35% moderated forums and Roman Urdu social media, 10% re-digitized Series Paper 1 notes, 10% published first-person narratives, and 10% Punjabi Shahmukhi supplementary material. These figures are design targets, not achieved collection totals. Because the genres differ substantially, source proportions will be documented at the document and token levels, and no source will be allowed to occur exclusively in only one class if a matched counterpart can be ethically and practically obtained.

3.3 Data Sources and Matching

Control material should be matched as closely as possible by register, platform type, language/script, approximate publication period, and document length. Series Paper 4 specifies equivalent scale and register matching and excludes material flagged for suicide-related content

ISSN (P): 3006-6522
ISSN (E): 3006-6530

during control sampling (Ishfaq & Sultan, 2026b). Matching is necessary because a classifier can otherwise learn source or genre characteristics rather than suicide-related linguistic patterns. The protocol therefore requires source-stratified descriptive tables before modelling and sensitivity analyses that separate within-source discrimination from cross-source generalization.

3.3.1 Operational Definition of the Classification Task

The primary unit of analysis will be a bounded textual document or communicative unit defined before sampling. For suicide notes and short posts, the unit will ordinarily be the complete text. For longer helpline conversations or narratives, segmentation rules will be fixed in advance; all segments from the same conversation or author-level source will remain grouped during partitioning. The positive label denotes suicide-related discourse according to the source and annotation criteria stated in Section 1.8, not a clinical diagnosis. Each candidate item should be independently reviewed against a written inclusion/exclusion codebook, with ambiguous cases retained for adjudication or excluded from the binary training set according to prespecified rules.

3.3.2 Sampling, Deduplication, and Leakage Control

Exact duplicates, reposts, quoted duplicates, near-duplicates, and template-generated material should be identified before train/validation/test splitting. Where user, thread, conversation, or source identifiers are available, group-aware splitting must prevent the same underlying individual or conversation from appearing in more than one partition. The final report should document deduplication thresholds, exclusion counts, class prevalence, source balance, and any oversampling or class weighting. These controls are essential because leakage can inflate apparent performance without improving generalization.

3.4 Ethical Governance

All primary data collection requires institutional ethics approval and, where applicable, formal data-sharing and data-sovereignty agreements before researchers access identifiable material. The final empirical report must state the approving institution, approval identifier, consent or waiver basis, source-specific permissions, and retention schedule; this protocol does not imply that those approvals have already been granted. Identifying information must be removed before modelling wherever feasible, raw transcripts should be retained only for the minimum approved period, and helpline-sourced material must not be repurposed beyond the approved research purpose. Annotator wellbeing procedures, exposure limits, debriefing, and access to support should be built into the annotation workflow. Demographic attributes such as gender must not be inferred from names, language style, or model predictions; subgroup analysis should use only ethically obtained and sufficiently reliable metadata.

3.5 Annotation Architecture

3.5.1 Semantic Domain Tagging

A bilingual Urdu-to-USAS domain-mapping lexicon should be constructed by trained bilingual annotators who assign USAS semantic-domain codes to high-frequency Urdu lemmas identified in pilot samples, followed by manual verification on held-out material. The mapping should use the UCREL Semantic Analysis System as its reference taxonomy (Rayson et al., 2004). Difficult or culturally specific items should be retained in an adjudication log rather than forced into an ill-fitting category. This procedure permits domain-compatible comparison with Series Paper 2 without requiring an entirely new semantic taxonomy.

3.5.2 Code Switch Annotation

Each token in code-switched material should be labelled as Urdu, Punjabi, English, ambiguous/mixed, or other according to a written codebook. The annotation should follow token-level conventions established in the CALCS shared-task tradition and adapt them explicitly to

ISSN (P): 3006-6522
ISSN (E): 3006-6530

Urdu-English and Punjabi-English material (Solorio et al., 2014; Molina et al., 2016). Romanized forms should be annotated for language identity before any optional normalization so that orthographic variation is not mistaken for code-switching.

3.5.3 Izzat Register

The izzat register will be constructed inductively from the qualitative categories established in Series Paper 1 and refined through pilot annotation. It will include honour-shame framing, relational guilt, family obligation, apology, coercion, forced circumstance, and domestic entrapment. Bilingual coders with relevant cultural and clinical familiarity will annotate pilot material, and the lexicon will be checked against the earlier thematic categories for construct validity. The codebook must include positive examples, counterexamples, boundary cases, and rules for figurative or quoted language so that the category reflects contextual meaning rather than simple keyword presence.

3.6 Inter Annotator Reliability

At least 15% of the corpus should be independently double-coded. Cohen's kappa should be calculated for nominal annotation layers such as code-switch labels and izzat categories (Cohen, 1960), with interpretation informed by established agreement benchmarks while avoiding mechanical reliance on a single cut-off (Landis & Koch, 1977). Series Paper 4 specifies a target of $\kappa \geq .75$ and notes consistency with the $\kappa = .81$ thematic-coding reliability reported in Series Paper 1. If the target is not achieved, the annotation scheme should be revised, coders retrained, and a new reliability sample evaluated before full annotation proceeds.

3.6.1 Annotation Codebook, Training, and Adjudication

Before full annotation, coders should complete a standardized training set and discuss disagreements against the codebook. A pilot reliability round should then be conducted on material not used for codebook examples. Disagreements in double-coded material should be resolved by consensus or by a designated senior adjudicator, and the adjudication decision should be logged without retroactively altering the reliability statistic. The final empirical report should provide the number and background of coders, training duration, reliability sample size, category prevalence, adjudication rate, and procedures used for ambiguous or distressing content.

3.7 Preprocessing

Preprocessing should preserve punctuation, repetitions, discourse markers, emojis, casing where meaningful, and spelling variation when these features may carry pragmatic or stylistic information. Two representations should be retained for Roman Urdu: the original user form and a normalized form created under documented rules. Script identification should distinguish Urdu Nastaliq, Punjabi Shahmukhi, Roman Urdu, English, and mixed-language text. Normalization must be performed only after the raw version has been archived so that robustness can be compared across original and normalized representations.

3.8 Baseline Models

Four baselines should be established: frequency-based linguistic indicators, TF-IDF logistic regression, a linear support-vector machine, and a feature-based machine-learning classifier. The feature set should include pain vocabulary, first-person reference, future orientation, absolutist language, negation, isolation, social affiliation, semantic domains, and culturally specific relational indicators. Baseline preprocessing, vocabulary thresholds, class weighting, regularization, and feature-scaling decisions should be fitted on the training partition only and reported in sufficient detail for replication.

ISSN (P): 3006-6522
ISSN (E): 3006-6530

3.9 Transformer Models

XLM-RoBERTa will serve as the principal multilingual transformer because it was designed for large-scale cross-lingual representation learning and has shown strong transfer performance across many languages (Conneau et al., 2020). Multilingual BERT will provide a second transformer comparison grounded in the BERT fine-tuning paradigm (Devlin et al., 2019). The primary implementations should use publicly identifiable checkpoints (for example, xlm-roberta-base and bert-base-multilingual-cased) with the exact library and model versions recorded in the final report. Where appropriate and legally available, English clinical language models may be included as secondary comparators. All models must be fine-tuned on the training partition only, with validation data reserved for hyperparameter selection, threshold analysis, and early stopping.

3.9.1 Training and Hyperparameter Protocol

The primary fine-tuning search should be prespecified and computationally feasible. A recommended initial grid is learning rates of 1e-5, 2e-5, and 3e-5; batch sizes of 16 or 32 where hardware permits; 3-5 epochs; weight decay of 0.01; and early stopping based on validation macro-F1 with a fixed patience rule. Maximum sequence length should be reported explicitly; if long documents are segmented into windows, the window length, overlap, and document-level aggregation rule must be fixed before test evaluation. Model selection must use validation data only. The final report should state tokenizer versions, software libraries, hardware, random seeds, selected hyperparameters, training time, and whether class weighting or sampling was used.

3.10 Culturally Calibrated Transformer

The proposed full model combines three information streams. Stream A is the contextual transformer representation. Stream B contains general linguistic features such as pronoun use, semantic domains, future orientation, absolutist language, pain, isolation, and negation. Stream C contains culturally situated features such as the izzat register, relational guilt, family obligation, honour-shame framing, coercion, entrapment, metaphor, script, and code-switch information. Structured features will be standardized using training-set statistics and combined with the contextual representation before final classification. The architecture should be described as culturally informed or culturally calibrated only after the empirical analysis demonstrates reliable annotation, incremental value, and stable performance across relevant language/source subgroups.

3.11 Validation Strategy

Four validation regimes are required: (1) a group-aware random stratified split, (2) source-held-out validation, (3) language-held-out validation, and (4) cross-script validation. For the primary random regime, the provisional split is 70% training, 15% validation, and 15% test, stratified by class and language where sample size permits. All segments from the same user, conversation, thread, or original document must remain in a single partition to prevent leakage. Near-duplicate detection should be completed before splitting. Source-held-out validation is particularly important because random document splitting can allow a classifier to learn platform or source characteristics rather than suicide-related linguistic patterns. Final results should be averaged across multiple fixed random seeds, with every seed reported.

3.12 Evaluation Metrics

Performance should be evaluated using precision, recall, F1 score, ROC AUC, precision-recall AUC, confusion matrices, and calibration measures including expected calibration error. Because modern neural networks can be poorly calibrated even when discrimination is strong, calibration should be evaluated explicitly and, where prespecified, temperature scaling may be fitted on the validation set only (Guo et al., 2017). Macro-averaged performance should be reported across languages so that English does not dominate the multilingual result. The default decision

ISSN (P): 3006-6522
ISSN (E): 3006-6530

threshold should be stated in advance; any validation-derived alternative threshold must be reported as a secondary analysis rather than silently substituted after viewing the test set.

3.12.1 Threshold Selection and Statistical Uncertainty

The primary classification threshold should remain 0.50 unless a different rule is prespecified. Secondary operating points may be selected on the validation set to prioritize recall, precision, F1, or a clinically motivated sensitivity target, but they must never be optimized on the test set. Ninety-five percent confidence intervals should be estimated for principal metrics using stratified bootstrap resampling or another documented procedure. Pairwise model comparisons should use paired resampling or an appropriate paired test, and the analysis should report repeated-seed variability so that small numerical differences are not presented as stable improvements.

3.13 Ablation Design

Model A: corpus linguistic lexical baseline.

Model B: transformer representation without structured features.

Model C: transformer plus general linguistic features.

Model D: transformer plus cultural features.

Model E: full transformer plus general linguistic and cultural features.

The principal ablation test is whether Models C, D, and E produce meaningful improvement over Model B across discrimination, calibration, and robustness measures. A failure to improve is also informative: it may indicate that contextual embeddings already capture the relevant signals, that structured features are noisy or redundant, or that the operationalization of cultural categories requires refinement. Ablation conclusions should therefore be based on confidence intervals and repeated-seed comparisons rather than on a single score difference.

3.14 Explainability

SHAP-based feature attribution may be used to examine how structured features contribute to model decisions, following the additive feature-attribution framework proposed by Lundberg and Lee (2017). Exploratory attention visualization may also be reported, but attention weights should not be treated as self-evident explanations of model reasoning because their explanatory status is contested (Jain & Wallace, 2019). Explanations should be interpreted contextually and verified against plausible linguistic constructs. A word associated with death, pain, or despair is not inherently suicidal and must not be treated as a deterministic indicator.

3.15 Ethical Boundary

The model should be described as a research-level indicator of suicide-related discourse rather than as a clinical diagnostic or autonomous suicide-prediction system. False positives can produce stigma, unnecessary escalation, and privacy harms, while false negatives can create unjustified perceptions of safety. Any future applied system must therefore retain a qualified human professional as the final decision-maker, document its intended use, and undergo prospective clinical validation before claims about individual-level risk are made.

4. PRESPECIFIED ANALYSIS AND INTERPRETATION PLAN

This chapter specifies how the completed Series Paper 5 experiments should be analysed and reported. Numerical model results are intentionally absent because the multilingual corpus has not yet been completed. Series Paper 4 explicitly states that data collection remained prospective; accordingly, the analyses below are prespecified procedures and interpretation rules rather than observed findings (Ishfaq & Sultan, 2026b). Any eventual empirical paper should clearly separate these prespecified analyses from exploratory analyses added after inspection of the data.

ISSN (P): 3006-6522
ISSN (E): 3006-6530

4.1 Planned Corpus Composition Reporting

The completed study should report achieved document and token counts by language, script, source, and class, together with exclusions, duplicates removed, missing metadata, and the number of grouped users/conversations. These figures should be compared with the design targets established in Series Paper 4. Any deviation from the proposed sampling frame should be documented and justified rather than silently corrected.

4.2 Planned Linguistic Analysis

The first empirical comparison should replicate the principal variables of Series Paper 2: pain and suffering vocabulary, first-person singular reference, future tense or future orientation, absolutist language, negation, and semantic-domain proportions. Series Paper 4 specifically proposes log-likelihood, Mutual Information, chi-square with Bonferroni correction, and Cohen's d for these variables and for the new izzat-register variable (Ishfaq & Sultan, 2026b). Effect sizes and uncertainty should accompany significance tests so that large samples do not turn trivial differences into substantively exaggerated findings.

4.3 Planned Cultural Analysis

The izzat register should be examined as a culturally specific feature rather than a universal marker. The analysis should determine whether honour-shame language, relational guilt, family obligation, coercion, and entrapment differ between suicide-related and control discourse and whether those patterns vary by language, script, and source. Subgroup analyses involving gender should be conducted only when reliable, ethically obtained metadata are available; missing or uncertain demographic information should be reported rather than inferred.

4.4 Planned Model Performance and Calibration Analysis

The completed study should report the performance of each baseline and transformer model under the same held-out partitions. Discrimination metrics should be accompanied by 95% confidence intervals, repeated-seed variability, and calibration results. Where temperature scaling or another calibration procedure is used, it must be fitted exclusively on validation data (Guo et al., 2017). Error analysis should separate false positives and false negatives by language, script, source, and text length, while avoiding demographic interpretation when reliable metadata are unavailable. Statistical comparisons should focus on the magnitude and stability of differences rather than on a single best score.

4.5 Comparative Interpretation

The central comparison will determine whether contextual multilingual modelling provides an advantage over classical linguistic features and whether structured features add value beyond the transformer representation. A higher F1 score alone should not be considered sufficient evidence. Improvements should be examined alongside calibration, source-held-out performance, cross-linguistic stability, subgroup robustness, confidence intervals, and explainability.

4.6 Cross-Linguistic Performance

Performance should be reported separately for English, Urdu Nastaliq, Punjabi Shahmukhi, Roman Urdu, and Urdu-English code-switched material, with macro averages accompanying overall scores. If English substantially outperforms Urdu or Punjabi, resource imbalance should be considered as one possible explanation, but alternative explanations such as label quality, source composition, script/tokenization effects, class prevalence, and domain mismatch must also be tested before attributing the difference to language resources alone.

4.7 Source-Held-Out Performance

The difference between random-split and source-held-out performance will provide a direct test of domain generalization. A large decline would indicate that performance is sensitive to

ISSN (P): 3006-6522
ISSN (E): 3006-6530

source and may reflect platform- or genre-specific cues rather than transferable suicide-related linguistic patterns. Such a result would not invalidate the model, but it would restrict the scope of claims that can be made about robustness and real-world transfer.

4.8 Ablation Results

The ablation study will test whether structured cultural and general linguistic features contribute information beyond the transformer representation. If the full model outperforms the transformer-only model across multilingual and source-held-out tests, the result would support the PFLSI proposition that structured features provide incremental information. If no improvement is observed, the result may indicate that contextual embeddings already encode those signals, that the structured representations are insufficiently reliable, or that their contribution is context-specific. Competing explanations should be evaluated rather than resolved in advance.

4.9 Explainability Results

Explanation analysis should examine whether important model features correspond to established PFLSI constructs while also identifying unexpected source, script, or formatting cues. Particular attention should be paid to pain and suffering, first-person reference, future depletion, absolutist language, isolation, relational guilt, family obligation, the izzat register, entrapment, metaphor, and code-switching. Explanations should be treated as diagnostic evidence about model behaviour, not as proof of psychological causation.

4.10 Discussion in Relation to Series 1

The first paper established qualitative evidence for apology, regret, self-worth, guilt, family obligation, coercion, and relational discourse (Ishfaq, Ahmad, & Sultan, 2025). Series Paper 5 should determine whether these culturally situated observations survive corpus-scale and computational modelling. The re-digitized Series Paper 1 notes are particularly valuable because they provide a bridge between the original qualitative evidence and the multilingual computational corpus, but their distinctive genre must be controlled so that models do not simply learn the style of suicide notes.

4.11 Discussion in Relation to Series 2

Series Paper 2 established the strongest corpus-linguistic evidence in the programme (Ishfaq & Sultan, 2025c). Series Paper 5 extends those variables into multilingual contextual modelling. The important test is whether pain, first-person reference, future orientation, absolutist language, and related features remain statistically and substantively meaningful outside the original English corpus and after source effects are controlled.

4.12 Discussion in Relation to Series 3

Series Paper 3 proposed a seven-layer PFLS architecture and identified cultural calibration as an essential component (Ishfaq, Sultan, & Bhatti, 2026). Series Paper 5 operationalizes this architectural claim by specifying tests of whether culturally structured information improves model performance, calibration, interpretability, and robustness.

4.13 Discussion in Relation to Series 4

Series Paper 4 identified the monolingual ceiling and proposed the M-SDC as the necessary empirical foundation for cultural calibration (Ishfaq & Sultan, 2026b). Series Paper 5 is therefore the computational continuation of that protocol. The model should not be described as culturally calibrated unless multilingual data, annotation reliability, ablation, and validation experiments demonstrate the relevant properties.

4.14 Interpretation of Hypotheses

Each hypothesis should be classified as supported, partially supported, or not supported using prespecified metrics, confidence intervals, and validation regimes. No hypothesis should be

ISSN (P): 3006-6522
ISSN (E): 3006-6530

accepted merely because a model achieves a high overall score. Cross-source and cross-language validation, calibration, and subgroup stability should form part of the interpretation, and exploratory findings should be labelled as such.

4.15 Overall Discussion

The anticipated theoretical contribution is a model of suicide-related discourse in which broadly observed and culturally situated signals can be tested together. General dimensions may include psychological pain, self-reference, future orientation, absolutist language, and social disconnection. Culturally situated dimensions may include honour-shame framing, relational guilt, family obligation, coercion, and entrapment. The transformer provides contextual representation, while the structured features preserve interpretability and allow direct tests of whether culturally motivated categories add value.

The most important methodological principle is that predictive performance and linguistic validity are separate constructs. A model may classify well for the wrong reasons, including platform, source, demographic, or script artifacts. The combination of source-held-out testing, cross-linguistic validation, ablation, calibration, repeated-seed analysis, and contextual explanation is therefore essential to distinguish transferable linguistic evidence from dataset-specific shortcuts.

5. CONCLUSION AND RECOMMENDATIONS

5.1 Summary of the Study

Series Paper 5 advances the PFLSI programme from corpus and architectural development toward a prespecified protocol for transformer-based computational validation. It builds directly on the sequence established by the preceding papers: culturally situated suicide notes in Series Paper 1, corpus-scale English analysis in Series Paper 2, theoretical and architectural integration in Series Paper 3, and multilingual corpus design in Series Paper 4. The present article defines what must be tested next; it does not claim that those tests have already been completed.

5.2 Principal Contribution

The principal contribution is the specification of a proposed culturally informed multilingual transformer framework and the validation procedures required to determine whether it can legitimately be described as culturally calibrated. The proposed system does not rely exclusively on lexical counts. It combines contextual language representations with linguistic and cultural indicators, including pain, first-person reference, future orientation, absolutist language, relational guilt, family obligation, the izzat register, metaphor, and code-switching.

5.3 Theoretical Contribution

Theoretically, the protocol shows how psychological and linguistic theories can be operationalized within a computational framework without treating the model as a direct measure of psychological state. Psychache, interpersonal processes, hopelessness, motivational and volitional processes, linguistic style, metaphor, and narrative crisis provide complementary theoretical anchors whose contribution can be tested rather than assumed.

5.4 Methodological Contribution

Methodologically, the study argues that suicide-discourse models should be evaluated through more than random train-test accuracy. Group-aware splitting, source-held-out validation, cross-linguistic testing, cross-script robustness, subgroup analysis, calibration, repeated-seed evaluation, uncertainty estimates, and ablation are required to establish whether the model has learned transferable linguistic signals rather than dataset artifacts.

ISSN (P): 3006-6522
ISSN (E): 3006-6530

5.5 Contribution to Urdu and Punjabi NLP

The proposed M-SDC provides a potential foundation for Urdu and Punjabi mental-health NLP. Its semantic-domain mapping, code-switch annotation, Roman Urdu normalization, and culturally specific izzat register may support future work beyond suicide discourse, including depression, anxiety, trauma, crisis communication, and mental-health support. Such extensions would require task-specific labels and validation rather than direct reuse of a suicide-discourse classifier.

5.6 Ethical Implications

Any future model should remain a research and decision-support component rather than an autonomous clinical system. Sensitive crisis data require strict anonymization, purpose limitation, data sovereignty, minimum retention, access controls, and researcher-wellbeing procedures. These requirements are central to the Series Paper 4 protocol and must be converted into a documented, approved governance process before data collection or modelling begins.

5.7 Limitations

- The multilingual corpus remains a design target until collection and annotation are completed.
- Suicide-related discourse is heterogeneous and cannot be equated automatically with suicidal ideation, intent, or prospective individual risk.
- Online, helpline, narrative, and suicide-note sources differ in communicative purpose and may create substantial domain and genre shift.
- Roman Urdu has substantial orthographic variation, creating both normalization and tokenization challenges.
- Punjabi NLP resources may not transfer directly across scripts, dialects, or source domains.
- Cultural annotation remains partly interpretive even when reliability thresholds are achieved; construct validity must be evaluated separately from agreement.
- Transformer explanations do not establish psychological causation, and attention weights alone should not be treated as explanations.
- Prospective clinical validation with clinically anchored outcomes remains necessary before any individual-level clinical application can be considered.

5.8 Recommendations

- Complete the M-SDC only after institutional ethical approval and data-governance agreements are in place.
- Maintain the proposed 500,000-token target where ethically and practically feasible, while documenting all deviations and exclusions.
- Preserve matched control sampling across source, register, language/script, and approximate time period.
- Complete double annotation, document coder training and adjudication, and require $\kappa \geq .75$ before full corpus annotation proceeds.
- Train both classical baselines and multilingual transformers under a fully documented, leakage-controlled protocol.
- Use source-held-out validation as a primary robustness test rather than relying on random splitting alone.
- Report language-specific and macro-averaged performance with confidence intervals and repeated-seed variability.

ISSN (P): 3006-6522
ISSN (E): 3006-6530

- Include calibration, threshold analysis, error analysis, and ethically defensible subgroup analysis in all final model reports.
- Do not describe a model as clinically validated or as predicting suicide without prospective clinical testing and an appropriate clinical reference standard.
- Keep a qualified human professional as the final decision-maker in any future applied system and document the limits of automated inference.

5.9 Future Research

Future research should investigate longitudinal language change, finer-grained crisis stages, dialect-sensitive modelling, parameter-efficient adaptation for low-resource languages, conversational turn-level analysis, and prospective clinical validation. A particularly promising direction is human-model collaboration, in which computational systems present linguistically interpretable evidence and calibrated uncertainty to trained professionals rather than making autonomous decisions. Future work should also compare culturally structured features with purely contextual representations across independent datasets to determine whether any observed gains generalize beyond the PFLSI corpus family.

6. CONCLUSION

The PFLSI programme has progressed from close reading of Pakistani suicide notes to corpus-scale analysis, theoretical architecture, multilingual resource design, and now a protocol for transformer-based validation. The central proposition of Series Paper 5 is that multilingual suicide-discourse detection should not merely recognize more languages; it should be evaluated for whether it represents linguistic and cultural variation without mistaking source artifacts for meaningful crisis signals.

The distinction is consequential. Translation does not guarantee cultural equivalence, and a high-performing classifier is not necessarily a valid, calibrated, or safe system. A responsible psycho-forensic linguistic model must demonstrate predictive performance alongside linguistic validity, cultural calibration, source generalization, interpretability, uncertainty awareness, and ethical governance.

Series Paper 5 therefore does not claim that a transformer can predict suicide. It specifies the methodological conditions under which associations between language and suicide-related discourse can be investigated responsibly and, only after further validation, considered for decision-support research. The long-term objective is not to replace clinical judgement with artificial intelligence but to make linguistically grounded evidence more visible, systematically testable, culturally sensitive, and useful to professionals whose judgement remains indispensable.

REFERENCES

Ahmad, A. I. S. (2026). Language of influence: A corpus-based lexical analysis of psychological power strategies in Robert Greene's The Laws of Human Nature. In Proceedings of the 2nd Riphah International Conference on Language, Literature, and Culture.

Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clinical Psychological Science, 6(4), 529–542.

Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014). I am borrowing ya mixing? An analysis of English-Hindi code mixing in Facebook. In Proceedings of the First Workshop on Computational Approaches to Code Switching (pp. 116–126).

ISSN (P): 3006-6522
ISSN (E): 3006-6530

Beck, A. T. (1963). Thinking and depression: I. Idiosyncratic content and cognitive distortions. Archives of General Psychiatry, 9(4), 324–333.

Beck, A. T., Weissman, A., Lester, D., & Trexler, L. (1974). The measurement of pessimism: The Hopelessness Scale. Journal of Consulting and Clinical Psychology, 42(6), 861–865.

Bialer, A., Izmaylov, D., Segal, A., Tsur, O., Levi-Belz, Y., & Gal, K. (2022). Detecting suicide risk in online counseling services: A study in a low-resource language. In Proceedings of the 29th International Conference on Computational Linguistics (pp. 4241–4250). International Committee on Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (pp. 51–60).

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 3267–3276).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Gkotsis, G., Oellrich, A., Hubbard, T., Dobson, R., Liakata, M., Velupillai, S., & Dutta, R. (2017). Characterisation of mental health conditions in social media using informed deep learning. Scientific Reports, 7, 45141.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321–1330). PMLR.

Ishfaq, A. (2025). Ethnolinguistic identity and cultural memory in the Dawoodi community. Annual Methodological Archive Research Review, 3(6), 185–208.

Ishfaq, A. (2026). Thinking through language: Linguistic foundation and advanced academic writing.

Ishfaq, A., & Bhatti, A. M. (2019). Language shift and imminent language death: A diachronic study of Dawoodi. Harf-o-Sukhan, 3, 1–14.

Ishfaq, A., & Bhatti, A. M. (2020). Lexical attrition and generational language competence in Dawoodi speakers. Harf-o-Sukhan, 4, 4–18.

Ishfaq, A., & Bhatti, A. M. (2021). From pidgin to creole to collapse: The evolutionary trajectory of Dawoodi language. Jahan-e-Tahqeeq, 4.

Ishfaq, A., & Bhatti, A. M. (2022). Linguistic hegemony and the silencing of Dawoodi: Power, stigma, and structural marginalization. Jahan-e-Tahqeeq, 5(4), 48–60.

Ishfaq, A., & Bhatti, A. M. (2023). Code-switching, borrowing, and linguistic dilution: Contact-induced change in Dawoodi. Jahan-e-Tahqeeq, 6(3), 577–592.

ISSN (P): 3006-6522
ISSN (E): 3006-6530

Ishfaq, A., & Bhatti, A. M. (2024). From heritage to liability: Language attitudes and identity reconstruction as drivers of obsolescence in the Dawoodi language. Al-Mahdi Research Journal (MRJ), 5(3), 1303–1336.

Ishfaq, A., & Sultan, N. (2024). Identification of different methodologies for treatment of autism in Urdu-speaking adolescents: An investigative report. Contemporary Journal of Social Science Review, 2(4), 1611–1618.

Ishfaq, A., & Sultan, N. (2025a). Artificial intelligence and human language cognition: An integrative comparative framework. Critical Review of Social Sciences and Humanities, 5(2), 57–71.

Ishfaq, A., & Sultan, N. (2025b). Cognitive control and executive function in advanced second-language writing. Sareer-a-Khama, 4(4).

Ishfaq, A., & Sultan, N. (2025c). Language of stress: A corpus-based study to detect early signs of suicide through lexical choice. Journal of Applied Linguistics and TESOL (JALT), 9(1), 862–880. https://doi.org/10.63878/jalt2310

Ishfaq, A., & Sultan, N. (2025d). Narrative and meaning in Surah Yūsuf: A critical hermeneutic analysis. AL-HAYAT Research Journal (AHRJ), 2(4), 11–23. https://doi.org/10.5281/zenodo.17091179

Ishfaq, A., & Sultan, N. (2025e). Trauma, resilience, and narrative healing: A psycho-hermeneutic reading of Surah Yūsuf. AL-JAMEI Research Journal, 3(1), 229–239.

Ishfaq, A., & Sultan, N. (2026a). Modeling meaning in generative AI: A corpus-assisted discourse analysis of coherence, framing, and persuasion. Pakistan Journal of Social Science Review, 5(3), 1147–1164.

Ishfaq, A., & Sultan, N. (2026b). Voices across scripts: A protocol for a multilingual Urdu–Punjabi suicide discourse corpus and the cultural calibration of psycho-forensic linguistic detection. Critical Review of Social Sciences and Humanities, 6(1), 43–56.

Ishfaq, A., Ahmad, S., & Sultan, N. (2025). Decoding despair: A multidisciplinary psycho-forensic linguistic approach to suicide notes. Journal of Psychology, Health and Social Challenges, 3(2), 83–89.

Ishfaq, A., Azim, M. U. (2025). Phono-semantics and translation: A cross-linguistic study of Urdu and Punjabi ideophones. International Research Journal of Arts, Humanities and Social Sciences, 2(3).

Ishfaq, A., Malik, A. H., & Sultan, N. (2025). Developing trauma-sensitive pedagogical practices for resilient learning in academia: A multidisciplinary approach of psycholinguistics and ELT. Al Aasar, 2(1), 171–189.

Ishfaq, A., Sultan, N., & Bhatti, A. M. (2026). The lexicon of crisis: A psycho-forensic linguistic architecture for real-time suicide risk detection integrating big data, corpus evidence, and theoretical frameworks. Annual Methodological Archive Research Review, 4(6), 1–27.

Ishfaq, A., Sultan, N., & Healy, B. (2025). Turn-taking, politeness, and identity: A conversational study of Speak Your Heart. Journal of Applied Linguistics and TESOL (JALT), 8(3), 1567–1581. https://doi.org/10.63878/jalt1146

Ishfaq, A., Sultan, N., Hassan, N., Aleem, F., & Maldonado, M. G. (2022). Elif Shafak's Forty Rules of Love: Contextual variation in adjectives. Turkish Online Journal of Qualitative Inquiry, 13(1), 2029–2036.

Izmaylov, D., Segal, A., Gal, K., Grimland, M., & Levi-Belz, Y. (2023). Combining psychological theory with language models for suicide risk detection. In Findings of the Association for

ISSN (P): 3006-6522
ISSN (E): 3006-6530

Computational Linguistics: EACL 2023 (pp. 2430–2438). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-eacl.184

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (pp. 3543–3556). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1357

Joiner, T. E. (2005). Why people die by suicide. Harvard University Press.

Khan, I. A., Khaled, F., & Ishfaq, A. (2025). Operation Bunyan un Marsoos: A critical analysis of human rights compliance: A study of the operation's adherence to human rights law and international humanitarian law. Dialogue Social Science Review (DSSR), 3(6), 173–184.

Khattak, A., Asghar, M. Z., Saeed, A., Hameed, I. A., Ahmad, S. A., & Hassan, M. (2021). A survey on sentiment analysis in Urdu: A resource-poor language. Egyptian Informatics Journal, 22(1), 53–74.

Kövecses, Z. (2015). Where metaphors come from: Reconsidering context in metaphor. Oxford University Press.

Lakoff, G., & Johnson, M. (1980). Metaphors we live by. University of Chicago Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30.

Menon, R. (2024). The narrative-crisis model: Understanding suicidal ideation through personal storytelling. Journal of Mental Health and Narratives, 12(3), 193–208.

Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., & Solorio, T. (2016). Overview for the second shared task on language identification in code-switched data. In Proceedings of the Second Workshop on Computational Approaches to Code Switching (pp. 40–49).

O'Connor, R. C. (2011). Towards an integrated motivational-volitional model of suicidal behaviour. In R. C. O'Connor, S. Platt, & S. Gordon (Eds.), International handbook of suicide prevention (pp. 181–198). Wiley-Blackwell.

Pennebaker, J. W. (2011). The secret life of pronouns: What our words say about us. Bloomsbury Press.

Rayson, P., Archer, D., Piao, S., & McEnery, T. (2004). The UCREL Semantic Analysis System. In Proceedings of the LREC Workshop on Beyond Named Entity Recognition.

Shneidman, E. S. (1993). Suicide as psychache: A clinical approach to self-destructive behavior. Jason Aronson.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., & Fung, P. (2014). Overview for the first shared task on language identification in code-switched data. In Proceedings of the First Workshop on Computational Approaches to Code Switching (pp. 62–72).

Sultan, N., & Ishfaq, A. (2026). Instagram captions as identity performance: A multimodal discourse analysis. AL-HAYAT Research Journal (AHRJ), 3(2), 9–21.

Teddlie, C., & Yu, F. (2007). Mixed methods sampling: A typology with examples. Journal of Mixed Methods Research, 1(1), 77–100.

World Health Organization. (2023). Suicide worldwide in 2019: Global health estimates. WHO Press.