

Výzvy pre monitoring médií za účelom analýzy zmien v zahraničnej politike

Challenges for Media Monitoring for Foreign Policy Analyses

Andrej Školkay

with contributions by **Konstantin Startsev**

The School of Communication and Media

Slovakia

Project code: 09I03-03-V04-00308, Call: 09I03-03-V04 Fellowships for excellent researchers R2-R4.

Acronym: EXPERIMENTAL DETECTOR. This project is funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00308.

Suggested Citation: Školkay, A. with contribution by K. Startsev (2026). *Výzvy pre monitoring médií za účelom analýzy zmien v zahraničnej politike* [Challenges for media monitoring for foreign policy analyses] (Project No. 09I03-03-V04-00308, EXPERIMENTAL DETECTOR). School of Communication and Media.

**Experimental Design,
Methodology and Outputs for
Detecting Foreign Policy Discourses**





ABSTRAKT - SLOVENSKÁ VERZIA

Táto správa systematicky zhrňuje a kriticky vyhodnocuje zistenia z predchádzajúcich troch čiastkových štúdií v rámci projektu EXPERIMENTAL DETECTOR: po prvé, Critical Review of Methodologies (Monitoring) task, publikovanej pod názvom *Analysing Media Analytics Tools for Foreign Policy Monitoring*; po druhé, Critical Review of Methodologies (Analyses) task, publikovanej pod názvom *Identification of Methodological Challenges in International Relations Studies: A Case Study of the 2008 Russo-Georgian War*; a po tretie, *Case Study: Framing the Russo-Ukrainian War. A Comparative Review of Media Coverage and Analytical Discourse. Methodological Lessons and Advances*. Správa je rozdelená do troch hlavných častí, ktoré majú jednotnú štruktúru: cieľ a prístup, kľúčové zistenia a odporúčania vyplývajúce z danej úlohy. Záverečná časť následne prináša všeobecnú analytickú syntézu s dôrazom na metodologické ponaučenia a na výskumné a politické odporúčania.

ABSTRACT - ENGLISH VERSION

This report systematically summarizes and critically evaluates the findings from the previous three partial studies within the EXPERIMENTAL DETECTOR project: first, the Critical Review of Methodologies (Monitoring) task, published under the title *Analysing Media Analytics Tools for Foreign Policy Monitoring*; second, the Critical Review of Methodologies (Analyses) task, published under the title *Identification of Methodological Challenges in International Relations Studies: A Case Study of the 2008 Russo-Georgian War*; and third, the *Case Study: Framing the Russo-Ukrainian War. A Comparative Review of Media Coverage and Analytical Discourse. Methodological Lessons and Advances*. The report is divided into three main sections, each with a uniform structure: objective and approach, key findings, and recommendations arising from the respective task. The final section then provides a general analytical synthesis, with an emphasis on methodological lessons and on research and policy recommendations.



Contents

I. Task 1: Analysing Media Analytics Tools for Foreign Policy Monitoring	1
Aim and Approach	1
Key Findings	1
Recommendations Arising from the Task	2
II. Task 2: Identification of Methodological Challenges in International Relations Studies — A Case Study of the 2008 Russo-Georgian War	2
Aim and Approach	2
Key Findings	3
Recommendations Arising from the Task	4
III. Task 3: Case Study — Framing the Russo-Ukrainian War: A Comparative Review of Media Coverage and Analytical Discourse	6
Aim and Approach	6
Key Findings	6
Recommendations Arising from the Task	7
IV. General Analytical Summary and Conclusions	8
Methodological Lessons Learned	8
Research Recommendations	14
Policy Recommendations	15
ATTACHMENTS	17
Attachment 1) An interview with analyst and former diplomat Balázs Jarábik, published on Postoj.sk, 18 June 2026	17
Attachment 2) Speech by the Minister of Foreign Affairs of the Russian Federation Sergey Lavrov in 2025	17
Attachment 3) A Pilot Experiment for Analysing Foreign Policies Changes Based on Media Reporting	19
REFERENCES	33

I. Task 1: Analysing Media Analytics Tools for Foreign Policy Monitoring

Aim and Approach

The first exploratory study examined media coverage and media monitoring of foreign policy, foreign affairs, and international relations, with the aim of identifying the methodological challenges associated with using media monitoring tools to detect emerging foreign policy. The research combined a systematic overview of past and recent research projects with practical experiments and direct communication with the creators and managers of monitoring tools in the public and private sectors. The research output exceeded 120 pages, included 20 appendices, and drew on more than 130 individual sources.

Key Findings

The study found that monitoring of media content in general — including a number of past and recent research projects in the areas of foreign affairs and international relations — had been widespread, but that no successful example of monitoring media content specifically for the purpose of revealing emerging foreign policy could be identified. Consequently, the research could not, at that stage, directly answer the question of which methodological challenges were associated with using media monitoring tools to identify emerging foreign policy. Nonetheless, it revealed inconsistencies in algorithmic dependability, coding biases, and problems of data collection and interpretation in existing media monitoring tools.

More specifically, despite considerable investment of financial resources and research effort into creating effective legacy media content monitoring tools — and, to a lesser extent, into identifying appropriate legacy media sources in the USA, the EU, and countries such as Switzerland — a truly reliable and current monitoring tool, especially for foreign policy, remained unavailable.

Worryingly, a few tools and research initiatives identified as directly or indirectly, ad hoc or permanently funded by the EU ultimately appeared to have ended in limbo. It remained questionable how scientifically or practically relevant their outputs — usually assessed as delivering excellent results — could be considered.

At the same time, the international situation had deteriorated in recent years, with low prospects of returning to the previous status quo. Careful monitoring of emerging and changing foreign policies in legacy media globally therefore appeared to be an urgent task, capable of usefully supplementing other available sources of analysis and information. The study argued that EU Member States should use the synergy of their efforts precisely in this way.

AI and machine learning (ML), despite the surrounding hype, could not solve this issue alone. It remained an open question whether the most recent large reasoning models (LRMs), specially trained to perform certain types of reasoning, could help. Instead, the study identified as a solution carefully selected and trained intelligence-type analysts, supported by AI- and ML-based monitoring tools, with researchers in the background providing conceptual materials and critical feedback. Cooperation with the private sector was seen as potentially providing a critical advantage and access to cutting-edge experience.

The key methodological lesson from this pilot study was that the only reliable way to assess the effectiveness of media monitoring tools in any field was to assign identical, specific tasks to their

creators or managers and then compare the results. Arriving at such a seemingly simple finding, however, first required producing a comprehensive overview of the kind undertaken here. Notably, the most that could be obtained from the contacted private and public project managers, leaders, and tools — through experiments and communication — were outputs based on sentiment analysis or frequency of mentions. This was clearly insufficient for comprehensive media monitoring of emerging foreign policies. Often, there was silence, or contradictory information was provided by researchers from publicly funded projects. The study therefore underscored the need for advancements in foreign policy monitoring and analysis, while remaining sceptical about the effectiveness of automated monitoring, analytical, and detection tools (with the exception of some partial tasks, such as automated abstract production) without substantial human oversight.

Recommendations Arising from the Task

On the basis of these findings, the study recommended:

Establishing a permanent interdisciplinary virtual or cross-border team, tasked with coordinating the development of comprehensive research and policy strategies for detecting new foreign policy trends in legacy media, and with integrating AI-powered and ML analysis tools, including large reasoning models (LRMs), with human expertise.

Prior to that, carrying out a thorough review of previously funded research projects in the field (largely identified within this task), with a specific focus on learning from failures and mistakes, and on examining the deeper reasons for the lack of continuation and further development of these projects and initiatives.

Enhancing current systems for detecting bias and identifying quality news reporting and scientific monitor of changes in news selection and coverage trends.

Establishing a dedicated European media monitoring system for foreign policy coverage and analysis. Strengthening collaboration within the EU and beyond, and improving interactions and consultations among key stakeholders in foreign policy, including: a shared resource database; efforts to improve, or at least to test, the professional competences of journalists in evaluating and interpreting often contradictory and incomplete information; regular online and offline forums for selected participants; rapid-response communication strategies to address unfolding international events; a strategic approach to exclusive information sharing with specific news outlets; and mechanisms to address coverage gaps concerning important niche events and localities.

II. Task 2: Identification of Methodological Challenges in International Relations Studies — A Case Study of the 2008 Russo-Georgian War

Aim and Approach

The second study explored a relatively minor international event, yet one of fundamental importance for international relations and one seen as having a long-term impact, shaping even the understanding of the current Russo-Ukrainian war. The research project exceeded 170 pages and drew on more than 250 individual sources, applying a structured comparative review of scholarly literature, media coverage, and legal-political documentation. Only selected findings of immediate importance were carried forward into this final report.

Key Findings

Among the more universal findings, the study highlighted the problematic contribution of authors with military backgrounds to scholarly discourse, noting that military training, education, and experience generally prioritised operational and strategic competencies over critical analytical frameworks. Similarly, military-defence organisations were found to share a specific culture nurturing priorities that may contradict the very purpose for which those organisations were established.

A secondary, yet highly relevant, finding was that the Russian Federation's foreign policies had evolved towards aggression only reluctantly. This thesis challenged the frequent assumption that Russia had always been an aggressive, neo-imperialist state — an erroneous assumption that appeared to prevail particularly in Poland and the Baltic states.

The central academic argument posited that epistemic bias (deeply rooted prejudices), factual ambiguity, specific theoretical perspectives, and evidence selectivity fundamentally compromised both the validity and the neutrality of IR analyses, especially in research addressing politically contentious conflicts.

With respect to media, the study found that the coverage and interpretation of the 2008 Russo-Georgian War in Western — more specifically, U.S. and European — media were largely shaped by national stereotypes, political leverage, and the interplay between political actors and journalists, as well as by structural features of the media, including hegemonisation and instrumentalisation. The misinterpretation of Russia as the initial aggressor in the 2008 war — despite evidence that Georgia had initiated a long-planned military operation aimed at the enforced integration of a separatist province, which was at that time under a specific legal regime and with peacekeeping forces on the ground, while Russia initially acted within its peacekeeping mandate — shaped Western perceptions in ways that continued to influence political, journalistic, and analytical discourse. This long-term narrative persisted even though Russia's 2022 invasion of Ukraine constituted a fundamentally different case of clear aggression.

Media coverage in the West, and parts of the Western academic literature following the 2008 conflict, contributed to portraying Russia as expansionist and unpredictable; in some cases, Russia's strong follow-up reaction was highlighted. However, at that time Russian foreign policy had not yet embraced the assertive military posture observed later. Misinterpretations stemming from the 2008 conflict endured in policy debates, journalism, and some academic works, contributing to skewed perceptions. The persistence of biased or inaccurate claims in older and even contemporary publications highlighted weaknesses in peer review and academic self-correction. Recognising these factual inaccuracies in no way negated the reality of Russia's aggression in Ukraine since 2022, or of possible hybrid activities elsewhere; however, persistent distortions concerning the 2008 conflict were found to influence rhetorical framings and, occasionally, foreign policy decisions in parts of Europe.

The study acknowledged its limitations, which included the restricted number of case studies, the limited response rate among journalists specialising in IR to the online questionnaire, the absence of replies from many foreign affairs ministries of EU Member States, the keywords used in selecting the research sample, and the reliance on a single academic repository. It was also noted that the original plan had envisioned a study of approximately 50–60 pages with a more modest scope. Several findings presented challenges with respect to scalability or transferability into broader policy frameworks, educational curricula, and national or international governance structures. Some

insights also appeared temporally constrained — knowledge that would have been of greater strategic or explanatory value had it circulated earlier — and certain arguments necessarily adopted perspectives that could be viewed as revisionist in epistemological terms. Nevertheless, the author did not regard the paradigm-challenging nature of these findings as problematic: although exploratory, the analysis was grounded in robust empirical evidence and methodologically rigorous inquiry, offering a defensible basis for questioning established assumptions and advancing alternative analytical perspectives. This was not presented as a claim of infallibility — minor methodological limitations or marginal errors may have persisted — yet, as widely acknowledged in the scholarly literature, exploratory studies, despite their inherent constraints, can generate empirically solid and theoretically significant contributions.

Recommendations Arising from the Task

The study recommended adopting critically reflexive approaches to analysing global news production, with explicit attention to narrative construction mechanisms and the systematic inclusion of devil's advocate and non-Western perspectives in international relations research. It further advocated that decision-makers prioritise a nuanced understanding of international actors' motivations over binary or distorted assessments, recognising that accurate comprehension — not endorsement — is crucial for strategic engagement. To overcome the significant obstacles posed by institutional distrust, geopolitical instability, and entrenched cognitive and personal biases, the study identified three prerequisites: systematic training in perspective-taking, in understanding theory's indirect impacts, and in recognising personal, professional, or national-level biases; institutional incentives favouring complexity over simplification, especially regarding factual and contextual saturation; and a long-term commitment to analytical rigour regardless of peer pressure, dominant IR theory, media reporting, or governmental rhetoric.

The study's detailed recommendations were grouped into four thematic clusters:

Avoiding governmental rhetoric, foreign policy formulations, or media coverage that may recast peacekeepers as aggressors. The study recommended preventing mischaracterisations in conflict narratives and, given the biased media representation of the 2008 conflict, developing more sophisticated systems to monitor emerging foreign policy trends and conflict dynamics. It acknowledged that absolute accuracy in media reporting during crises was unattainable, but argued that methodological and technological innovations could help mitigate systematic distortions. A clear legal, political, and analytical distinction should be drawn between aggression *per se* and excessive intervention by peacekeeping forces in response to aggression (or military attack).

Exercising caution when approaching studies in international relations. To reduce bias in the analysis of contentious IR issues, the study recommended that practitioners consult non-Western sources and employ *pentagulation*: verification through at least five independent assessors or theoretical–empirical perspectives or sources drawn from diverse or neutral contexts across Europe, Asia, and Latin America — extending beyond conventional triangulation, which may insufficiently capture divergent interpretations.

For example, in an interview published on Postoj.sk on 18 June 2026, the analyst and former diplomat Balázs Jarábik advanced a critical assessment of European strategic capacity vis-à-vis the war's resolution. Jarábik, who from 2019 to 2023 headed the policy, analysis and coordination unit of the EU Advisory Mission (EUAM) in Ukraine and had previously served as an adviser to the Ukrainian government on EU integration, identified three interrelated deficits in European policy: the absence of a substantive strategic debate concerning the terms on which the war might conclude

and the corresponding future architecture of EU–Russia relations; the lack, at both EU and member-state level, of an articulated strategy specifying the means and objectives through which the conflict is to be brought to an end; and, notwithstanding recent Ukrainian territorial gains, the assessment that Russia retains its strategic objectives in Donbas and shows no indication of relinquishing them.

Analytically, Jarábik disaggregated the conflict into three distinct dimensions, or "layers": a war of narratives, in which he assesses Ukraine to currently hold the advantage; a war of attrition, in which the balance of advantage at the front is judged to favor Russia; and a war of adaptability, in which Ukraine is characterized as performing comparatively well, a performance attributed in significant part to sustained Western material and political support (see Attachment 1).

Another example highlighting often downplayed or ignored perspectives is Russian Foreign Minister Sergey Lavrov's justification of the war primarily in security terms, centering Russia's demands on preventing Ukraine's NATO accession and further alliance enlargement, which he framed as a broken Western promise. He invoked Ukraine's 1991 non-aligned Declaration of Independence as the condition under which its territorial integrity had originally been recognized, implying this condition had lapsed. Notably, Lavrov explicitly minimized territorial gain as a war aim, framing the operation instead as protecting populations he described as culturally and linguistically Russian — a claim worth scrutinizing given the concurrent occupation and annexation of Ukrainian territory (see Attachment 2).

Where possible, drafts or published papers should undergo review by experts from neutral countries (e.g., Austria, Switzerland) and non-Western regions to counterbalance prevailing narratives. The study further recommended approaching cautiously, on contested issues, publications authored by military personnel, by those with military-oriented education, or appearing in military-affiliated outlets — particularly in highly sensitive cases such as Gaza or the West Bank in 2025, where institutional cultures and specific education may affect analytical independence.

Raising the standards of IR research. The study recommended that IR research explicitly state its theoretical framework — be it neorealism, neoliberalism, constructivism, or another lens — since many IR articles lacked a methodology section altogether, leaving readers to infer underlying assumptions and risking the uncritical reproduction of mainstream or regional biases, particularly in contested cases such as the 2008 Russo-Georgian War. Even a single sentence situating the analysis within a recognised framework was found to enhance transparency significantly. In this context, the study observed that instead of a liberal international order, the world had moved into a realist international order, or a variant thereof — a fundamental transformation reflected in many key challenges and conflicts of the past decade and one that became fully transparent in 2025. The study also recommended: incorporating perspectives from all principal actors and clearly articulating assessments of initial triggers and aggressors (a basic journalistic standard often absent in IR scholarship — perspectives from South Ossetian actors and from the Russian peacekeeping contingent were notably absent in many of the studies examined); introducing a 'contested findings' label in academic databases such as JSTOR, which often retained older works containing inaccuracies without contextual annotation; and subjecting peer review procedures themselves — both internal and external — to independent assessment, with greater linguistic, geographic, and disciplinary diversity among reviewers, the ultimate goal being the inclusion of a devil's advocate perspective.

Including core analytical competencies in IR education. Based on the examined literature, and acknowledging that some elements may already be implemented in certain institutions, the study recommended that IR students: acquire a nuanced understanding of international law, particularly

regarding the limited role treaties play in state recognition processes, since political, strategic, and diplomatic factors often outweigh legal ones — as illustrated by the recognition of Palestine by several states in 2025, and by Israel’s recognition of Somaliland in 2025, becoming the first and only state to do so after almost three decades of Somaliland’s de facto independence; become familiar with politico-military strategy — for example, recognising that the 2008 Russo-Georgian War involved substantial military preparation by both sides, as evidenced by Georgia’s pre-war military build-up; and develop media literacy for IR, understanding how journalism operates, including agenda-setting, framing, and priming, while avoiding extremes — neither assuming media neutrality nor embracing conspiracy-like narratives of collusion.

III. Task 3: Case Study — Framing the Russo-Ukrainian War: A Comparative Review of Media Coverage and Analytical Discourse

Aim and Approach

The third study comparatively reviewed media coverage and analytical discourse on the framing of the Russo-Ukrainian war, with the explicit purpose of drawing methodological lessons for the anticipatory monitoring of subtle discourse shifts. The review confronted computational approaches (topic modelling, integrated pipelines) with human-coded framing analysis and audience-side data, and included a micro-pilot study testing AI-supported analysis of nuanced discourses at the local media level.

Key Findings

Three cross-cutting methodological conclusions followed from the review. First, no single methodological family was sufficient: the most credible findings in the reviewed literature emerged where computational scale was validated by human coding, or where media-side data were confronted with audience-side data. Second, methodological choice was not neutral with respect to substantive conclusions — Verbytska (2024) demonstrated that two topic models applied to the same corpus prioritised different frames, a sobering result for any monitoring system built on a single model. Third, for the project’s purpose of anticipatory monitoring of subtle discourse shifts, the integrated-pipeline family offered the most promising template, but only if its identified weaknesses — peer-review deficits, monolingual bias, and unvalidated engagement metrics — were corrected by design: through multilingual corpora, human validation samples, and triangulation with survey data.

At the theoretical level, Ye and Aron (2025, pp. 36–37) proposed an updated model of global media influence, termed Alliance-Linked Narrative Dynamics (ALND). The model suggested that narratives were conditioned by three interacting forces: (1) domestic media institutions setting the baseline style and constraints; (2) alliance ties synchronising narrative production across borders; and (3) platform-level dynamics diffusing bloc-consistent frames rapidly across systems, with structural asymmetries (resources, wire services, language hubs) further amplifying some narratives while muting others. What was new in this model was not that agendas converge — agenda-setting, indexing, and cascading activation had already shown that — but that convergence and divergence mapped systematically onto alliance blocs. Alignment appeared when events created shared salience, while divergence appeared in sub-angles shaped by bloc-specific political economies and institutional routines. Engagement added a crucial layer: publics responded most strongly when

coverage matched their bloc's narrative culture, highlighting the co-production of alignment between media and audiences. The framework thereby extended prior accounts of media imperialism and news flow, and complemented network agenda-setting models by clarifying that networks are not flat but bloc-structured.

Taken together, the review supported five overall conclusions. First, the framing of the Russo-Ukrainian war was structured less by individual national idiosyncrasies than by alliance-bloc membership, with domestication as the dominant mechanism in non-aligned systems — a finding that any monitoring methodology must build in as a stratification variable. Second, the dominant Western causal frame (Russian imperial aggression) and the marginalised co-responsibility frame (NATO expansion) coexisted in nearly all Western media systems, but at ratios so unequal that the diversity requirement of impartial coverage was arguably not met; conversely, the co-responsibility literature itself required critical handling. Third, vocabulary functioned as a leading indicator of framing shifts — the migration from 'crisis' to 'war' to 'invasion', or from 'special military operation' to civilisational war — which made lexical monitoring one of the most tractable early-warning signals for the anticipatory purposes of the project. Fourth, media narratives and public opinion were only loosely coupled, so any monitoring system restricted to the supply side would systematically misestimate discourse shifts; engagement and survey data therefore had to be integrated by design. Fifth, setting aside war-induced ideological-political divisions, the selection between the two main explanatory frameworks identified by Zollmann (2024) was not purely empirical: it was ontologically grounded — whether through a realist or a liberal IR lens — and, consciously or not, this worldview ultimately shaped geopolitical interpretation.

Methodologically, the way forward indicated by the strongest reviewed studies was a hybrid architecture: multilingual computational pipelines for scale and speed, validated by human-coded framing analysis for conceptual accuracy, and triangulated with events data and official-statement corpora — precisely the kind of AI-facilitated, real-time, globally comparative monitoring that the project's original aims anticipated and that recent advances had brought within reach. The micro-pilot study (Attachment 3) indicated that AI-supported tools enabled precise and rapid analysis of nuanced discourses at the local media level. Media in authoritarian regimes, in particular, strongly reflected official government stances; insights or cross-checking from local experts — or, initially, from the European diaspora — could strengthen the validity and reliability of media monitoring in countries such as Iran. No new core staff appeared necessary; experts across the EU (some cited in the study) could be pooled and could work remotely.

Regarding tools, most of this output was produced using US-based Anthropic AI tools (Claude; the best appeared to be Fable 5) and the Chinese-based DeepSeek. Using a Europe-based AI such as Mistral (Le Chat) or Lumo would avoid many concerns associated with foreign-hosted proprietary models — such as data residency, cloud compliance, or vendor lock-in — yet Mistral was not, at the time of writing, the best AI tool, or even among the top performers, on demanding tasks.

Finally, the findings cautioned academics, governments, and their contractors against relying uncritically on press-report datasets, as reporting probabilities varied by event type, location, and time — potentially reflecting media biases rather than actual phenomena. Nonetheless, the findings suggested that, in light of both geopolitical challenges and AI progress, the timing was right to introduce a novel method for monitoring foreign policy discourses.

Recommendations Arising from the Task

The study recommended, as the next research step, a comparison of how approximately twenty

current core AI tools handle identical tasks in monitoring emerging foreign policies. This comparison was to build on the micro-pilot study conducted within the research, drawing on insights gained throughout the project, especially from Task 1. It was proposed that the pilot study be carried out in two phases — or two distinct periods — to allow both the identification of potential errors and the monitoring of AI advancements over time. The task was estimated to be realisable at limited cost (up to EUR 35,000, including the purchase of access to AI tools) and rather quickly (five person-months).

IV. General Analytical Summary and Conclusions

Methodological Lessons Learned

Across the three tasks, several convergent methodological lessons emerged. First, the project repeatedly confirmed a gap between claimed and demonstrable performance: neither commercial nor publicly funded monitoring tools could demonstrate the capacity to detect emerging foreign policy, and outputs officially assessed as excellent frequently proved unverifiable or reducible to sentiment analysis and mention counts (Task 1). The most robust way of testing such tools proved deceptively simple — assigning identical tasks and comparing the results — yet reaching that conclusion required a systematic overview that had not previously existed.

Second, the project demonstrated that methodological and epistemic choices were never neutral. In IR scholarship, epistemic bias, factual ambiguity, theoretical perspective, and evidence selectivity fundamentally compromised the validity and neutrality of analyses of contested conflicts (Task 2); in computational media research, even the choice between two topic models applied to the same corpus produced different substantive priorities (Task 3). In both domains, the underlying ontological lens — realist or liberal — shaped interpretation, whether acknowledged or not.

Third, all three tasks converged on the insufficiency of any single method or actor. Reliable findings emerged only where computational scale was validated by human coding, where media-side data were confronted with audience-side data, and where diverse — including non-Western and devil's advocate — perspectives were systematically included. Human oversight, in short, remained indispensable, while AI tools evolved from objects of scepticism (Task 1) into demonstrably useful, though still supervision-dependent, instruments (Task 3) over the course of the project itself.

Fourth, the project documented the long institutional half-life of early misinterpretations. Distorted framings of the 2008 Russo-Georgian War persisted in journalism, policy debates, and academic literature for well over a decade, exposing weaknesses in peer review and academic self-correction; analogously, lexical shifts in the coverage of the Russo-Ukrainian war proved to be leading indicators of framing changes. Both findings pointed to the same conclusion: early, careful, and methodologically transparent analysis of conflict narratives (and vice-versa) has consequences that extend far beyond the immediate event.

Fifth, the project proposed **a shift from retrospective evaluation to prospective monitoring**. The quality criteria developed for assessing completed framing analyses do not transfer to a monitoring instrument designed to detect emerging shifts in foreign-policy discourse without modification. Certain criteria carry over unchanged; others are effectively inverted; and a number of requirements with no counterpart in retrospective scholarship become decisive in a monitoring context. The tentative table below maps each retrospective criterion onto its monitoring analogue and

characterizes the relationship between them, so that a monitoring design is not mistaken for a methodologically weaker variant of a conventional study. The relationship is coded using three values: *carries* (the criterion transfers, in some cases in extended form), *inverts* (the monitoring requirement runs counter to the retrospective one), and *new* (no retrospective counterpart exists).

Table 1. Retrospective criteria and their monitoring analogues

Retrospective criterion	Monitoring analogue	Relation
1. Theoretical grounding		
The study declares which conception of the frame it employs and applies it consistently.	The instrument declares the same conception, and additionally specifies the sub-frame observables through which movement in that frame is expected to become visible first.	Carries (extended)
Frames must not be conflated with topics, themes, or sentiment.	The same prohibition holds, and a further one is added: an indicator is not a frame. A shift in designation lexicon is evidence of frame movement, not the movement itself.	Carries (extended)
Generic and issue-specific frames are distinguished, with consequences for comparability.	Both layers are required simultaneously: generic indicators supply stable baselines and cross-case comparability; issue-specific indicators supply sensitivity to the case at hand.	Carries
Frame-building and frame-setting are theorized as process (Scheufele, 1999).	Cascading activation (Entman, 2003) supplies a discriminant rather than a background theory: an authentic repositioning should display origination at elite level and downward propagation. Simultaneous or bottom-up appearance is treated as probable noise.	Carries (operationalized)
2. Design and sampling		
Constructed-week sampling is the efficient standard for estimating period parameters (Riffe et al., 1993).	Census capture of a deliberately narrow outlet set. Constructed-week sampling discards days by design; the signal may be a two-day anomaly.	Inverts

Retrospective criterion	Monitoring analogue	Relation
Population, period, outlets, and retrieval terms are defined for a single closed corpus.	A frozen burn-in baseline period is defined, followed by an open-ended rolling window, together with a stated policy governing when and how the baseline may be revised. (<i>a pre-specified, immutable opening period whose data establish the reference values for all subsequent measurement, and which is deliberately excluded from hypothesis testing</i>). Two practical implications worth checking in whatever document this comes from: (1) it should state <i>how long</i> the burn-in runs and on what basis that length was chosen (convergence criterion, inter-coder reliability threshold, fixed calendar span); (2) it should state what happens if the system materially changes mid-study — normally that requires declaring a new baseline and reporting both, not silently amending the old one.	Carries (temporalized)
Outlets are selected for reach, ownership diversity, and ideological range.	Outlets are selected for baseline stability and cascade position. Ritualized or state-affiliated discourse supports anomaly detection against a tight baseline; plural systems support dispersion measures across outlets.	Inverts
The unit of analysis is declared (article, paragraph, proposition, image).	Two units are declared: a measurement unit below the frame (clause, designation, source mention) and a decision unit at which alerts are raised (typically outlet-week).	Carries (extended)
The corpus must be reconstructable by a third party.	Corpus composition is itself tracked as a variable — outlet closures and launches, paywall changes, editorial turnover, platform distribution — and observed change is decomposed into within-outlet change and compositional change.	New
Translation decisions are disclosed as framing operations.	Translation is not performed prior to coding. Parallel language-specific indicator sets are anchored to a common construct, and measurement equivalence across languages is tested rather than assumed.	Inverts
3. Measurement quality		
Krippendorff's $\alpha \geq .80$ is established once on a double-coded subsample (Krippendorff, 2004).	Reliability is re-established periodically on a frozen calibration set, and coder drift is reported as a time series rather than certified at a single point.	Carries (temporalized)
Reliability is reported per frame rather than as an aggregate.	Reliability is reported per indicator <i>and</i> per period, since a stable overall coefficient can conceal a category that degraded mid-series.	Carries (extended)

Retrospective criterion	Monitoring analogue	Relation
Validity is established by deriving the operationalization from framing theory (Matthes & Kohring, 2008).	Construct validity remains necessary but is no longer sufficient. Criterion validity against an independent policy series — assembly voting, sanctions decisions, treaty and memorandum signature, arms-transfer authorizations, diplomatic personnel movements — becomes the governing test.	Carries (raised)
—	Backtesting over historical periods containing known repositionings, reporting hit rate, false-alarm rate, and detection lead time.	New
—	False-discovery-rate control across the indicator $\times$ outlet $\times$ period grid (Benjamini & Hochberg, 1995). A dashboard performing thousands of implicit tests annually will otherwise detect shifts continuously.	New
—	Specificity is prioritized over sensitivity, with the alert threshold justified against the assumed base rate of genuine repositioning.	New
4. Computational procedures		
Preprocessing is documented, and framing-bearing lexical variants are preserved.	Preprocessing is additionally frozen for the life of the series; any alteration forces a documented discontinuity and a parallel run across an overlap window.	Carries (extended)
Topics are not to be equated with frames (DiMaggio et al., 2013; Walter & Ophir, 2019).	The same holds, with a stronger consequence: unsupervised topic solutions are unstable across re-estimation and are therefore unsuitable as the primary detector. They remain useful for exploration and codebook revision.	Carries (strengthened)
Model name, version, and access date are reported for reproducibility.	Version pinning becomes a continuity requirement rather than a citation courtesy. Silent vendor updates break a time series without producing any error signal; a frozen calibration set is re-run at every version change, and any resulting offset is reported as an artefact.	Inverts in status
Classifier performance is validated against a manually coded gold standard, with precision, recall, and F1 reported per frame (Grimmer & Stewart, 2013).	Two gold standards are maintained: a frozen set used to detect model and coder drift, and a periodically refreshed set used to measure current performance under concept drift.	Carries (extended)

Retrospective criterion	Monitoring analogue	Relation
Sentiment is a framing device, not a frame.	Emotional-trajectory data are admitted as a covariate and as context, but never as an alert trigger.	Carries
5. Interpretation and context		
The study explains why a given frame emerged when it did.	The instrument distinguishes exogenous event-driven spikes, genuine drift, and compositional change, maintaining an event register as a control series.	Carries (recast)
Indexing theory predicts that the frame range tracks the range of elite debate (Bennett, 1990).	The prediction becomes a live indicator: contraction or expansion of the quoted elite range is among the earliest observable traces of consensus movement.	Carries (operationalized)
Source structure is analysed — who speaks, who is paraphrased, who is absent.	Anonymous-sourcing density is tracked as a leading indicator, on the reasoning that trial balloons precede declared positions.	Carries (extended)
Silences are analysed against a counterfactual standard of what could have been reported.	Silence detection requires an explicit model of the expected utterance, and is tractable only where the baseline is formulaic. The disappearance of a standing qualifying clause is treated as a first-class signal.	Carries (formalized)
—	Lead–lag estimation against the policy series. If the discourse indicator does not precede the policy indicator, the instrument is detecting published news rather than anticipating movement.	New
6. Reflexivity, ethics, and output		
Limitations are acknowledged rather than minimized.	A running record of false alarms and missed events is published with each release. An instrument that does not publish its misses is not an instrument.	Carries (formalized)
The analysis maintains neutral language and does not adopt the frames under study.	The same, with the addition of analytic tradecraft conventions: evidence is separated from assessment, probability language is calibrated and used consistently, and each assessment states what would disconfirm it.	Carries (extended)
Wartime corpora are shaped by strategic communication, censorship, and disinformation.	The same, plus adaptation: actors who learn the indicator set can adjust surface features while holding the underlying position. Indicator sets are rotated periodically and are not published in full.	Carries (extended)

Retrospective criterion	Monitoring analogue	Relation
A dual-reading protocol treats belligerent-state sources as both secondary literature and primary artefact.	Retained, and extended by explicit cascade-position coding of each source.	Carries
Ethical distance from the object of study is maintained.	Purpose limitation is stated in the design phase. Monitoring outputs on foreign-policy discourse are read by policy and security audiences, which places the analyst inside the process; anticipatory media literacy must be distinguished explicitly from intelligence production.	New
Findings are published as an article, with claims stated as demonstrated results.	Findings are released as a periodic bulletin, graded as signal, watch, or noise, with confidence expressed in calibrated terms.	Inverts

Note on the three consequential inversions

Sampling. Constructed-week sampling is so well established in content analysis that its rejection reads at first as methodological carelessness. The justification is that the two designs estimate different quantities. Constructed-week sampling is an efficient estimator of a *period parameter* and achieves that efficiency precisely by discarding days. A monitor does not estimate a period parameter; it tests for departure from a baseline at the earliest feasible moment, and a discarded day is a discarded opportunity for detection. The efficiency gain is purchased with exactly the property the monitor needs.

Outlet selection. Selecting outlets for baseline stability rather than representativeness appears to sacrifice external validity. It does so deliberately. The monitor's target is not a description of the media system but the detection of movement, and detectability is a function of the signal-to-baseline-variance ratio. Low-variance, ritualized discourse yields high information from small departures; the classic craft of inferring what was happening inside a closed political system from indirect, often trivial surface signals — because direct evidence was unavailable - result rests on this property rather than on any special insight into the texts. Where the media system is plural and no outlet has a tight baseline, the measurable object shifts from wording to distribution — the appearance of a previously marginal position in a centre outlet, or a contraction in the spread of positions across the field — and requires dispersion statistics rather than anomaly detection.

Translation. Translating before coding is standard practice in comparative work and is nonetheless disqualifying here. The signal in emerging repositioning resides substantially in designation lexicon, modality, and agency attribution, all of which translation normalizes. A weakening from obligation to recommendation, or a substitution of one designation for a near-synonym, will typically survive translation only by accident. Parallel indicator sets are more laborious to construct and require explicit equivalence testing, but they preserve the measurement.

Sixth, alternative approach can be seen as presented in the Attachment 3.

Research Recommendations

Building on these lessons, future research should proceed along the following lines:

1. **Comparative benchmarking of AI tools.** A two-phase pilot study should compare how approximately twenty current core AI tools handle identical foreign policy monitoring tasks, enabling both error identification and the tracking of AI advancement over time. This is realisable at limited cost (up to EUR 35,000, including tool access) and within approximately five person-months work.
2. **A hybrid monitoring architecture.** Future monitoring systems should combine multilingual computational pipelines (for scale and speed) with human-coded framing analysis (for conceptual accuracy), triangulated or pentagulated with events data, official-statement corpora, and audience-side engagement and survey data. Alliance-bloc membership should be built in as a stratification variable, and lexical monitoring should be developed as an early-warning instrument.
3. **A systematic review of previously funded projects.** Before new investments are made, a thorough review of previously funded research projects (by and large collected within this research) in the field should be carried out, focusing on learning from failures and on the deeper reasons why these projects and initiatives were not continued or further developed.
4. **Higher methodological standards in IR research.** IR scholarship should explicitly state its theoretical framework, incorporate the perspectives of all principal actors, clearly articulate assessments of initial triggers and aggressors, and adopt extended verification procedures such as pentagulation — review through at least five independent assessors or theoretical—empirical perspectives from diverse and neutral contexts. Peer review procedures themselves should be subject to independent assessment, and academic databases should consider introducing a ‘contested findings’ label for disputed empirical claims.
5. **Strengthened analytical education.** IR curricula should include core analytical competencies: a nuanced understanding of international law and state recognition, basic politico-military strategy, and media literacy covering agenda-setting, framing, and priming.
6. **Operationalising the analytical categories.** The close-reading categories identified so far — asymmetric standards, outcome prestructuring, inferential extension, preservation of alternatives, and inferential traceability — should be turned into explicit coding rules and tested for intercoder agreement before their prevalence across the corpus is estimated.
7. **Corpus-wide analysis.** The broader, search-retrieved corpus should be completed and frozen under predefined inclusion criteria, with human-validated coding in which AI assists structured extraction rather than acting as sole judge of interpretive variables.
8. **Measuring auditability.** A corpus-wide study should track how often publications disclose the sampling frames, coding rules, denominators, reliability statistics, and specifications needed to check their empirical claims independently.
9. **Comparative designs within controversial (e.g. in authoritarian states) academic publishing.** Within-author, within-journal, and matched cross-institutional comparisons before and after critical periods could test whether political sensitivity shifts research questions, methods, and inferential practice, while holding disciplinary and individual

differences constant.

10. **Linguistic orientation as a testable hypothesis.** Comparison with less politically sensitive media research should clarify whether the prominence of lexical, metaphorical, and conceptual analysis reflects the discipline's general structure or intensifies specifically around war-related studies.
11. **Movement towards “safer” research objects.** Matched comparisons across authors, journals, and periods could test whether greater political sensitivity pushes scholarship towards foreign, historical, or narrowly linguistic objects and away from direct examination of causality, responsibility, verification, and audience effects.
12. **Distance between evidence and claims.** Claim-level coding could test whether political sensitivity widens the gap between the level of available evidence — text, digital trace, survey, or experiment — and the level of the conclusion drawn, and whether this gap tracks the claim's political direction.
13. **International comparison.** A parallel international corpus, coded to the same rules, is needed to establish whether any differences reflect methodological quality, disciplinary composition, inferential practice, or political directionality — or disappear once discipline, genre, and method are controlled for.

Policy Recommendations

For policymakers at the national and EU levels, the project's findings support the following measures:

1. **Establish a dedicated European media monitoring capability for foreign policy.** A permanent interdisciplinary, virtual or cross-border team should coordinate comprehensive research and policy strategies for detecting new foreign policy trends in legacy media, integrating AI- and ML-based tools — including large reasoning models — with carefully selected and trained intelligence-type analysts, supported by researchers providing conceptual materials and critical feedback. Cooperation with the private sector can provide a critical advantage and access to cutting-edge experience; expertise can be pooled remotely across the EU without new core staff.
2. **Strengthen collaboration and stakeholder infrastructure.** This should include a shared resource database, regular online and offline forums for selected participants, rapid-response communication strategies for unfolding international events, a strategic approach to exclusive information sharing with specific news outlets, mechanisms addressing coverage gaps in important niche events and localities, and initiatives to test and improve journalists' professional competences in evaluating contradictory and incomplete information.
3. **Guard against narrative mischaracterisation.** Governments and public communicators should avoid rhetoric, foreign policy formulations, or media framings that may recast peacekeepers as aggressors, and should prioritise nuanced understanding of international actors' motivations over binary or distorted assessments — recognising that accurate comprehension, not endorsement, is crucial for strategic engagement. It should be clearly differentiated, legally, politically, and analytically, between aggression *per se* and excessive intervention by peacekeeping forces in response to aggression (or military attack).

4. **Exercise caution with sources.** Policymakers and their contractors should treat press-report datasets with caution, as reporting probabilities vary by event type, location, and time and may reflect media biases rather than actual phenomena; they should likewise approach military-affiliated publications on contested issues with heightened scrutiny, and consult non-Western and neutral-country sources wherever feasible.
5. **Consider European AI capacity strategically.** While Europe-based AI tools (e.g., Mistral Le Chat, Lumo) would mitigate concerns over data residency, cloud compliance, and vendor lock-in, their current performance on demanding analytical tasks lags behind leading US- and China-based models. Procurement and capacity-building decisions should weigh this trade-off explicitly and support the development of competitive European alternatives.

In sum, the project demonstrated that effective anticipatory monitoring of foreign policy discourses is now technically within reach, but only through a deliberately hybrid design that combines AI-facilitated scale with sustained human analytical judgement, methodological transparency, and institutionalised openness to dissenting perspectives. The geopolitical urgency of such a capability, and the maturity of the available tools, make the present moment the right one to act.

ATTACHMENTS

Attachment 1) An interview with analyst and former diplomat Balázs Jarábik, published on Postoj.sk, 18 June 2026

Jarábik was as a long-serving EU diplomat who, from 2019–2023, headed the policy, analysis and coordination unit of the EU Advisory Mission (EUAM) in Ukraine, having previously advised the Ukrainian government on EU integration.

The key messages that anchor the whole interview are:

- a) **Europe is paralyzed by the absence of a genuine strategic debate about how the war should end and what future relations with Russia should look like,**
- b) **EU/European governments lack any clear strategy for how, and toward what end, to conclude the war,**
- c) **that Russia will not give up its Donbas objective despite recent Ukrainian gains.**

Jarábik's divides the war into **three separate "layers": a war of narratives (Ukraine currently winning), a war of attrition (Russia has the edge at the front), and a war of adaptability (Ukraine performing strongly, aided by massive Western support).**

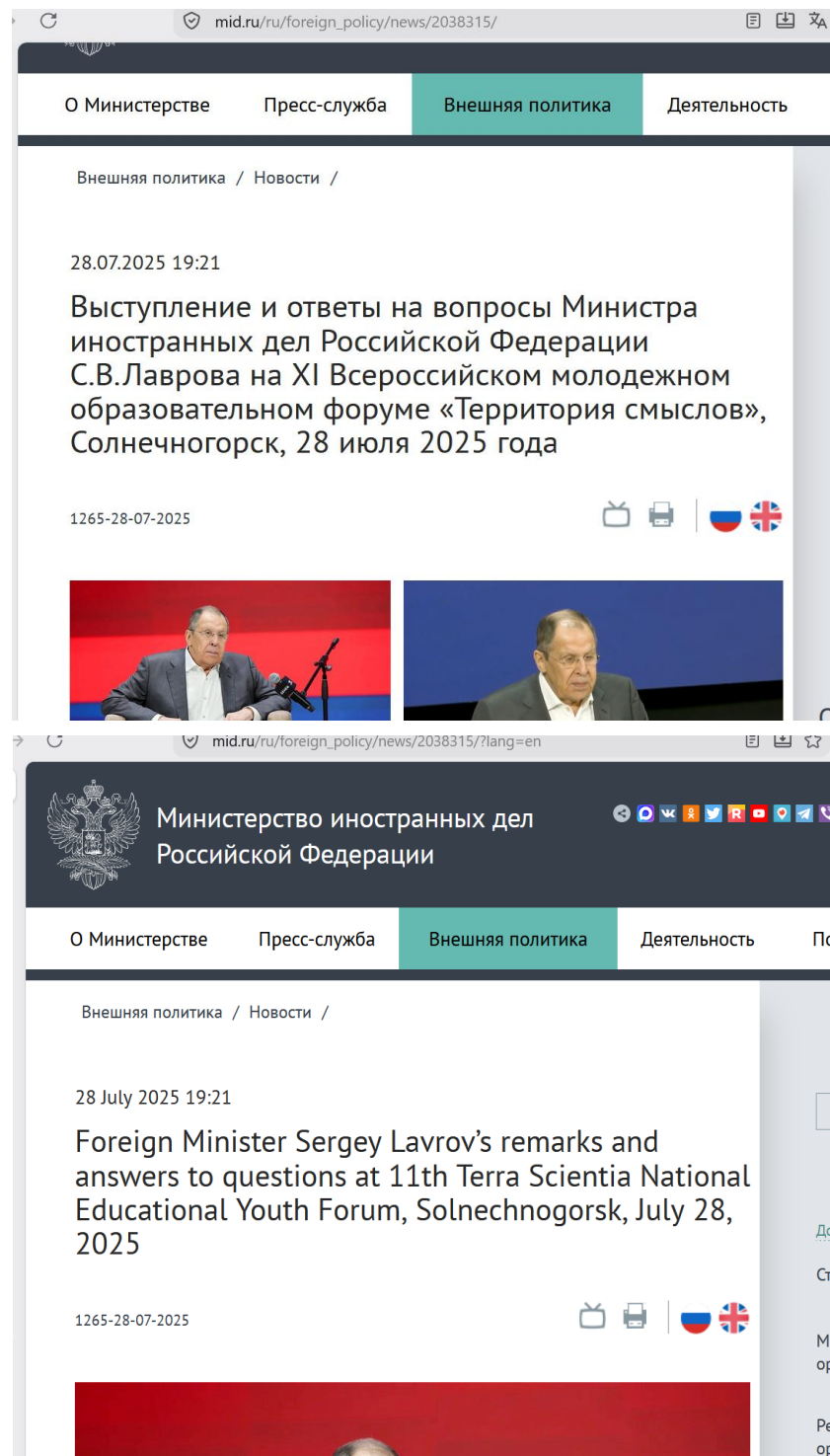
<https://www.postoj.sk/196061/europska-taktika-voci-rusku-je-nepochopitelna-kopirujeme-bidenovsku-politiku-akurat-bez-bidena-a-s-trumpom-na-obzore>

Attachment 2) Speech by the Minister of Foreign Affairs of the Russian Federation Sergey Lavrov in 2025

As put by the Minister of Foreign Affairs of the Russian Federation Sergey Lavrov in 2025:

*“In Europe, the Ukrainian issue is a manifestation of the Western line to inflict a “strategic defeat of Russia.” They are not embarrassed to confirm that they have been preparing for this for a long time... We insist on what is our legitimate demand: **ensuring our security. First, no involvement of Ukraine in NATO and no expansion of the alliance at all** (it is already in spite of all the promises and documents that were adopted, expanded close to our borders). Secondly, everyone says: “Russia must return to the borders of 1991.” In 1991, when Ukraine was recognized as an independent state, it had as the main principle its Declaration of Independence, which said: “a bloc-free, nuclear-free, **neutral state**.” It was in this capacity that the territorial integrity of Ukraine was recognized. When they began to destroy, destroy everything Russian, we could not stand aside, tried to persuade, **negotiations - it did not work out**. Therefore, there was no alternative to the beginning of a special military operation. The territories are not important to us. Sometimes they say, “They have seized the territory, we need to liberate them.” **These territories are not important to us – we have the largest country in the world**. It is important for us that people who have lived there for centuries and are native to Russian culture, language, education, who want to raise their children in the same culture, so that they are not exterminated and protected by their rights. This is a perfectly legal requirement.”*

Figure 1. Speech and answers to questions of the Minister of Foreign Affairs of the Russian Federation Sergey Lavrov at the XI All-Russian Youth Educational Forum “Territory of Meanings”, Solnechnogorsk, July 28, 2025, 28.07.2025 19:21, https://mid.ru/ru/foreign_policy/news/2038315/



Attachment 3) A Pilot Experiment for Analysing Foreign Policies Changes Based on Media Reporting

Countries Discourses Compared: India and Pakistan

Period Analysed: June 2026 (selected days).

Period of Sampling for Experiment: June 2026

Media Selected for Analysis (design vs. realized sample): *Designed sample:* Daily Jang and Express (Urdu, Pakistan); Dainik Jagran and Dainik Bhaskar (Hindi, India).

Realized sample: Daily Jang and Express, with same-date substitutions from Nawa-i-Waqt and Daily Pakistan where Jang yielded no retrievable item; on the Indian side, AajTak Hindi (aajtak.in, a TV-affiliated news portal) and Prabhasakshi (Hindi online daily) replaced the two flagship print dailies, whose archives proved non-extractable by date through open web search. All substitutions are marked in the text.

Tool used: Claude Fable 5 (Anthropic), the first model of the Claude 5 family and Anthropic’s most capable generally available model at the time of the experiment (see Anthropic’s announcement, <https://www.anthropic.com/news/claude-fable-5-mythos-5>).

Instructions, Step 1: Select an online Pakistani newspaper in a major local language. Extract articles from the last month, but select only every Monday issue. Focus on relations between India and Pakistan only. Analyse whether one can notice positive, neutral or negative developments in these relations, identifying words such as progress, regress, tensions, friendship, cooperation, challenges, obstacles and similar positive, negative or neutral terms. Provide a chart that includes positive, negative or neutral descriptions of India-Pakistan relations during that period. Provide a table that will include the number of positive, negative and neutral terms as per above.

Results:

AI has compiled these data from Daily Jang (jang.com.pk), Pakistan’s oldest and largest-circulation Urdu daily, for June 2026 (Mondays: 1, 8, 15, 22, 29 June). One important methodological caveat before the findings: Jang’s Monday print issues are only available as an image-based e-paper (e.jang.com.pk), which cannot be text-mined remotely. AI therefore worked with the retrievable full-text items from Jang’s online edition dated on or immediately adjacent to each Monday, supplemented — where Jang yielded no retrievable India-related item for a given Monday — by same-date items from other major Urdu dailies (Nawa-i-Waqt, Daily Pakistan), each clearly marked below. Term counts are based on the retrieved article texts, not complete issues, so one should treat this as a pilot-level approximation of sampling design rather than a strict replication.

The June 2026 context provided by AI (revised)

The month sat within a post-May-2025 conflict environment: India reaffirmed that the Indus Waters Treaty remains suspended in response to what it calls Pakistan’s cross-border terrorism, a suspension dating from the April 2025 Pahalgam attack, while Pakistan’s airspace remained closed to Indian aircraft, with the closure extended by NOTAM. Against this negative structural backdrop, late June brought *reports* of tentative openings: an informal Track-2/Track-1.5 dialogue in Colombo reportedly discussed water, airspace, crisis-communication mechanisms and the avoidance of future escalation. A Daily Pakistan report of 21 June additionally described Pakistan–India technical-level water talks at Bürgenstock, Switzerland, framed as “important progress.” **This claim must be treated as contested:** the 21 June Bürgenstock talks were primarily the US–Iran negotiations

mediated by Pakistan under the Islamabad Memorandum of Understanding (signed 17 June), and no India–Pakistan water track was ever officially confirmed. Pakistan’s Ministry of Water Resources explicitly denied knowledge of any progress on 28 June, and New Delhi dismissed the Colombo contact as a private initiative with no official standing, repeating that terrorism and talks cannot go together. The confirmed India–Pakistan contact of late June was therefore the informal Colombo meeting only.

Sensitivity note. Week 4 coding partly rests on the contested 21 June report. Excluding the contested item reduces Week 4 from $P = 6$ to approximately $P = 4$ and the monthly totals from 15/13/15 to approximately 13/13/15. The direction of the overall findings (oscillation around a negative structural baseline; religious-cultural exchange as the most consistently positive strand) is unchanged; the size of the Week-4 positive spike is not robust.

Monday-by-Monday coding

Week 1 (Mon 1 June). No retrievable Jang online item on bilateral relations dated 1 June; nearest Jang item (4 June) was a descriptive piece on Modi becoming India’s longest-serving PM, with a side reference that a ceasefire was aimed at reducing tension in the region. Tone: predominantly neutral, one de-escalatory positive marker (کشیدگی کم کرنا – “reducing tension”).

Week 2 (Mon 8 June). Substitute: Nawa-i-Waqt column of 8 June on Azad Kashmir unrest, which framed India adversarially — suggesting India was gloating over the unrest and describing the situation as “India’s game” (بھارت کا کھیل، محاذ آرائی). Counterweight: Daily Pakistan (8 June) on comprehensive hospitality arrangements for Indian Sikh pilgrims arriving via Wagah on 10 June. Tone: negative-dominant with a positive religious-diplomacy strand.

Week 3 (Mon 15 June). Jang item of 16 June: singer-actor Haroon Shahid claimed Pakistan is more developed than India and that Indian media has constructed a distorted image of Pakistan — competitive/negative image framing (مخصوص تاثر). Pilgrims’ itinerary reaching Lahore 15–16 June continued the positive strand. Tone: mixed, mildly negative.

Week 4 (Mon 22 June). The most positive Monday. 337 Indian Sikh pilgrims arrived via Wagah to a warm welcome, with exemplary arrangements that Indian Sikh delegations praised and thanked Pakistan for (پرتپاک استقبال، مثالی انتظامات). This coincided with the contested 21 June report of B rgerstock water talks, framed in that report as پيش رفت (“progress”); see the revised context and Sensitivity note above — this element of the Week-4 positive count is flagged CT (contested) in the codebook sense. Tone: positive-dominant (upper bound).

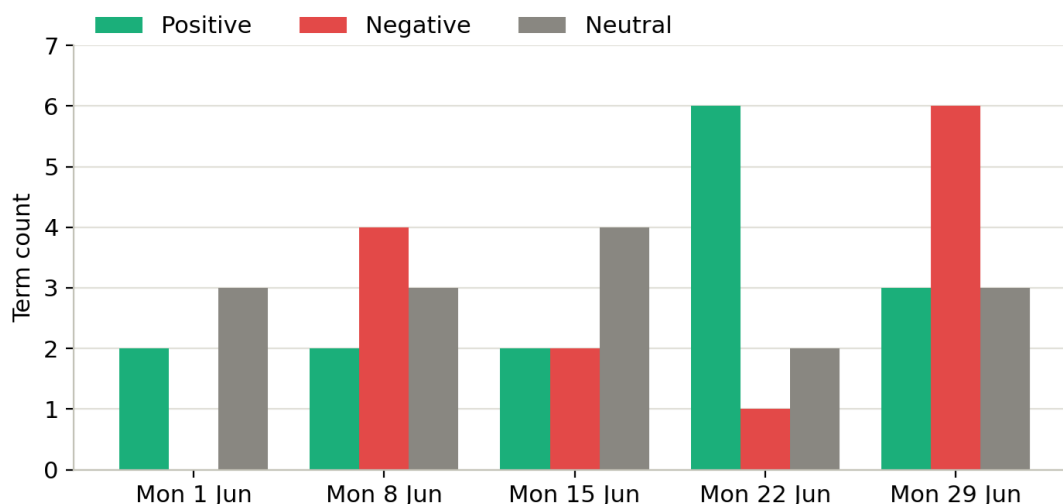
Week 5 (Mon 29 June). Jang’s 28 June lead: the Ministry of Water Resources said it was unaware of any progress on the Track-2 proposal to restore the Indus Waters Treaty in exchange for reopening airspace — dialogue vocabulary (مذاكرات) but explicit denial of پيش رفت. Surrounding coverage was heavily conflictual: a federal minister stating that 6,000 Pakistanis had died because of India and that no country can unilaterally suspend the treaty, alongside seminar rhetoric that the treaty is the foundation of peace, stability and water cooperation in the region. Tone: negative-dominant with peace/cooperation vocabulary deployed defensively.

Table 2 presents the term counts with the actual Urdu markers identified in the retrieved texts; Figure 2 charts the tonal distribution across the five Mondays.

Table 2. Positive, negative and neutral terms in the Daily Jang–centred Monday sample, June 2026

Monday (week)	Positive terms (examples)	Negative terms (examples)	Neutral terms (examples)	P	N	Neu
1 June	جنگ بندی (ceasefire), کشیدگی کم کرنا (reducing tension)	—	وزیراعظم, ریکارڈ, بیان (descriptive political reporting)	2	0	3
8 June	مہمان نوازی (hospitality), سہولت (facilitation) — Sikh yatra	بھارت کا کھیل (“India’s game”), محاذ آرائی (confrontation), بدامنی (unrest), بغلیں بجانا (gloating)	مذاکرات, مطالبات, صورتحال	2	4	3
15 June	استقبال (welcome), یاترا انتظامات (pilgrimage arrangements)	مخصوص تاثر (constructed negative image), حقیقت سے مختلف تصویر (distorted portrayal)	موازنہ, تاثرات, پوڈکاسٹ, ناظرین	2	2	4
22 June	پرتپاک استقبال (warm welcome), مثالی انتظامات (exemplary arrangements), شکر (gratitude), محبت/رواداری (love/tolerance), اہم مذاکرات (talks), اہم پیش رفت (important progress; flagged CT)	کشیدگی (tension — backdrop reference)	تکنیکی سطح (technical level), وفد (delegation)	6	1	2
29 June	امن (peace), استحکام (stability), تعاون (cooperation) — all flagged DP	معطلی (suspension), بٹ دھرمی (obstinacy), دہشت گردی (terrorism), آبی جارحیت (water aggression), بندش (closure), جاں بحق (deaths attributed to India)	مذاکرات, ٹریک ٹو, مؤقف, ترجمان	3	6	3
Total				15	13	15

Figure 2. Tonal distribution of India–Pakistan relations coverage across the five Mondays of June 2026 (term counts; Daily Jang–centred sample with marked substitutions). The Week-4 positive count is an upper bound (see Sensitivity note).



Interpretation by AI

The month shows no linear trend but rather an oscillating pattern around a structurally negative baseline. The conflict-era architecture (treaty suspension, airspace closure, terrorism accusations) generates a steady stream of negative vocabulary, and the sharpest negative spike comes at month’s end precisely because a positive development (the Track 1.5 opening) forced both governments into hardline public denials — India insisting terrorism and talks cannot go together, Pakistan’s ministry disclaiming knowledge of any progress. Meanwhile the most consistently positive register throughout June was not diplomacy at all but religious-cultural exchange (Sikh pilgrimages), where an unambiguous friendship vocabulary — messages of love, peace and tolerance — appeared. Notably, much of the “positive” vocabulary (امن, تعاون, استحکام) is deployed defensively, in arguments about what India’s treaty suspension threatens, which is a pattern worth flagging in any framing analysis: valence at the lexical level and valence at the relational level diverge (see Codebook, category DP).

Three caveats: (1) the counts derive from retrievable online texts, not complete Monday print issues; (2) two Monday data points relied on substitute Urdu dailies where Jang yielded no dated item; and (3) a five-Monday, single-outlet sample is illustrative, not generalizable.

Instructions, Step 2: Prepare and execute a constructed-week sample across Jang and Express. And extend this to a second Urdu daily for comparison.

Results:

Design and retrieval outcome

Constructed week (fixed a priori): Mon 1 (W1), Tue 9 (W2), Wed 17 (W3), Thu 25 (W4), Fri 12 (W2), Sat 20 (W3), Sun 28 (W4) — June 2026, two outlets: Daily Jang (news-led national daily) and Express (express.pk, which in the retrievable sample contributed mainly opinion columns).

Cell fill rate: 6 of 14 cells codable, four of them via ± 1 –4 day nearest-issue substitution (marked ★). Neither outlet’s early-June online archive yielded retrievable India-relations items for the exact sampled dates, and Jang’s print e-paper remains image-locked. Crucially, non-retrieval $\neq$ non-coverage: the empty cells reflect archive searchability, not editorial silence, so the totals below characterize retrievable coverage only.

Table 3. *Constructed-week cell grid, Daily Jang and Express, June 2026*

Cell	Jang	Express
Mon 1	★ 4 June: Modi to become longest-serving PM on 10 June; passing ceasefire/de-escalation reference	— no item retrieved
Tue 9	—	—
Fri 12	—	—
Wed 17	★ 16 June: Haroon Shahid claims Pakistan more developed than India; Indian media accused of constructing a distorted image of Pakistan	—
Sat 20	—	★ 21 June: Bürgenstock live report incl. JD Vance’s quip that his two most important people are his Indian-born wife and Pakistan’s Field Marshal Munir
Thu 25	—	✓ 25 June: Dr Tauseef Ahmed Khan column urging Pakistan to leverage its diplomatic success to normalize relations with India, offer talks, and pursue peaceful relations, against a backdrop of frostiness, suspended treaty, divided families and interference accusations
Sun 28	✓ 28 June: Water Ministry unaware of any progress on the Track-2 airspace-for-treaty-restoration proposal after the Sri Lanka talks	★ ~26 June column: relations under Modi described as continuously destabilized, with the path of dialogue closed

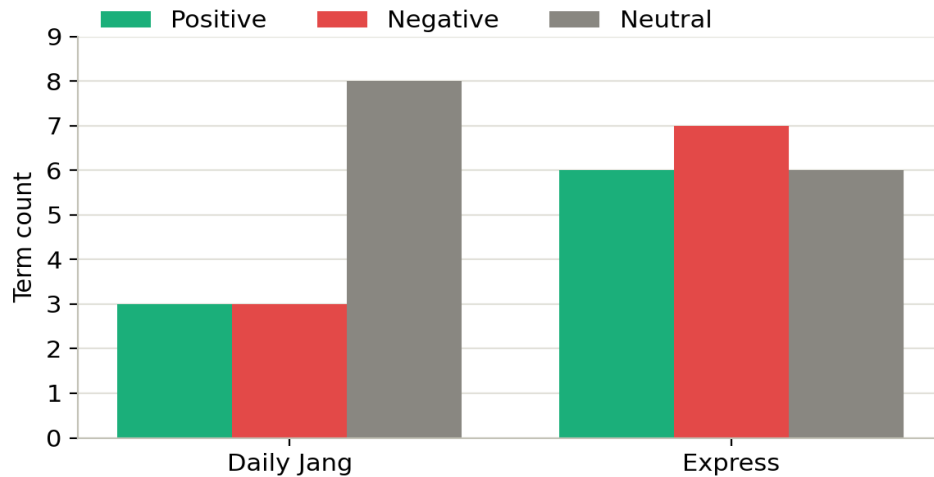
An adjacent Express column outside the sampled cells — framing May 2025 as India’s dramatic military humiliation, a crushing blow and successive diplomatic setbacks — is noted qualitatively as it typifies Express’s triumphalist strand, but is excluded from the counts.

Comparative term counts (retrievable constructed-week corpus)

Table 4. *Positive, negative and neutral terms by outlet, constructed-week sample, June 2026.*

Outlet	Positive terms (examples)	Negative terms (examples)	Neutral terms (examples)	P	N	Neu
Jang (3 items, news register)	جنگ بندی (ceasefire), فضائی حدود کھولنے / بحالی کی تجویز (reopening/restoration proposal)	مخصوص تاثر (constructed negative image), حقیقت سے مختلف تصویر (distorted portrayal), پیش رفت کا علم نہیں (progress denied)	مذاکرات, ٹریک ٹو, ترجمان, ریکارڈ, موازنہ, وزارت	3	3	8
Express (3 items, column-heavy)	تعلقات معمول پر لانا (normalization), مذاکرات کی پیشکش (offer of talks), پرامن تعلقات (peaceful relations), مثبت جواب (positive response), مشترکہ حکمت عملی (joint strategy)	سرد مہری (frostiness), مداخلت (interference), دہشت گردی کی واردات (terror attack), انتہاپسند حکومت (extremist government), غیر مستحکم (destabilized), راستہ بند (dialogue closed), کشیدگی (tension)	بائی کمشنر, ویزے, مکالمہ, تعلقات, وفود	6	7	6
Total				9	10	14

Figure 3. *Positive, negative and neutral term counts in Daily Jang versus Express coverage of India–Pakistan relations, constructed-week sample, June 2026.*



Register, not just outlet, drives the difference. Jang's retrievable items are news-desk reporting: 57% of its coded terms are neutral-institutional (spokespersons, track-two, negotiations), with

valence entering mainly through quoted actors. Express’s retrievable sample is column-dominated, and its valenced share (68%) is roughly double Jang’s — but strikingly, Express supplies both poles of the June discourse: the strongest pro-normalization advocacy in the corpus (the 25 June call for Pakistan to offer talks and pursue peaceful relations) and, just outside the sampled cells, the harshest adversarial framing (India’s “military humiliation”). This is a genre effect familiar from the peace-journalism framework: opinion pages are where both war-journalism and peace-journalism vocabularies concentrate, while news pages default to elite-indexed neutrality (very much Bennett’s indexing at work — the tone of Jang’s news items tracks official statements almost mechanically).

One correction to previous analysis that this second pass surfaced: the 21 June Bürgenstock technical talks were primarily the US–Iran negotiations mediated by Pakistan under the Islamabad MoU signed 17 June; a Daily Pakistan report of 21 June did describe Pakistan–India water-related technical talks at the same venue, but this India–Pakistan track was never officially confirmed — Pakistan’s Water Ministry explicitly denied knowledge of progress on 28 June, and the confirmed India–Pakistan contact of late June was the informal Colombo meeting, which Delhi dismissed as a private initiative with no official standing, repeating that terrorism and talks cannot go together. So the “positive” spike coded for the week of 22 June rests partly on contested reporting — itself an analytically interesting instance of hopeful framing outrunning verifiable diplomacy in the Pakistani press. (The Step 1 context passage and Table 2 have been revised accordingly; see the Sensitivity note.)

Limitations for the write-up as suggested by AI

The constructed week achieved only 43% cell fill from open web archives, with substitution in four cells, so outlet totals are indicative, not inferential; term counts derive from retrieved full texts and extended snippets rather than complete issues; the Jang/Express comparison is confounded by genre composition (news vs. columns); and single-coder lexical classification without intercoder reliability (e.g., Krippendorff’s α) limits replicability. For a publishable version one would want e-paper access (or press-archive subscription equivalents for Pakistani dailies), a two-constructed-week sample per outlet, and separate coding strata for news versus opinion.

.....

Instructions, Step 3: Select two online Indian newspapers in a major local language. Extract articles from the last month, execute a constructed-week sample. Focus on relations between India and Pakistan only. Analyse whether one can notice positive, neutral or negative developments in these relations, identifying words such as progress, regress, tensions, friendship, cooperation, challenges, obstacles and similar positive, negative or neutral terms. Provide a chart that includes positive, negative or neutral descriptions of India-Pakistan relations during that period. Provide a table that will include the number of positive, negative and neutral terms as per above.

Results:

For the Indian side AI used Hindi — the largest local-language press market — with the two highest-circulation dailies: Dainik Jagran (jagran.com) and Dainik Bhaskar (bhaskar.com). To keep the comparison symmetrical with the Pakistani study, AI applied the identical constructed week for June 2026: Mon 1, Tue 9, Fri 12, Wed 17, Sat 20, Thu 25, Sun 28. One finding has to lead the AI write-up, because it changes the outlet selection: the two flagship Hindi print dailies AI targeted first, Dainik Jagran and Dainik Bhaskar, proved non-extractable by date through open web search — their Hindi archives are JS-rendered/app-first and poorly indexed for date-stamped retrieval, the same class of problem as Jang’s image-based e-paper. Rather than silently misattribute, AI

substituted the two Hindi outlets whose June 2026 India–Pakistan items were retrievable with verifiable dates: AajTak Hindi (aajtak.in — India’s largest Hindi news portal, TV-affiliated) and Prabhasakshi (Hindi online daily). AI also suggested possible database access (Factiva, LexisNexis, Media Cloud).

The Hindi-language corpus shows even stronger temporal clustering than the Pakistani one: virtually all retrievable India–Pakistan items fall in the last week of June, triggered by the Colombo Track-2 revelations and the Indus Waters Treaty exchanges. Early- and mid-June cells (Mon 1 through Sat 20) returned no retrievable India–Pakistan items from either outlet — again, non-retrieval, not necessarily non-coverage. The codable corpus:

Thu 25 June (context source, DW-Hindi syndication): Pakistan’s draft budget raises defence spending 18% amid regional uncertainties, with analysts citing rapidly changing military technologies and emerging security threats in South Asia — security-threat framing, no positive markers. Used as contextual anchor for the Thursday cell, not coded into outlet totals.

Thu 25–Fri 26 June ★ (Prabhasakshi): Track-2 report — despite sharp rhetoric and a tense atmosphere, former diplomats, retired officers and academics from both countries launched a new informal dialogue initiative in Colombo and Bangkok, aiming to keep communication channels open and keep tension under control; the same piece recounts that relations froze after 2019, trade is negligible, and India holds that meaningful dialogue is impossible while terrorism continues.

Sat 27 June ★→Sun 28 cell (AajTak): commentary-inflected news analysis framing Pakistan as playing a victim card in a narrative war over the Indus waters, citing provocative rhetoric from Pakistani ministers linking water security to war, and recalling Operation Sindoor and Modi’s “blood and water cannot flow together” line.

Sun 28 June (corroborating MEA-denial reporting, cross-checked in two further Hindi portals, not coded): the Foreign Secretary clarified the Colombo meeting had no official standing, participants’ views were personal, and it should not be read as government policy — while reports noted the meeting did discuss reducing military tension. Notably, Pakistani-media claims of an airspace-for-treaty swap offer were denied by a participant on the Indian side.

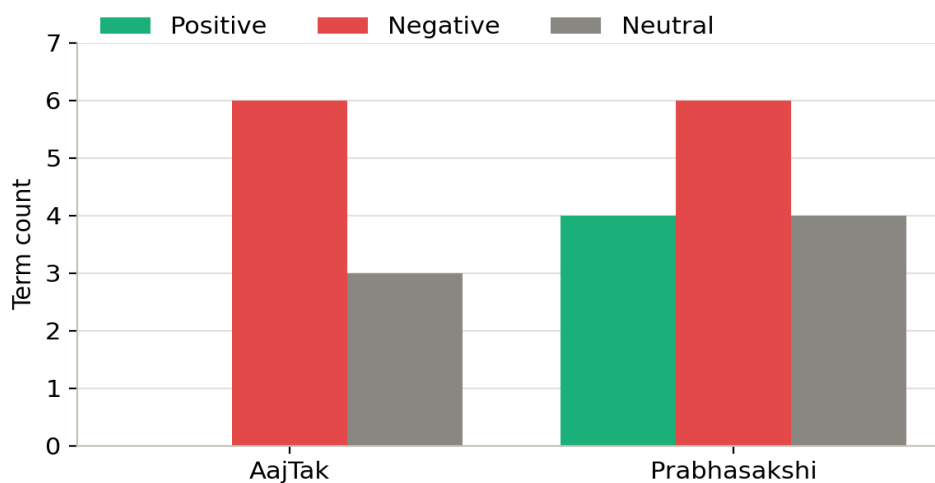
Term counts — Hindi outlets

Table 5. *Positive, negative and neutral terms by outlet, Hindi sample, June 2026.*

Outlet (items)	Positive terms (examples)	Negative terms (examples)	Neutral terms (examples)	P	N	Neu
AajTak (1)	—	तनाव (tension), उकसावे वाली बयानबाजी (provocative rhetoric), विक्टिम कार्ड/नैरेटिव वॉर (victim card/narrative war), जंग (war), खून और पानी साथ नहीं	संधि (treaty), जल प्रवाह (water flow), स्टोरेज क्षमता (storage capacity)	0	6	3

		(blood-and-water), आतंकी हमला (terror attack)				
Prabhasakshi (1)	संवाद की नई पहल (new dialogue initiative), रास्ते खुले रखना (keeping channels open), तनाव नियंत्रण (tension control), मैत्रीपूर्ण आदान- प्रदान (friendly exchanges)	तीखी बयानबाजी (sharp rhetoric), तनावपूर्ण (tense), आतंकी हमला, संघर्ष (conflict), संबंध ठंडे (frozen ties), नगण्य व्यापार (negligible trade)	ट्रेक टू, बैठक (meeting), प्रतिनिधि (representatives), सूत्र (sources)	4	6	4
Total				4	12	7

Figure 4. Positive, negative and neutral term counts in AajTak and Prabhasakshi coverage of India–Pakistan relations, June 2026.



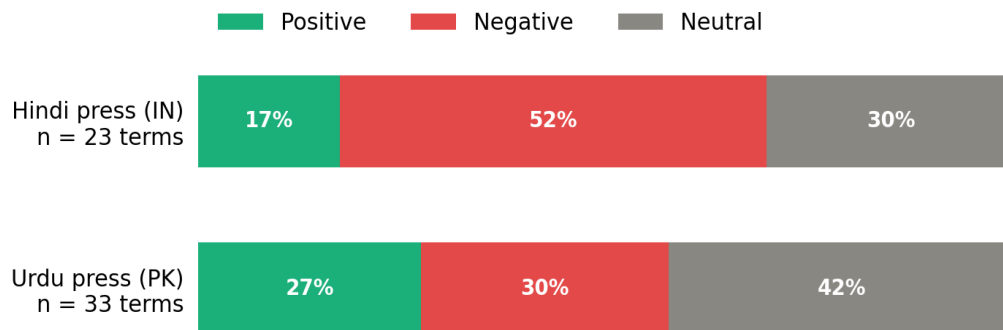
COMPARISON BY AI

Comparison baseline. The Urdu comparator throughout this section is the **constructed-week sample** (Table 4: P = 9, N = 10, Neu = 14; n = 33 coded terms), not the Monday sample of Table 3. The shares are: Hindi sample — positive 17%, negative 52%, neutral 30% (n = 23); Urdu sample — positive 27%, negative 30%, neutral 42% (n = 33). See Figure 5.

The asymmetry is the headline finding. In the Hindi sample, negative terms make up roughly half of all coded vocabulary, versus under a third in the Urdu sample, while the positive share is markedly lower. This mirrors the two countries' structural positions rather than mere editorial temperament: Pakistan is the demandeur of normalization (so its press hosts the vocabulary of talks, restoration, peace, and hospitality, even when deployed defensively), whereas India's official line makes engagement conditional — terrorism and talks cannot go together, and the treaty stays in abeyance

until cross-border terrorism credibly ends — so Indian coverage is saturated with conditionality and threat vocabulary even when reporting dialogue. Strikingly, both press systems covered the same late-June Colombo opening, but with inverted emphasis: the Urdu press foregrounded a possible airspace-for-treaty breakthrough, while Hindi coverage foregrounded the official denial — the government does not attach official importance to such private meetings — and an Indian participant’s denial that any swap offer was made at all.

Figure 5. Tone shares of coded terms: Hindi-press sample (India) versus Urdu-press constructed-week sample (Pakistan), June 2026. Percentages of all coded terms per sample.



Within the Hindi sample, the same genre gradient reappears as in both earlier studies: the straight-news/reportage register (Prabhasakshi’s Track-2 piece) carries a mixed valence profile, including the corpus’s only positive terms, while the commentary-inflected register (AajTak’s “victim card” analysis) is uniformly negative — with a distinctive Indian-press feature absent from the Urdu corpus: explicit meta-framing vocabulary (नैरेटिव वॉर, विक्टिम कार्ड) that delegitimizes the adversary’s discourse itself, not just its actions. That is analytically close to what a cognitive-warfare framework would classify as second-order framing. One further note: the single most inflammatory episode of the period — Bilawal Bhutto invoking Pakistan’s nuclear doctrine over the Indus waters, covered wall-to-wall in Hindi media with mocking idioms like गीदड़भभकी (“jackal’s bluster”) — falls on 1 July, just outside the June window, and is therefore excluded from the counts, though any longer study should capture it.

Study Limitations Identified by AI

Beyond the standard caveats carried over from the previous two studies (single coder, retrieved-text basis, no intercoder reliability), two are specific here: the flagship-outlet substitution described above, and a severe fill rate — effectively only the final constructed-week cells (Thu 25/Sun 28) were codable, so this Hindi sample characterizes the late-June news cycle, not the month. The India–Pakistan cross-country comparison in Figure 5 should accordingly be read as an illustrative contrast of retrievable late-June coverage, not a month-level population estimate; with archive access, a proper two-constructed-week design per country would be the minimum for publishable claims.

Feedback on Study by a Local (Pakistani) Expert Mohammed Sadji

(We are not certain how much of this is original contribution and how much has been written by AI).

This pilot study offers a tentative yet useful indication of how selected Pakistani and Indian media framed foreign-policy issues during the observed period. The central observation—that Pakistani coverage more frequently advocates dialogue and normalization, while Indian coverage emphasizes security, terrorism, and strategic caution—is consistent with known editorial tendencies in each media landscape.

The primary limitation lies in the substantial divergence between the realized sample and the original research design. On the Pakistani side, substitute newspapers were employed; on the Indian side, the intended comparison with *Dainik Jagran* and *Dainik Bhaskar* was replaced by *AajTak Hindi* and *Prabhasakshi*. These outlets differ in editorial stance, ownership, and audience, rendering the results non-representative of the flagship Hindi print press.

Overall, the draft is empirically grounded and consistent with established understandings of Indo-Pakistani foreign-policy discourse. Beyond the acknowledged uncertainty surrounding the Bürgenstock talks, no major factual inaccuracies are evident. The principal concerns relate to sampling, archival access, outlet substitution, and genre imbalance—not to substantive interpretation. As such, the pilot is suitable as an exploratory study, provided its conclusions remain appropriately qualified and not extended beyond the retrieved sample.

Comments on the Proposed Data Collection and Analysis Procedure by Konstantin Startsev

Limitations of Data Collection Using LLMs

An LLM cannot independently collect data in cases where a website generates content dynamically, via scripts executed directly in the browser. This problem can, however, be addressed using a specialized tool. I would recommend a program written in Python, running in the Google Colab cloud environment, which uses the Playwright library to emulate a standard Chromium browser. To recognize image-based electronic documents (for example, texts in Urdu script), it would be advisable to supplement this with optical character recognition (OCR) tools.

In this way, all relevant pages can be saved and a stable, closed corpus can be constructed. This corpus can subsequently be analyzed using various methods, made publicly available, and thereby allow independent verification of the analysis. At the same time, this approach resolves the question of scale — instead of dozens of articles, it becomes possible to collect thousands to millions of texts.

The "Black Box" Problem

A generative model, even one as capable as GPT-5, functions as a black box: its internal decision-making processes are neither observable nor independently verifiable. For this reason, the model does not necessarily produce an identical output when given the same input twice; if the same text is submitted again tomorrow, the resulting classification may differ. Moreover, the provider (Anthropic) continuously updates the model without necessarily notifying the user. Formally, it therefore remains "the same" model, but in practice its version may differ from one day to the next.

Conflation of Multiple Research Tasks into a Single Step

In the proposed pilot design, the LLM simultaneously retrieves data, translates it, classifies it, and analyzes it. This constitutes a single, non-transparent step whose outputs, moreover, cannot be independently verified without knowledge of Urdu and Hindi.

Recommended Procedure

I would propose the following procedure: a specialized tool would first be used to assemble a stable corpus, with a preliminary keyword-based selection of texts carried out at the same stage. The Claude Fable model would then, via the API (using a fixed, precisely identified version), perform coarse classification of a large volume of texts. A random sample would be independently coded by a human coder (for example, via the Prolific platform), and agreement between the model's classification and the manual coding would be assessed using the F1 metric or Krippendorff's alpha. Given a sufficient level of agreement, a smaller classification model (for example, XLM-RoBERTa) would be trained on these annotations, with frozen weights, so that it would always produce the same output for the same input; this model would then perform the final classification of the entire corpus.

This procedure is more time-consuming at the outset, but its result is a reproducible and verifiable measurement — not a collection of the model's impressions about search-engine outputs. I would be happy to discuss with you how to apply this procedure concretely to your material.

CODEBOOK

C1. Units of analysis. Three nested units are used. The *design unit* is the day-cell (a calendar date assigned a priori by the Monday-sample or constructed-week design). The *sampling unit* is the individual news item (article or column) retrieved for that cell and dealing with India–Pakistan relations. The *coding unit* is the valence-bearing lexical item — a single word or fixed multi-word expression (e.g., خون اور پانی / खून और पानी constructions) — occurring in the retrieved text and predicating the bilateral relationship or an actor's conduct within it.

C2. Relevance rule. A term is coded only if it refers to the India–Pakistan relationship, to one country's actions or statements directed at the other, or to bilateral instruments (treaty, airspace, talks, pilgrimages). Identical vocabulary used in a purely domestic context (e.g., مذاکرات about internal Azad Kashmir negotiations, तनाव about domestic politics) is excluded.

C3. Valence categories. *Positive (P)*: terms denoting or connoting improvement, rapprochement, amity or constructive engagement (progress, restoration, normalization, peace, cooperation, friendship, welcome, hospitality, confidence-building, ceasefire, de-escalation). *Negative (N)*: terms denoting deterioration, hostility, blockage or threat (tension, confrontation, obstinacy, terrorism, aggression, suspension, closure, war, humiliation, interference, frostiness, provocative rhetoric, distorted portrayal). *Neutral (Neu)*: institutional-procedural vocabulary without inherent valence (talks/negotiations as a label, track-two, meeting, delegation, spokesperson, treaty, statement, position). Borderline rule: مذاکرات / वार्ता ("talks") is coded Neutral when used as a procedural label, Positive only when explicitly framed as an achievement, offer or opening (e.g., مذاکرات کی پیشکش, "offer of talks").

C4. Treatment of repeated terms. Terms are counted as *types per article*, not tokens: an identical term repeated within one article is counted once, unless successive occurrences carry distinct

referents or distinct valence framings (e.g., مذاکرات as neutral label in one paragraph and مذاکرات کی پیشکش as a positive act in another), in which case each distinct framing is counted once. Across articles, the same term is counted anew per article.

C5. Journalist voice versus quoted elite speech (tag QS). Valence terms occurring in direct or indirect quotation of officials, politicians or experts are coded and included in totals — the media's relay of elite valence is part of the tonal environment the study measures — but each is tagged QS so that indexing effects (Bennett) can be analysed by stratifying journalist-voice terms against quoted-speech terms. In the present pilot, the large neutral share of Jang's news items and the concentration of valence inside quoted official statements is consistent with the indexing hypothesis.

C6. The “defensive positive” category (tag DP). Positive-valence lexemes (peace, stability, cooperation, تعاون, استحکام, امن) that occur inside an argument about threats to the values they name — typically in the pattern “X endangers the peace/stability/cooperation that the treaty guarantees” — are coded Positive at the lexical level but tagged DP, because the relational meaning of the passage is adversarial. DP terms are reported inside P totals but must be identified in any interpretation; in the Monday sample, all three Week-5 positive terms are DP. The P/DP distinction operationalizes the divergence between lexical valence and relational valence noted in the interpretation.

C7. Contested-event flag (tag CT). Terms whose positive or negative valence rests on factual claims that were subsequently denied or never officially confirmed (e.g., اہم پیش رفت attached to the 21 June Bürgenstock report) are coded normally but tagged CT; headline results must be reported both with and without CT items (see Sensitivity note under Step 1).

C8. Genre stratum. Every sampling unit is classified as news report, news analysis/explainer, or opinion column; all outlet-level comparisons must report genre composition, since genre confounds outlet tone (Step 2 finding).

C9. Reliability provision (for the full study). Two coders working independently from this codebook on the full corpus; Krippendorff's α computed at the coding-unit level for the P/N/Neu decision and for each tag (QS, DP, CT); $\alpha \geq .80$ required, with disagreement adjudication documented. The present pilot is single-coded and reports no reliability coefficient.

CONCLUSIONS:

The pilot offers a reasonable exploratory indication of how selected Pakistani and Indian outlets framed foreign-policy issues during the sampled period. Its findings, however, should be read as evidence of *media framing*, not as a direct measure of either state's foreign-policy stance. The coding broadly reproduces the dominant narratives associated with India–Pakistan coverage since the deterioration in bilateral relations following the 2019 constitutional changes, the 2025 security crisis, and the ongoing dispute over the Indus Waters Treaty. The general finding — that Pakistani coverage carries more language advocating dialogue and normalization, while Indian coverage foregrounds security concerns, terrorism, and strategic caution — seems to be consistent with the known editorial tendencies of both media environments.

The report is transparent about several of its own limitations: archive access, the use of substitute outlets, and the divergence between news and opinion genres. This transparency strengthens the report's credibility and appropriately restrains its claims; the decision to flag the reported Bürgenstock water talks as contested and to run a sensitivity analysis around them is sound

methodological practice. There are several underlying problems.

1. Sample realization diverges from design, and the divergence is non-random. On the Pakistani side, some observations rely on substitute newspapers; on the Indian side, the intended comparison with *Dainik Jagran* and *Dainik Bhaskar* effectively becomes a comparison involving AajTak Hindi and Prabhasakshi. These outlets differ from the originally planned newspapers in editorial style, ownership, and audience, so results should not be presented as representative of the flagship Hindi print press. As the analytical summary notes, this is compounded by a structural retrieval bias: search-based collection systematically over-returns event-driven, recent, high-engagement material, which is almost certainly why both corpora cluster in the final week of June rather than distributing across the month. The low retrieval rate is therefore not simply a data-availability problem but a *sampling bias with a direction* — the pilot likely over-represents peak-salience, high-affect coverage and under-represents routine reporting. This is stated explicitly as a limitation on generalizability and not necessarily folded into a general caveat about archive access.

2. Genre imbalance is a validity threat. A non-trivial share of coded items are commentary or opinion rather than straight news. Opinion columns carry stronger evaluative language by construction, which confounds cross-outlet comparison unless genre is modeled separately. The report acknowledges that its substantive conclusions should be re-stated conditional on genre, since the observed India–Pakistan asymmetry may partly reflect *which genre each outlet over-samples* rather than a genuine difference in editorial position. The pilot's own data — a genre gradient in which opinion pages carry both poles of sentiment while news pages remain closer to official framing — is itself an informative finding, but it undercuts, rather than merely qualifies, the outlet-level comparison as currently presented.

3. Lexical counts are an insufficient instrument, particularly given tooling constraints in Urdu/Hindi. Word-frequency analysis does not reliably capture stance: terms such as "peace," "cooperation," or "dialogue" can be deployed critically or defensively rather than affirmatively, and the pilot's lexical/relational divergence finding — peace vocabulary used in a defensive register — demonstrates this directly. The problem is sharper because no validated valence lexicon of comparable quality to English tools currently exists for Urdu or Hindi journalistic register. In its absence, the pilot's lexical layer cannot be assumed to carry the same discriminant validity that equivalent English-language dictionaries would. Two paths follow: either a purpose-built, human-validated lexicon should be constructed, or classification should be delegated to a model with a documented, auditable validation procedure — sentence- or clause-level classification checked against independent human coding, rather than word counts alone.

4. Indexing, not just tone, needs to be modeled explicitly. On security topics, Indian coverage frequently reproduces official government framing rather than an independent editorial position — a pattern consistent with Bennett's indexing hypothesis, which appeared to operate mechanically in at least some of the Jang items reviewed. This means the instrument, as currently constructed, partly measures *how government discourse is relayed by the press* rather than an independent signal of bilateral relations. This is a validity concern, not merely an interpretive nuance, and it should condition any causal or quasi-causal claims the study might later be tempted to make about "foreign policy change" on the basis of tone shifts. Tone may lag rather than lead policy; triangulation with events data (in the tradition of COPDAB/WEIS, GDELT tone measures) or with official-statement corpora would be required before such claims are defensible.

5. Methodological transparency should extend to the classification procedure itself, not only to sampling. If any part of classification or translation was performed by a large language model, this

introduces a further, distinct source of unreliability: outputs are not perfectly reproducible across repeated runs of the same input, and providers update models without notice, so a "same model" queried on different days is not guaranteed to behave identically. Reporting the exact model version used, together with an inter-coder-style reliability check (e.g., F1 or Krippendorff's α between model output and independent human coding on a random subsample), is a minimal transparency requirement — comparable to reporting inter-coder reliability for human coders — and its absence is flagged as a limitation on par with the sampling issues already discussed. A more robust design would separate the tasks the pilot currently collapses into a single step (retrieval, translation, classification, analysis), since collapsing them into one LLM pass makes the process difficult to audit, particularly for readers without Urdu or Hindi.

6. The pilot's design generalizes reasonably well to other active, high-salience interstate rivalries with substantial vernacular online media (comparable dyads include India–China, Armenia–Azerbaijan, Israel–Iran, Russia–Ukraine, and a broader set of perhaps 40–60 currently active cases). It generalizes less straightforwardly to fully state-controlled press systems, where the object of study shifts from "media tone" to "official framing" — a shift that should be declared as a design choice rather than treated as an incidental limitation if the method is extended to such cases. Similarly, if broadcast media were ever added to this design, lexical coding alone would systematically undercount negativity, since valence in broadcast formats is carried substantially through non-lexical channels (prosody, visuals, on-screen tickers); this is not a concern for the present print/digital pilot but is worth noting as a boundary condition on any future scaling.

Overall assessment suggests that the pilot is factually consistent with the broader scholarly and journalistic understanding of India–Pakistan relations, and no major factual inaccuracies are evident beyond the already-acknowledged uncertainty around the Būrgenstock reports. Its principal weaknesses are methodological rather than substantive: a non-random, salience-biased sample; unaddressed genre confounding; reliance on lexical counts without a validated Urdu/Hindi instrument; insufficient treatment of indexing as a validity threat; and, an unaudited LLM classification step. The study is an exploratory, hypothesis-generating pilot. Its conclusions should be read as provisional, explicitly conditioned on genre and sampling window, and should not be generalized to the flagship print press or treated as a proxy for the two states' foreign-policy positions without further, methodologically strengthened work along the lines outlined above.

REFERENCES

- Bennett, W. L. (1990). Toward a theory of press-state relations in the United States. *Journal of Communication*, 40(2), 103–125.
- DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6), 570–606. <https://doi.org/10.1016/j.poetic.2013.08.004>
- Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267–297. doi:10.1093/pan/mps028
- Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3), 411–433. <https://doi.org/10.1111/j.1468-2958.2004.tb00738.x>
- Matthes, J., & Kohring, M. (2008). The content analysis of media frames: Toward improving

reliability and validity. *Journal of Communication*, 58(2), 258–279. <https://doi.org/10.1111/j.1460-2466.2008.00384.x>

Riffe, D., Aust, C. F., & Lacy, S. R. (1993). The effectiveness of random, consecutive day and constructed week sampling in newspaper content analysis. *Journalism Quarterly*, 70(1), 133–139. <https://doi.org/10.1177/107769909307000115>

Startsev, Konstantin (2026): Comments on the Proposed Data Collection and Analysis Procedure. Personal communication.

Sadji, Mohammed (2026). Feedback on the micro-pilot study A pilot experiment for analysing foreign policies changes based on media reporting. Personal communication.

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of Communication*, 49(1), 103–122. <https://doi.org/10.1111/j.1460-2466.1999.tb02784.x>

Walter, D., & Ophir, Y. (2019). News frame analysis: An inductive mixed-method computational approach. *Communication Methods and Measures*, 13(4), 248–266. <https://doi.org/10.1080/19312458.2019.1639145>

Coded corpus — Pakistan (Urdu).

Daily Jang. (2026, June 16). پاکستان بھارت سے زیادہ ترقی یافتہ ہے: ہارون شاہد [Pakistan is more developed than India: Haroon Shahid]. *Daily Jang*. <https://jang.com.pk/news/1592049>

Daily Pakistan. (2026, June 8). جہلم، رنجیت سنگھ کی برسی پر سکھ یاتریوں کیلئے انتظامات مکمل [Arrangements completed for Sikh pilgrims for Jor Mela and Ranjit Singh's death anniversary]. *Daily Pakistan*. <https://dailypakistan.com.pk/08-Jun-2026/1936251>

Daily Pakistan. (2026, June 21). پانی کے تکنیکی امور پر پاک بھارت مذاکرات آج سوئٹزر لینڈ میں ہونگے [Pakistan–India talks on technical water matters to be held in Switzerland today]. *Daily Pakistan*. <https://dailypakistan.com.pk/21-Jun-2026/1939617>

Daily Pakistan. (2026, June 22). 337 بھارت سے سکھ یاتری پاکستان پہنچ گئے [337 Sikh pilgrims from India arrive in Pakistan]. *Daily Pakistan*. <https://dailypakistan.com.pk/22-Jun-2026/1939826>

Express News. (2026, June 21, approx.). ایران امریکا مذاکرات: فریقین کانفرنس ہال پہنچ گئے [Iran–US talks: Both sides arrive at the conference hall]. *Express*. <https://www.express.pk/story/2818598/iran-us-talks-both-sides-arrived-conference-hall-2818598>

Express. (2026, June 26, approx.). بھارت سے مکالمہ اور ایران سے تجارت [Dialogue with India and trade with Iran]. *Express*. <https://www.express.pk/story/2819665/bharat-se-mukalma-aur-iran-se-tijarat-2819665>

Khalid, K. M. (2026, June 4). مودی 10 جون کو بھارت میں طویل عرصے رہنے والے وزیراعظم بن جائینگے [Modi will become India's longest-serving prime minister on 10 June]. *Daily Jang*. <https://jang.com.pk/news/1588762>

Khan, T. A. (2026, June 25). کامیاب سفارتکاری اور بھارت سے تعلقات [Successful diplomacy and relations with India]. *Express*. <https://www.express.pk/story/2819141/kamyab-safarat-kari-aur-bharat-se-talluqa-2819141>

Mustafa, K. (2026, June 28). وزارتِ آبی وسائل ٹریک ٹو سطح پر سندھ طاس معاہدے سے متعلق [Ministry of Water Resources unaware of any Track-II progress on the

Indus Waters Treaty]. *Daily Jang*. <https://jang.com.pk/news/1595007>

Nawa-i-Waqt. (2026, June 8). آزاد کشمیر میں مذاکرات، محاذ آرائی اور بھارت کا کھیل [Negotiations, confrontation and India's game in Azad Kashmir]. *Nawa-i-Waqt*. <https://www.nawaiwaqt.com.pk/08-Jun-2026/1999732>

Coded corpus — India (Hindi).

AajTak. (2026, June 27). भारत से जंग की चाहत? सिंधु के पानी पर विक्टिम कार्ड खेल रहा पाकिस्तान [A desire for war with India? Pakistan playing the victim card over Indus waters]. *AajTak*. <https://www.aajtak.in/world/story/pakistan-narrative-war-indus-water-treaty-sindhu-jal-sandhi-indian-pak-war-operation-sindoor-mnrd-dskc-2587213-2026-06-27>

Prabhasakshi NewsRoom. (2026, June 26, approx.). जुबानी जंग के बीच तीसरे देश में India और Pakistan ने की गुप्त बातचीत [Amid a war of words, India and Pakistan hold quiet talks in a third country]. *Prabhasakshi*. <https://www.prabhasakshi.com/national/india-pakistan-track-two-dialogue-colombo-bangkok-talks-pahalgam-operation-sindoor>

Contextual and corroborating sources (not coded).

Daily Pakistan. (2026, June 20). بھارت سے آئے سکھ یاتری مذہبی رسومات کے بعد واپس روانہ [Sikh pilgrims from India return after religious rites]. *Daily Pakistan*. <https://dailypakistan.com.pk/20-Jun-2026/1939388>

Deutsche Welle (Hindi). (2026, June 25). क्षेत्रीय तनाव के बीच पाकिस्तान बढ़ा सकता है रक्षा खर्च [Pakistan may raise defence spending amid regional tension]. *Webdunia Hindi*. https://hindi.webdunia.com/latest-deutsche-welle-news/pakistan-defence-budget-2026-increase-india-tension-drone-warfare-imf-126062500002_1.html

Express. (2026, June 23, approx.). भारत की ड्रामائی عسकری بزیمت [India's dramatic military humiliation]. *Express*. <https://www.express.pk/story/2819642/bharat-ki-dramai-askari-hazeemat-2819642>

JBT News. (2026, June 28, approx.). क्या भारत-पाकिस्तान के बीच हुई थी सीक्रेट मीटिंग? विदेश मंत्रालय ने किया बड़ा खुलासा [Was there a secret India–Pakistan meeting? Ministry of External Affairs clarifies]. *JBT News*. <https://www.thejbt.com/india/india-pakistan-secret-meeting-in-colombo-mea-vikram-misri-denies-official-track-2-diplomacy-news-313673>

Juraat. (2026, June 27). پاکستان اور انڈیا کے درمیان کولمبو میں بالواسطہ مذاکرات کی خبریں [Reports of indirect Pakistan–India talks in Colombo]. *Juraat*. <https://juraat.com/27/06/2026/165414/>

Mashriq Kashmir. (2026, June, approx.). [Delhi distances itself from Track-Two contacts]. *Mashriq Kashmir*. <https://mashriqkashmir.com/archives/96697>

Mashriq Kashmir. (2026, July, approx.). [‘Position on the suspension of the Indus Waters Treaty maintained’]. *Mashriq Kashmir*. <https://mashriqkashmir.com/archives/96848>

Raksha Samachar. (2026, June, approx.). India-Pakistan Track-2 Talks: कोलंबो ट्रैक-2 मीटिंग [India–Pakistan Track-2 talks: The Colombo Track-2 meeting]. *Raksha Samachar*.

<https://www.rakshasamachar.com/geopolitics/india-pakistan-track-2-talks-colombo-bangkok-ram-madhav/>

Post-window episode (excluded from counts, noted in text).

AajTak. (2026, July 1). बिलावल भुट्टो को फिर भारत पर आया गुस्सा, सिंधु के पानी पर दे दी परमाणु धमकी [Bilawal Bhutto angry at India again, issues nuclear threat over Indus waters]. *AajTak*.

<https://www.aajtak.in/world/story/indus-water-treaty-bilawal-bhutto-zardari-india-pakistan-relations-nuclear-doctrine-ntc-mnrd-dskc-2591584-2026-07-01>