Spektrum Iran | Jg. 38 | Nr. 2 | Sommer und Herbst-2025 | pp. 143-157 |

Original-Forschungsarbeit

Gefühlsasymmetrien in der KI: Sentiment-bias zwischen Englisch und Persisch in harmonisierten LLM-Pipelines

Michael W. Totaro[1], **Leila Gheisi**[2], **Ehsan Shahghasemi**[3]*

[1] *Professor, Fachbereich Informatik, Universität Louisiana at Lafayette, USA*

[2] *Doktorandin in Kommunikation, University of Louisiana at Lafayette, USA*

[3] *Außerordentlicher Professor, Department of Communication, Universität Teheran, Iran*

Empfangen: 5. März 2025 **Akzeptiert:** 8. Juni 2025

Zusammenfassung:

Diese Studie untersucht, wie Sprache die Sentiment-Klassifikation in Ausgaben eines multilingualen großen Sprachmodells (LLM) namens Grok beeinflusst. Basierend auf Langdon Winners Theorie der technologischen Politik, die besagt, dass Technologien inhärent nicht neutral sind und strukturelle Verzerrungen einbetten, wird geprüft, ob Sentiment-Verteilungen auch bei einer vollständig harmonisierten Analysepipeline systematisch zwischen Sprachen variieren. Die Analyse basiert auf einem Korpus von 4.799 Beiträgen (Englisch: n = 2.399; Persisch: n = 2.400), die mit identischen Aufforderungen erzeugt wurden. Sentiment-Ausgaben wurden auf ein gemeinsames dreistufiges Schema (Negativ, Neutral, Positiv) abgebildet, wobei sowohl diskrete Klassenzuweisungen als auch kontinuierliche Wahrscheinlichkeitswerte berücksichtigt wurden. Strukturelle Merkmale wie Satz-, Wort- und Zeichenanzahl wurden berechnet und als Kontrollvariablen einbezogen, um oberflächliche textuelle Unterschiede zu berücksichtigen. Die Ergebnisse zeigen eine deutliche sprachübergreifende Divergenz in Sentiment-Mustern. Englische Ausgaben konzentrieren sich überwiegend auf Neutralität und weisen eine vergleichsweise geringere affektive Intensität auf, während persische Ausgaben eine starke Verschiebung hin zu positivem Sentiment und größere Streuung zeigen. Diese Unterschiede bleiben auch nach Kontrolle struktureller Merkmale statistisch signifikant, was nahelegt, dass die Sprachzugehörigkeit und nicht Textlänge oder Segmentierung der Hauptfaktor für die beobachteten Sentiment-Unterschiede ist. Auf Wahrscheinlichkeitsniveau zeigen englische Verteilungen eine engere Konzentration nahe Neutralität, während persische Verteilungen flacher und stärker positiv verzerrt sind, mit höheren Intensitätswerten. Diese Ergebnisse haben wichtige Implikationen für mehrsprachige Sentiment-Analysen und LLM-Audits. Ohne explizite Modellierung und Kalibrierung von Spracheffekten könnten vergleichende Analysen sprachliche Variation mit affektiver Absicht verwechseln, was zu verzerrten Schlussfolgerungen über Ton, Haltung oder emotionale Valenz führt. Die Studie betont die Bedeutung der Berichterstattung sowohl von Label- als auch Wahrscheinlichkeitsmetriken, die Anwendung sprachspezifischer Kalibrierungsprotokolle und die Berücksichtigung von Sprache als primäre Messdimension in der sprachübergreifenden Inhaltsanalyse.

Schlüsselwörter: sentiment-analyse, mehrsprachige NLP, sprachbias, große sprachmodelle, sprachübergreifender vergleich

* Korrespondierender Autor ✉ shahghasemi@ut.ac.ir https://orcid.org/0000-0002-8716-5806

Wie dieser Artikel zu zitieren ist:
Totaro, M.W., Gheisi, L., & Shahghasemi, E. (2025). Affective asymmetries in AI: Sentiment bias between English and Persian in harmonized LLM pipelines. *Spektrum Iran*, 38(2), 143-157.
https://doi.org/10.22034/spektrum.2026.563602.1052

اشپکتروم ایران | جلد ۳۸ | شماره ۲ | تابستان و پاییز ۱۴۰۴ | ۱۵۷-۱۴۳

مقاله پژوهشی

عدم تقارن‌های عاطفی در هوش مصنوعی: سوگیری احساسی بین زبان‌های انگلیسی و فارسی در جریان‌های یکپارچهLLM

مایکل و. توتارو[۱]، لیلا غیثی[۲]، احسان شاه‌قاسمی[۳]*

۱ استاد، دپارتمان علوم کامپیوتر و انفورماتیک، دانشگاه لوئیزیانا در لافایت، ایالات متحده

۲ دانشجوی دکتری ارتباطات، دانشگاه لوئیزیانا در لافایت، ایالات متحده

۳ دانشیار، دانشکده ارتباطات، دانشگاه تهران، تهران، ایران

دریافت: ۱۴۰۳/۱۲/۱۵ **پذیرش:** ۱۴۰۴/۳/۱۸

چکیده:

این مطالعه بررسی می‌کند که چگونه زبان بر دسته‌بندی احساسی در خروجی‌های تولیدشده توسط یک مدل زبانی بزرگ چندزبانه (LLM) به نام Grok تأثیر می‌گذارد. با تکیه بر نظریه سیاست‌های تکنولوژیک لانگدون وینر، که معتقد است فناوری‌ها به‌طور ذاتی خنثی نیستند و سوگیری‌های ساختاری را در خود جای می‌دهند، این پژوهش بررسی می‌کند که آیا توزیع‌های احساسی حتی در یک جریان تحلیلی کاملاً یکپارچه به‌طور سیستماتیک بین زبان‌ها متفاوت است یا خیر. تحلیل بر اساس یک مجموعه داده شامل ۴۰۷۹۹ پست) انگلیسی$n = 2,399$:؛ فارسی ($n = 2,400$: تولیدشده با استفاده از درخواست‌های یکسان انجام شد. خروجی‌های احساسی به یک طرح سه‌رده‌ای مشترک (منفی، خنثی، مثبت) نگاشت شدند و تحلیل‌ها شامل هر دو معیار انتساب دسته و امتیازهای احتمال پیوسته بود. برای کنترل تفاوت‌های سطحی متنی، ویژگی‌های ساختاری شامل تعداد جملات، کلمات و کاراکترها محاسبه و به‌عنوان متغیر کنترل وارد تحلیل شد. نتایج نشان‌دهنده یک تفاوت بین‌زبانی قوی در الگوهای احساسی است. خروجی‌های انگلیسی عمدتاً در رده خنثی متمرکز هستند و شدت عاطفی نسبتاً کمتری دارند، در حالی که خروجی‌های فارسی گرایش قوی به احساس مثبت همراه با پراکندگی بیشتر نشان می‌دهند. مهم این که، این تفاوت‌ها حتی پس از کنترل ویژگی‌های ساختاری نیز از نظر آماری معنادار باقی می‌مانند و نشان می‌دهند که وابستگی به زبان، نه طول متن یا بخش‌بندی آن، مهم‌ترین عامل مرتبط با تغییرات احساسی مشاهده‌شده است. در سطح احتمال، توزیع‌های انگلیسی تمرکز بیشتری نزدیک به خنثی دارند، در حالی که توزیع‌های فارسی مسطح‌تر و به سمت مثبت متمایل هستند و شاخص‌های شدت بالاتری دارند. این یافته‌ها پیامدهای مهمی برای تحلیل احساسی چندزبانه و بررسی مدل‌های LLM دارند. در صورتی که اثرات زبانی به‌صراحت مدل‌سازی و کالیبره نشوند، تحلیل‌های مقایسه‌ای ممکن است تفاوت‌های زبانی را با قصد عاطفی اشتباه بگیرند و منجر به برداشت‌های تحریف‌شده درباره لحن، موضع یا ارزش عاطفی شوند. این مطالعه اهمیت گزارش هر دو معیار برچسب و احتمال، اتخاذ پروتکل‌های کالیبراسیون مخصوص زبان و در نظر گرفتن زبان به‌عنوان یک بعد اندازه‌گیری درجه‌اول در تحلیل محتوا میان‌زبانی را تأکید می‌کند.

واژگان کلیدی: تحلیل احساسات، پردازش زبان طبیعی چندزبانه، سوگیری زبانی، مدل‌های بزرگ زبان، مقایسه بین‌زبانی

* نویسنده مسئول
https://orcid.org/0000-0002-8716-5806
shahghasemi@ut.ac.ir
https://doi.org/10.22034/spektrum.2026.563602.1052

Spektrum Iran | Vol. 38 | Issue 2 | Summer & Autumn 2025 | pp. 143-157 |

Original Research Paper

Affective asymmetries in AI: Sentiment bias between English and Persian in harmonized LLM pipelines

Michael W. Totaro[1], **Leila Gheisi**[2], **Ehsan Shahghasemi**[3]*

[1] *Department of Computer Science & Informatics, University of Louisiana at Lafayette, USA*
[2] *PhD student in Communication, University of Louisiana at Lafayette, USA*
[3] *Associate Professor, Department of Communication, The University of Tehran, Tehran, Iran*

Received: Mar. 05, 2025 **Accepted:** Jun. 08, 2025

Abstract

In the contemporary landscape of crisis management, decision-makers are increasingly overwhelmed by the sheer volume, velocity, and variety of media data generated during emergencies. Traditional manual analytical methods are often insufficient to process this influx effectively, necessitating a paradigm shift toward advanced computational approaches. The primary goal of this study is to bridge the gap between technical data science and practical crisis communication by establishing a clear analytical link between specific machine learning (ML) paradigms and their operational capabilities. This article utilizes a narrative review methodology, underpinned by a theoretical framework grounded in machine learning. The study systematically synthesizes existing literature to categorize and analyze how distinct ML architectures—specifically supervised, unsupervised, and deep learning—are applied within the domain of media data analysis to support decision-making processes during crises. The analysis confirms that artificial intelligence significantly enhances crisis management effectiveness by automating media monitoring and generating actionable real-time insights. The findings delineate specific roles for different algorithms: supervised learning serves as the theoretical foundation for rapid misinformation detection and precise crisis classification. Conversely, unsupervised learning and deep learning are identified as critical tools for detecting data anomalies and recognizing emerging patterns, which are essential for the functionality of proactive early warning systems. While AI offers transformative potential, this study provides a critical reflection on significant implementation challenges. It highlights the "black box" problem—characterized by a lack of algorithmic interpretability—and inherent data biases as major ethical hurdles that can compromise accountability and fairness in crisis response. The present study contributes a structured framework for understanding AI's role through a theoretical lens. It concludes that future implementation must prioritize explainable AI to balance computational efficiency with ethical responsibility.

Keywords: artificial intelligence, media data analysis, crisis management, crisis communication, machine learning

* Corresponding Author ✉ shahghasemi@ut.ac.ir https://orcid.org/0000-0002-8716-5806

How to Cite this Article:
Totaro, M.W., Gheisi, L., & Shahghasemi, E. (2025). Affective asymmetries in AI: Sentiment bias between English and Persian in harmonized LLM pipelines. *Spektrum Iran*, 38(2), 143-157.
https://doi.org/10.22034/spektrum.2026.563602.1052

1. Introduction

Bias in communication has been a central concern across centuries of scholarly inquiry, and the historical debate around this topic reflects persistent anxieties about the accuracy, fairness, and intent behind how messages are transmitted and received. From classical rhetoric to contemporary media studies, scholars have long examined the role of communicative agents, languages, mediums, and institutions in shaping the content and reception of messages. In early rhetorical traditions, figures like Aristotle explored the persuasive functions of ethos, pathos, and logos, which implicitly acknowledge the subjective framing that can influence how information is presented and interpreted. The rise of print culture, the evolution of mass media, and now the proliferation of algorithmically mediated forms of communication have only deepened interest in understanding how structural, cultural, and technological conditions contribute to communication biases. Across disciplines—from sociology and political science to media studies and linguistics—bias in communication is studied as both a deviation from neutrality and as a mechanism through which power and ideology are encoded and reproduced (Entman, 2007; van Dijk, 1993).

Bias is not merely a failure of impartiality; it is often embedded in the architectures of communication itself. As early as the mid-20th century, communication scholars began interrogating how different media channels structure the conditions for discourse, privileging certain temporal or spatial biases depending on their material and institutional configurations. A particularly influential contribution came from Canadian theorist Harold Innis, whose concept of "bias" in communication technologies laid the groundwork for later media ecology frameworks. Innis (1951) proposed that communication media exhibit either a bias toward time or space. Time-biased media, such as stone tablets or oral traditions, facilitate the preservation of culture across generations but are limited in their spatial reach. Space-biased media, such as paper or digital text, enable communication across large distances but tend to be ephemeral and less durable over time. Innis's framework emphasizes that media are not neutral conduits; they condition what kinds of knowledge are preserved, circulated, and prioritized, thus shaping the trajectory of civilizations. His work highlighted how the very format of a medium predisposes it to certain political and cultural effects.

It should be noted that, For Harold Innis, the concept of "bias" in communication differed significantly from the later, more politicized understandings of bias found in Marxist or critical theory traditions. Innis's use of the term was less concerned with ideological favoritism or systematic discrimination against particular social groups, and more focused on the structural and temporal properties of communication media. His notion of bias referred to the way certain media favor the transmission of knowledge across either time or space, shaping the stability and expansion of civilizations accordingly.

Extending Harold Innis's foundational contributions, subsequent scholars elaborated the analysis of communicative bias across more specific empirical and theoretical domains. Marshall McLuhan, a student of Innis, famously asserted that "the medium is the message," foregrounding the claim that the affordances and constraints of communication technologies generate effects that exceed the semantic content they convey (McLuhan, 1964). Later theorists such as Pierre Bourdieu (1991) shifted attention to the operation of symbolic power through language itself, arguing that linguistic capital and habitus structure not only what can be said, but also who is authorized to speak and how utterances are received. In a complementary vein, Norman Fairclough (1995) developed critical discourse analysis to expose how ideological formations are reproduced through texts that appear ostensibly neutral, while Teun van Dijk (1993) examined the manifestation of systemic racism in media representations and political discourse. Taken together, these lines of inquiry converge on a common premise: communication is never free from bias, and such bias is frequently structurally embedded, culturally sustained, and technologically mediated.

Within this intellectual lineage, Langdon Winner (1980) introduced a critical refinement by explicitly theorizing the politics of artifacts, with particular attention to communication technologies. Winner contended that technologies are not neutral instruments but material instantiations of power relations and social organization. By posing the provocative question "Do artifacts have politics?" he argued affirmatively that certain technologies are intrinsically political, either because they presuppose or reinforce specific social arrangements or because their effects systematically advantage particular groups. Applied to communication, Winner's framework suggests that technologies do more than transmit messages: they actively configure

the conditions of expression, access, and interpretation. This perspective reorients analysis away from content and toward infrastructure, shifting emphasis from discourse to design. Such an approach is especially salient in digital media environments, where algorithms, interfaces, and data architectures increasingly govern what information becomes visible, authoritative, or amplified.

Winner's theory of technological politics dovetails with contemporary concerns about algorithmic bias and automated decision-making. In the era of large language models (LLMs), the processes by which information is classified, ranked, and labeled are often opaque, raising questions about how technological infrastructures encode preferences, assumptions, or systemic inequalities. Scholars such as Noble (2018) have shown how search engines reproduce racial and gendered stereotypes, while Eubanks (2018) documented how algorithmic systems used in public services often penalize marginalized populations. These critiques underscore Winner's insight that technologies, particularly those used in communication, are not apolitical—they are designed and deployed within specific power structures that shape their effects.

Artificial intelligence, particularly in the domain of language modeling and sentiment classification, introduces subtle yet consequential forms of bias into communication (Sabbar & Habib Zadeh Khiyaban, 2023). As evidenced in the present study, large language models (LLMs) like Grok exhibit systematic affective divergence across languages even when analytic pipelines are fully harmonized. These divergences—manifested as higher positivity and greater sentiment intensity in Persian compared to English—persist after controlling for structural variables such as word or sentence count, indicating that the bias is not merely a function of input length or segmentation, but of deeper linguistic, script, or model calibration effects. Such asymmetries highlight how AI systems, through their training data, tokenization mechanisms, and classifier thresholds, do not neutrally interpret or replicate linguistic inputs; instead, they encode and reproduce culturally contingent affective norms that shape interpretive outcomes. This phenomenon aligns with Langdon Winner's (1980) theory of technological politics, which posits that artifacts—including computational models—embody social and political values and can enforce specific regimes of interpretation depending on their embedded assumptions and affordances.

The risk, therefore, is not only technical misclassification but also epistemological distortion: AI-driven communication systems may systematically reshape how tone, sentiment, and stance are understood across linguistic groups (Salehi et al., 2026). For instance, as Noble (2018) and Eubanks (2018) have argued in adjacent contexts, algorithmic systems often reflect and exacerbate structural inequalities by encoding dominant cultural perspectives into seemingly objective technologies. In multilingual sentiment analysis, this dynamic manifests when affective intensity is over- or under-represented based on language-specific priors, potentially skewing comparative analyses in media studies, public opinion research, or international communication scholarship (Hanna et al., 2025; Pessach & Shmueli, 2022; Tejani et al., 2024; Shahghasemi, 2025). Without robust calibration protocols that treat language as a first-order measurement dimension, LLMs risk privileging affective neutrality in one language (e.g., English) while amplifying positivity in another (e.g., Persian), not due to actual differences in tone but due to model artifacts.

The study of sentiment analysis provides a concrete example of how communicative technologies can introduce or exacerbate bias (Venkit & Wilson, 2021; Thelwall, 2018; Díaz et al., 2018; Bhanvadia et al., 2024; Venugopal et al., 2024; Rozado, 2020; Shahghasemi et al., 2025; Radaideh et al., 2025). Automated sentiment classifiers, especially those powered by LLMs, are trained on corpora that reflect existing linguistic, cultural, and ideological biases. As such, these models may encode not only lexical or syntactic patterns but also implicit judgments about tone, emotion, and stance. When applied across languages, such models face the added challenge of cross-linguistic variation in pragmatics, grammar, and script. For instance, a model trained on English may interpret hedging, politeness, or affirmation differently than one trained on Persian, leading to divergent sentiment classifications even when the underlying messages are equivalent in intent or tone. This raises profound implications for comparative media research, public opinion analysis, and digital governance.

Indeed, cross-lingual disparities in sentiment analysis may not simply reflect differences in expression but rather artifacts of the technological pipeline itself. This aligns with Winner's contention that technological systems can enforce or conceal forms of bias depending on how they are structured and operationalized. In multilingual applications of LLMs,

differences in tokenization, script (e.g., Latin vs. Perso-Arabic), and classifier calibration can produce systematic differences in affective outputs, even when inputs are semantically aligned. Thus, communication technologies—particularly those powered by machine learning—do not merely mediate human expression; they participate in shaping it, often in subtle and consequential ways.

Building on Winner's theory, the present study follows this tradition of inquiry by empirically investigating how language-specific characteristics in large language model outputs may introduce bias into sentiment analysis. Specifically, the research examines the outputs of a multilingual LLM (Grok) when prompted to generate content in English (EN) and Persian (FA). By harmonizing the sentiment classification scheme across both languages—reducing native sentiment labels to a common three-class system (Negative/Neutral/Positive)—the study isolates language membership as a potential driver of divergent sentiment outcomes. The design controls for structural features of the text (such as sentence and word counts) to disentangle linguistic from purely formal differences.

This work is guided by a series of research questions that probe the relationship between language and sentiment classification in a harmonized analytic pipeline:

RQ1. Do structural features of Grok's outputs (sentences, words, characters per post) differ between English and Persian?

RQ2. Holding the label space constant (Negative/Neutral/Positive), does the distribution of sentiment classes differ by language?

RQ3. Are the mean sentiment probabilities P(Negative), P(Neutral), P(Positive) systematically different across languages?

RQ4. Does language (EN vs. FA) predict sentiment outcomes under a harmonized pipeline?

RQ5. Do confidence characteristics (e.g., the allocation of probability mass toward neutrality vs. positivity) vary by language in ways that could bias interpretation?

RQ6. Are observed differences substantively meaningful in terms of effect sizes, beyond mere statistical significance, for cross-language comparisons of tone?

RQ7. After accounting for structural text differences, do language-based

differences in sentiment persist, indicating a genuine cross-lingual shift rather than an artifact of message length or segmentation?

2. Methodology

We constructed a time-bounded corpus of Grok outputs in two languages, English (EN) and Persian (FA), generated with identical prompts within a single day. After removing empty and duplicate entries, we retained balanced samples of roughly 2,400 posts per language. Preprocessing served two purposes. First, to characterize text structure, we computed sentence, word, and character counts directly from the raw text so that punctuation and segmentation cues were preserved; only light normalization (e.g., link removal, whitespace cleanup) was applied to avoid altering length or punctuation patterns. Second, to characterize sentiment, both corpora were aligned to a common three-class space (Negative/Neutral/Positive). English posts were scored with a standard three-class transformer; Persian posts were scored with a widely used Persian model and then collapsed from five native categories to the same three classes (Negative = Furious+Angry; Neutral = Neutral; Positive = Happy+Delighted). For both languages, we retained per-class probabilities and the hard label (argmax), and we derived an intensity index defined as 1−P(Neutral).

To ensure cross-lingual comparability, we held the label space, decision rules, and output format constant across EN and FA. All subsequent analyses apply the same procedures to both samples; this enabled a clean assessment of whether language membership (EN vs. FA) is associated with systematic differences in sentiment under a harmonized pipeline.

3. Findings

Using the harmonized three-class scheme (Negative/Neutral/Positive), we summarized sentiment at both the hard-label and probability levels, and tracked a continuous intensity index. The class composition differs markedly by language. In the English corpus, posts are predominantly Neutral, with 1,637 items (≈68.2%) assigned to that class; Negative and Positive account for 558 (≈23.3%) and 204 (≈8.5%) posts, respectively. In the Persian corpus, the distribution shifts toward the Positive pole: 1,006 posts (≈41.9%) are labeled Positive, 917 (≈38.2%) Neutral, and 477 (≈19.9%) Negative.

Table 1. Cross-lingual disparities in sentiment analysis.

Class	EN Count	EN %	FA Count	FA %
Negative	558	23.3	477	19.9
Neutral	1,637	68.2	917	38.2
Positive	204	8.5	1,006	41.9

The analytic corpus comprises two balanced language sets, English (EN; n = 2,399) and Persian (FA; n = 2,400), processed under a harmonized three-class sentiment scheme (Negative/Neutral/Positive). Descriptive statistics were computed for two families of variables: structural features of the text (number of sentences, words, and characters per post) and sentiment outputs at both the label and probability levels, including a continuous intensity index. All structural measures were derived from the raw message body to preserve punctuation and segmentation cues, and sentiment probabilities are those emitted by the language-specific classifiers after mapping both languages to the common three-class space.

Table 2. Descriptive statistics for structural text features in the English (EN) and Persian (FA) corpora

Metric	EN n	EN M	EN SD	FA n	FA M	FA SD
Sentences per post	2,399	3.04	0.89	2,400	3.49	1.00
Words per post	2,399	42.68	5.62	2,400	50.77	9.21
Characters per post	2,399	268.53	25.46	2,400	259.08	40.98

For structural features, English posts are shorter in sentences and words but slightly longer in characters. English contains on average M = 3.04 sentences per post (SD = 0.89), whereas Persian averages M = 3.49 (SD = 1.00); medians in both languages are 3, and the empirical distributions are discrete with narrow interquartile ranges, especially in English. Word counts follow the same pattern: English averages M = 42.68 words (SD = 5.62), compared with M = 50.77 words (SD = 9.21) in Persian, with a visibly broader spread in

the Persian sample. Character counts invert this relationship: English averages M = 268.53 characters (SD = 25.46) versus M = 259.08 (SD = 40.98) in Persian, with identical medians (276) but heavier tails in Persian. These distributional shapes are consistent with script and tokenization differences: the Persian sample partitions content into more orthographic tokens and sentences without increasing character length proportionally, while English concentrates more tightly around a stable character count.

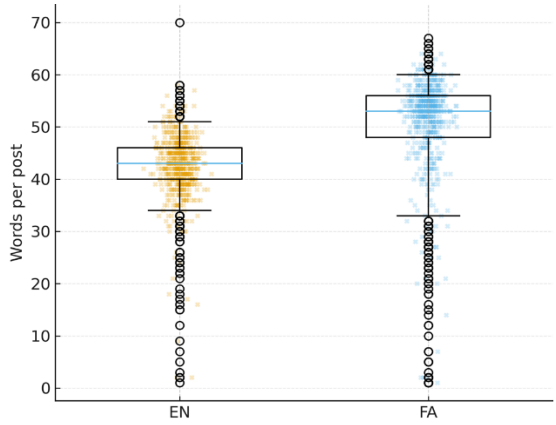

Figure 1. Distribution of words per post in the English (EN) and Persian (FA) corpora, shown as boxplots with overlaid jittered observations

At the label level, the composition of sentiment classes differs markedly between the two languages. In English, Neutral dominates the distribution (1,637 of 2,399; ≈68.2%), with smaller shares for Negative (558; ≈23.3%) and Positive (204; ≈8.5%). In Persian, the mass shifts toward Positive (1,006 of 2,400; ≈41.9%), with Neutral reduced (917; ≈38.2%) and Negative somewhat lower (477; ≈19.9%). These shares are reflected in the underlying probability profiles. The mean neutrality probability is substantially higher in English than in Persian (M = .613, SD = .259 vs. M = .384, SD = .394), indicating that English outputs concentrate probability near the neutral class and do so with comparatively low dispersion. The positive dimension shows the converse: Persian assigns more probability to positivity and exhibits greater spread (M = .414, SD = .409) relative to English (M = .138, SD = .208). On the negative dimension, English carries a modestly higher mean than Persian (M = .249,

SD = .281 vs. *M* = .201, *SD* = .325), while the Persian distribution again shows wider dispersion. Visual diagnostics—including overlaid histograms, boxplots, and violin plots—reinforce the same pattern: English curves are steeper around neutrality, whereas Persian curves are flatter and longer-tailed toward positivity, with more mass away from the neutral center.

The intensity index condenses these probability patterns into a single continuous measure of affective strength. English posts exhibit lower and less variable intensity (*M* = .387, *SD* = .259), while Persian posts are both higher on average and more dispersed (*M* = .616, *SD* = .394). Empirical cumulative distribution functions for intensity and for the maximum class probability (a proxy for classification confidence) display corresponding differences, with English curves rising more steeply (concentration near moderate intensity and higher neutrality) and Persian curves rising more gradually (greater heterogeneity, with more high-intensity, positive-leaning posts).

Table 3. Sentiment probability outputs and the continuous sentiment intensity index in the English (EN) and Persian (FA) corpora.

Metric	EN n	EN M	EN SD	FA n	FA M	FA SD
P(Negative)	2,399	0.249	0.281	2,400	0.201	0.325
P(Neutral)	2,399	0.613	0.259	2,400	0.384	0.394
P(Positive)	2,399	0.138	0.208	2,400	0.414	0.409
Sentiment intensity	2,399	0.387	0.259	2,400	0.616	0.394

Table 4. Between-language differences in structural and sentiment-related measures, reporting mean differences (ΔM = EN − FA), 95% confidence intervals, permutation-test p-values, and effect sizes (Cliff's δ)

Metric	ΔM (EN−FA)	95% CI ΔM	p (perm, mean)	Cliff's δ
Sentences per post	−0.4500	[−0.5000, −0.3900]	< .001	0.000
Words per post	−8.1000	[−8.5400, −7.6600]	< .001	0.007
Characters per post	+9.4600	[+7.5400, +11.4300]	< .001	0.045

Metric	ΔM (EN−FA)	95% CI ΔM	p (perm, mean)	Cliff's δ
P(Negative)	+0.0477	[+0.0306, +0.0650]	< .001	0.302
P(Neutral)	+0.2287	[+0.2098, +0.2475]	< .001	0.332
P(Positive)	−0.2764	[−0.2972, −0.2547]	< .001	−0.299
Sentiment intensity	−0.2287	[−0.2475, −0.2098]	< .001	−0.332

4. Conclusion

Evidence from descriptive profiles and harmonized outputs points to a clear pattern: language systematically shapes Grok's expressed affect. English responses concentrate probability mass around neutrality and, to a lesser extent, negativity; Persian responses shift mass toward positivity and exhibit higher intensity (i.e., lower neutrality probabilities). These differences persist when sentiment is summarized as both hard labels and probability distributions, indicating that the phenomenon is not an artifact of a particular decision rule. In practical terms, this means that direct, cross-language comparisons of tone can be misleading unless the measurement pipeline explicitly accounts for language effects.

Regarding RQ1 (structural variation), English-language posts are marginally shorter in terms of sentences and words but slightly longer when measured by character count, whereas Persian-language posts exhibit greater length and variability at the token level. Importantly, these structural differences are relatively small and do not correspond to the observed direction of sentiment divergence. Persian texts, despite containing more words and sentences, are associated with higher levels of positivity and affective intensity, while English texts—characterized by more compact character distributions—tend toward greater neutrality. This decoupling between structural properties and affective outcomes indicates that superficial measures of text length are unlikely to account for the cross-linguistic sentiment gap.

Addressing RQ2–RQ4 (differences in class distributions and probabilities; language as a predictor), the distributions diverge in straightforward ways: English outputs cluster in Neutral, Persian in Positive. Because the label space and decision rules were held constant across languages, the most

plausible explanations are (a) cross-lingual differences in pragmatics (how affirmation, hedging, and stance are expressed); (b) script and tokenization effects that shape model attention and priors; and (c) model calibration differences across the English and Persian classifiers. The intensity profiles reinforce this interpretation: Persian's lower neutrality and broader spreads are consistent with a classifier (and discourse register) that distributes affect more decisively away from the center, while English appears more tightly centered around neutrality.

Regarding RQ5 (confidence characteristics), English exhibits steeper concentration near the neutral probability region, whereas Persian displays broader, more heterogeneous probability mass, especially on the positive pole. This asymmetry has interpretive consequences: confidence thresholds calibrated to English neutrality distributions will overstate "decisiveness" in English and understate it in Persian, unless calibrated separately.

Turning to RQ6 (substantive meaning of differences), the size and stability of the gaps visible at both the label and probability levels are large enough to matter for communication research. In audits of LLM behavior, newsroom analytics, or comparative political communication, a neutrality-skew in English versus a positivity-skew in Persian can produce different storylines about tone even when the underlying prompts are identical. Substantively, the safest reading is that language membership is not just a nuisance covariate; it is a consequential measurement dimension in its own right.

For RQ7 (persistence after accounting for structure), the persistence of the affective gap alongside only small structural differences argues against a length-based explanation. Even when one controls analytically for sentences/words/characters by examining distributions conditional on length bands or by focusing on probability space rather than hard labels, the neutrality-versus-positivity contrast remains the defining axis across languages. This points to linguistic and modeling factors rather than superficial length as the primary sources of divergence.

Implications. For multilingual content analysis, three practices are advisable. First, report both hard labels and probability summaries, and predefine language-specific calibration checks (e.g., reliability curves, threshold sensitivity). Second, adopt harmonized label spaces and decision rules, but allow post-hoc calibration by language (temperature/Dirichlet

scaling), so neutrality is not implicitly privileged in one language. Third, incorporate measurement-invariance diagnostics, e.g., replicate key contrasts using a multilingual model or a translate-to-pivot approach with human adjudication on a stratified subsample to distinguish linguistic pragmatics from model artifacts.

Taking the research questions together, structural differences between English and Persian are modest and do not account for the affective gap; the distribution of sentiment classes and the underlying probability profiles differ systematically by language; and language membership functions as a meaningful predictor of sentiment outcomes under a harmonized pipeline. For scholars and practitioners, the practical upshot is straightforward: in multilingual analyses of LLM outputs, language must be modeled, calibrated, and reported as a first-order measurement dimension, not treated as noise to be averaged away.

References

Bhanvadia, S., Radha Saseendrakumar, B., Guo, J., Spadafore, M., Daniel, M., Lander, L., & Baxter, S. L. (2024). Evaluation of bias and gender/racial concordance based on sentiment analysis of narrative evaluations of clinical clerkships using natural language processing. *BMC medical education*, *24*(1), 295. https://doi.org/10.1186/s12909-024-05271-y

Bourdieu, P. (1991). *Language and symbolic power* (J. B. Thompson, Ed.; G. Raymond & M. Adamson, Trans.). Harvard University Press.

Díaz, M., Johnson, I., Lazar, A., Piper, A. M., & Gergle, D. (2018, April). Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1-14). https://doi.org/10.1145/3173574.3173986

Entman, R. M. (2007). Framing bias: Media in the distribution of power. *Journal of Communication*, *57*(1), 163–173. https://doi.org/10.1111/j.1460-2466.2006.00336.x

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Fairclough, N. (1995). *Critical discourse analysis: The critical study of language*. Longman.

Hanna, M. G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., ... & Rashidi, H. H. (2025). Ethical and bias considerations in artificial intelligence/machine learning. *Modern Pathology*, *38*(3), 100686. https://doi.org/10.1016/j.modpat.2024.100686

Innis, H. A. (1951). *The bias of communication*. University of Toronto Press.

McLuhan, M. (1964). *Understanding media: The extensions of man*. McGraw-Hill.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, *55*(3), 1-44. https://doi.org/10.1145/3494672

Radaideh, M. I., Kwon, O. H., & Radaideh, M. I. (2025). Fairness and social bias quantification in Large Language Models for sentiment analysis. *Knowledge-Based Systems*, 113569. https://doi.org/10.1016/j.knosys.2025.113569

Rozado, D. (2020). Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. *PloS one*, *15*(4), e0231189. https://doi.org/10.1371/journal.pone.0231189

Sabbar, S., & Habib Zadeh Khiyaban, S. (2023). Algorithms of Displacement: Emotional and Rhetorical Responses to AI-Driven Job Loss in Digital Public Discourse. *International Journal of Advanced Multidisciplinary Research and Studies, 3*(4), 1324-1331. https://doi.org/10.62225/2583049X.2023.3.4.5012

Salehi, K., Habib Zadeh Khiyaban, S., Sabbar, S. (2026). Artificial Intelligence and Crime Detection: A Critical Review. *Cyberspace Studies. 10*(1): 181-197. https://doi.org/10.22059/jcss.2025.402206.1179

Shahghasemi, E. (2025). AI; A Human Future. *Journal of Cyberspace Studies*, *9*(1), 145-173. doi: 10.22059/jcss.2025.389027.1123

Shahghasemi, E., Gholami, F. & Alikhani, Z. (2025). Global patterns of social media use and political sentiment. *Discover Global Society*, 3, 36. https://doi.org/10.1007/s44282-025-00171-y

Tejani, A. S., Ng, Y. S., Xi, Y., & Rayan, J. C. (2024). Understanding and mitigating bias in imaging artificial intelligence. *Radiographics*, *44*(5), e230067. https://doi.org/10.1148/rg.230067

Thelwall, M. (2018). Gender bias in sentiment analysis. *Online Information Review*, *42*(1), 45-57. https://doi.org/10.1108/OIR-05-2017-0139

van Dijk, T. A. (1993). Principles of critical discourse analysis. *Discourse & Society*, *4*(2), 249–283. https://doi.org/10.1177/0957926593004002006

Venkit, P. N., & Wilson, S. (2021). Identification of bias against people with disabilities in sentiment analysis and toxicity detection models. *arXiv preprint arXiv:2111.13259*. https://doi.org/10.48550/arXiv.2111.13259

Venugopal, J. P., Subramanian, A. A. V., Sundaram, G., Rivera, M., & Wheeler, P. (2024). A Comprehensive Approach to Bias Mitigation for Sentiment Analysis of Social Media Data. *Applied Sciences*, *14*(23), 11471. https://doi.org/10.3390/app142311471

Winner, L. (1980). Do artifacts have politics? *Daedalus*, *109*(1), 121–136. http://www.jstor.org/stable/20024652