Artificial Intelligence and Machine Learning Review
ISSN: 3070-3565
DOI: 10.69987/AIMLR.2025.60304

Cross-lingual Robustness of Behavioral Coordination Signals for Detecting State-Linked Information Operations on Twitter: A Comparative Empirical Study

Shang Wen
International Affairs, Science and Technology, Georgia Institute of Technology, USA

Keywords

coordinated inauthentic behavior; cross-lingual transfer; information operations; user similarity networks

Abstract

Coordinated inauthentic behavior on social media has become a primary vector through which state-linked information operations seek to shape public political discourse. Computational detection work to date has been built and validated mainly on English-language operations, leaving open whether the behavioral coordination signals the field relies on—co-retweet, co-hashtag, co-URL, temporal synchronization, and textual similarity—generalize to non-English political environments. This study presents a comparative empirical assessment of the cross-lingual robustness of these five canonical signals using two publicly released Twitter datasets: 3,841 IRA accounts (English-dominant, 2018 release) and 23,750 PRC-attributed accounts (Chinese-dominant, June 2020 takedown). User similarity networks are constructed following the framework standardized in prior literature, the textual-similarity backbone is swapped among mBERT, XLM-R, and LaBSE, and unsupervised detection quality is measured against time- and topic-matched organic control samples. Co-retweet and temporal synchronization transfer most reliably across languages, with F1 gaps under 0.05; textual similarity drops by 0.12 on Chinese data even with multilingual encoders. The findings offer empirical guidance for cross-border information operation monitoring and reveal the unequal language coverage of current detection toolchains. The contribution is empirical rather than methodological, the scope is restricted to two operations on a single platform, and the results are intended as a calibration baseline for English-trained pipelines redeployed in Chinese-language political environments.

1. Introduction

1.1 Background and Motivation

Social media now serves as a primary venue for political communication, and the same openness that supports civic exchange also creates room for coordinated manipulation at scale. Coordinated inauthentic behavior refers to networks of accounts acting in concert to amplify or distort narratives while keeping the coordination itself hidden, and it has drawn sustained research attention as a threat to information integrity online. Network-based detection methods build user similarity graphs from shared low-level actions such as co-retweet, co-hashtag, co-URL, and synchronized posting, and identify clusters whose joint behavior departs from an organic baseline defined by null models or matched controls [1][2]. This framework has been applied across a wide range of case studies and platforms over the past five years, and the field has settled on a small set of canonical behavioral signals that now form the operational core of cross-platform monitoring.[3][4]

Cases of inauthentic political amplification have been documented across a wide range of political systems, and similar operational signatures recur even when the targets and language environments change [5]. More recent work shows that coordinated communities behave less like static actors and more like evolving organizations, with archetypes and influence that shift over the lifecycle of political events [6]. Even with this maturation, the empirical base of the field is still tilted heavily toward English. The most extensively studied case is the Russian Internet Research Agency (IRA) campaign that Twitter released in October 2018, containing 3,841 accounts and around 8.7 million tweets, the bulk of them in English. Subsequent public takedowns have covered seventeen countries and more than 200 million tweets, yet Chinese-language operations have so far received comparatively little algorithmic scrutiny.

1.2 Research Gap in Cross-lingual Coordinated Behavior Detection

The largest Chinese-language disclosure to date is the June 2020 PRC takedown, which covered 23,750 core accounts pushing content on the Hong Kong protests, the Guo Wengui dissident narrative, and the early COVID-19 period. This English-versus-Chinese imbalance is consequential because the dominant detection signals were built and tuned on English data, and posting rhythms, hashtag use, retweet conventions, URL ecosystems, and tokenization all differ enough between the two language environments that clean transfer cannot be assumed. Practical monitoring of cross-border information operations depends on knowing which signals survive that shift and which do not. Without such calibration, English-trained pipelines redeployed in Chinese-speaking online publics will tend to under-detect coordinated activity, with downstream consequences for threat assessment and for operational prioritization in multilingual analytic settings.[7][8]

1.3 Research Questions and Contributions of This Study

No new detection model is introduced here. The goal is to take the five canonical behavioral signals and measure how their detection quality changes when the same pipeline is applied uniformly to a matched English operation and a matched Chinese operation, with controls drawn from time- and topic-aligned organic discourse. Only the textual-similarity component varies across runs, where the backbone encoder is swapped among three publicly released multilingual models while every other parameter is held fixed. The work contributes a per-signal cross-lingual characterization that records which signals carry across languages and which do not, an encoder ablation that measures how much of the text-signal gap stronger multilingual representations can recover, and a discussion of the linguistic and platform-cultural sources behind the residual gap. Two operations and a single platform make for a deliberately narrow evidence base, the pipeline is the standard one in the field rather than a new architecture, and the claims are bounded to the signals actually tested. What the results offer is a calibration baseline for practitioners porting English-trained detection tools into non-English political environments, not a definitive ranking of detection methods.[9][10]

2. Related Work

2.1 Network-Based Detection of Coordinated Inauthentic Behavior

Computational research on coordinated inauthentic behavior has developed along two complementary tracks. The network-signature track focuses on the structural traces of coordination, building user similarity graphs from shared low-level actions and pruning the result down to dense components that exceed a null-model baseline. Pacheco, Flammini, and Menczer illustrated this strategy on the White Helmets disinformation campaign and laid out the bipartite-projection-and-thresholding pipeline that later studies have adapted and extended [11][12]. Nizzoli and colleagues extended the framework to multi-action settings in their analysis of the 2019 UK general election, finding that joint co-retweet and rapid-retweet views surface coordination patterns missed by either view alone, and they released a benchmark of 11.26 million tweet identifiers that has been widely reused since [13][14]. Magelinski, Ng, and Carley proposed a synchronized-action framework anchored on an explicit five-minute window, and argued that

deployment carries ethical constraints because organic activism is at non-trivial risk of being miscategorized as coordinated manipulation [15][16].

2.2 Behavioral Signals and Latent-Process Coordination Models

The second track treats coordination as a latent process to be inferred rather than directly observed. Sharma and colleagues represented account activity as continuous-time event sequences with attentive mixture density networks, relaxed the assumption that coordination must surface as exact action overlap, and recovered hidden influence among IRA-attributed accounts even where literal co-retweets did not occur [17][18]. Work in the security-research community has applied network-analytic methods to map disinformation campaigns at scale, tying retweet-graph statistics to campaign-level indicators of inauthenticity [19][20]. Cross-operation generalization has been the focus of a parallel strand: Gabriel and colleagues trained graph-learning models on operations attributed to one state and evaluated them on held-out operations attributed to other states, finding that the abstract structural signatures of coordinated manipulation carry across campaigns of Russian, Chinese, and Iranian origin [21][22]. Knowledge-informed temporal point-process models such as VigDet combine prior coordination knowledge with neural sequence models and have been validated on COVID-19 misinformation, indicating that semi-supervised configurations remain workable even when account activity records are sparse [23][24].

2.3 Cross-lingual Transfer and the Open Language Gap

Several patterns recur across this body of work: behavioral signals do carry meaningful information about coordination across operations, no single signal dominates every campaign type, and the strength of cross-operation transfer varies substantially. Co-hashtag analyses are sensitive to drift in hashtag conventions and to topical churn during fast-moving political events, and co-URL analyses depend on link ecosystems that themselves vary by language community.[25-27] Temporal-synchronization signals are platform-mechanical rather than linguistic, and prior work has argued that they are the most operation-agnostic of the established set. The language dimension of cross-operation transfer remains the least developed of these strands. Most published evaluations either focus on English-language operations or aggregate across languages without separating per-signal contributions. Cross-lingual representation learning has matured in parallel through pre-trained multilingual encoders that map many languages into a shared semantic space, yet their interaction with the text-similarity component of coordination detection has not been benchmarked systematically in a paired English-versus-Chinese state-linked-operation setting. The work reported here targets that gap directly, running a standard detection pipeline with standard multilingual encoders and quantifying how the established signals transfer to a Chinese-dominant operation under conditions otherwise matched to an English-dominant baseline.[28][29]

3. Data and Methodology

3.1 Datasets and Matched Organic Controls

The English coordination sample comes from the IRA dataset that Twitter released in October 2018 through its Information Operations Hub. That release contains 3,841 accounts attributed to the Internet Research Agency and roughly 8.7 million tweets spanning 2009 through October 2018, dominated by English-language posts with a long tail of Russian-language and other-language tweets. Accompanying fields include account-level metadata, full tweet text, retweet relationships, hashtag and URL columns, and timestamps at second-level granularity. The Chinese coordination sample comes from the June 2020 PRC takedown that Twitter disclosed under its state-backed information operations program, which contains 23,750 core accounts and roughly 348 thousand tweets written mostly in simplified and traditional Chinese, organized around three thematic clusters: the Hong Kong protest movement, the Guo Wengui dissident narrative, and the early phase of the COVID-19 pandemic. Both datasets are publicly distributed with platform-released documentation, and only the public versions are used here, with no enrichment from third-party scrapers or hydration of suspended account histories.

Organic control samples are built to match each coordination sample on time, topic, and language. The English baseline for the IRA sample is drawn from the publicly released #Election2020 collection of over 1.2 billion tweet identifiers covering 2019 through 2021, filtered to thirty-day windows centered on the median IRA posting date for each topic and restricted to accounts that posted in English during the matching window. The Chinese baseline for the PRC sample is drawn from public Twitter collections covering the same Hong Kong and COVID-19 timelines, restricted to Chinese-language accounts active in the matching window. Following the matched-control protocol used in earlier cross-operation evaluation work [30], control accounts are sampled to approximate the IO accounts on tweet-volume and account-age distributions, and any account later flagged in subsequent Twitter takedowns is dropped from the control pool. Per-language sample sizes after filtering are reported in Table 1.

Table 1. Summary statistics of the matched coordination and organic control samples.

Sample	Source	Language	Accounts	Tweets	Date range
IO-EN	IRA (Twitter, 2018)	English-dom.	3,841	≈ 8.7 M	2009–2018
IO-ZH	PRC June 2020 takedown	Chinese-dom.	23,750	≈ 348 K	2017–2020
Ctrl-EN	#Election2020 (filtered)	English	4,200	≈ 1.1 M	2019–2021
Ctrl-ZH	Public ZH baseline	Chinese	4,500	≈ 420 K	2019–2020

Tweet counts in Table 1 refer to post-filter samples, and account counts refer to the matched pool produced after age- and volume-based balancing against the coordinated samples. All four sources are released for academic research under the data-use terms associated with the original publications.

3.2 Coordination Signal Extraction and Network Construction

Five behavioral coordination signals are extracted under identical settings on both languages. Co-retweet connects pairs of users who retweet the same source tweet inside the dataset. Co-hashtag connects pairs that share the same hashtag on at least three occasions. Co-URL connects pairs that share at least one expanded URL after tracking parameters are stripped and the link is normalized to its registered domain. Temporal synchronization connects pairs whose posting timestamps fall inside a five-minute window at a frequency above a chance baseline. Textual similarity connects pairs whose tweet embeddings exceed a cosine threshold of 0.85. For each signal, the user-action bipartite graph is folded into a weighted user-user network through TF-IDF-weighted cosine similarity, and edges below the 99th percentile of the language-specific null distribution are pruned, following the standard procedure in prior network-based coordination work.

Textual similarity is the only component whose language-handling configuration varies across runs. Three multilingual sentence encoders are benchmarked here. mBERT is a 12-layer transformer pre-trained on Wikipedia in 104 languages and serves as the standard early baseline for cross-lingual representation [31][32]. XLM-R is pre-trained on 2.5 terabytes of CommonCrawl text spanning 100 languages, and the original release reports an average gain of 13.8 accuracy points over mBERT on the XNLI cross-lingual natural language inference benchmark [33][34]. LaBSE is a dual-encoder trained with translation-pair contrastive ranking, supports 109 languages, and is explicitly optimized for cross-lingual sentence similarity [35][36]. Embeddings are produced at the tweet level after a standard preprocessing pass that strips URLs, masks user mentions, and lowercases Latin-script text, and the same cosine threshold is applied across encoders to keep their outputs comparable. Signal-extraction parameters appear in Table 2, and the full pipeline is summarized in Figure 1.

Table 2. Behavioral coordination signal extraction parameters.

Signal	Action atom	Threshold	Network folding	Null-model pruning
Co-retweet	shared source tweet ID	≥1 shared retweet	TF-IDF cosine	99th percentile
Co-hashtag	shared hashtag	≥3 shared hashtags	TF-IDF cosine	99th percentile
Co-URL	shared registered domain	≥1 shared URL	TF-IDF cosine	99th percentile
Temporal sync	5-minute posting window	observed > chance	Pointwise mut. info.	99th percentile
Text similarity	cosine of sentence emb.	≥ 0.85	encoder-specific	99th percentile

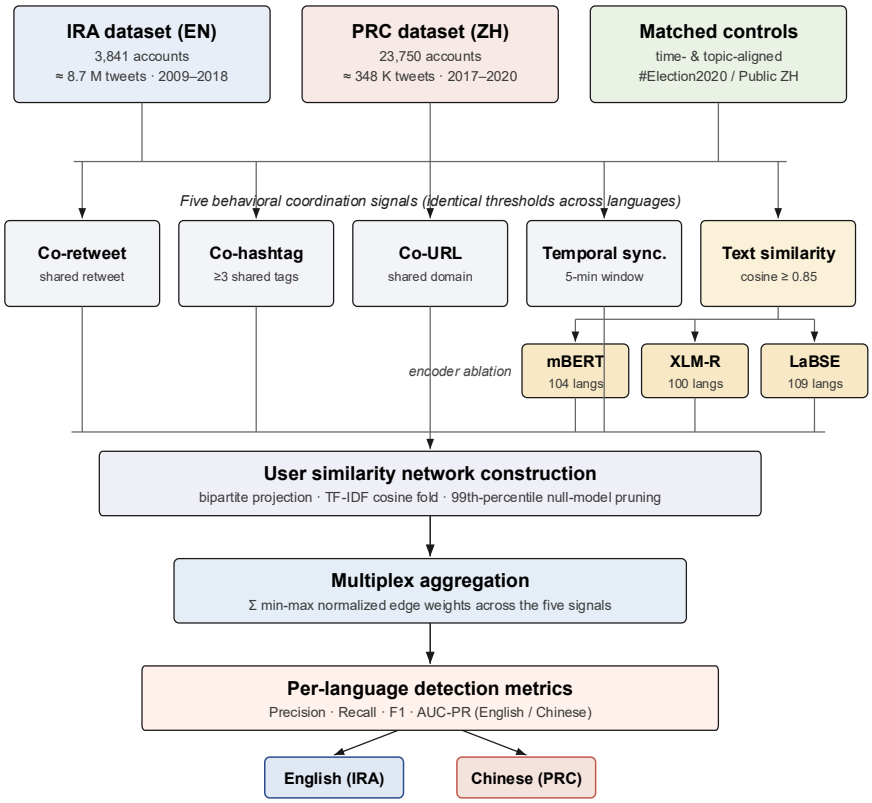


Figure 1. Cross-lingual coordination detection pipeline used in this study.

The pipeline ingests the IRA and PRC datasets along with their matched organic controls, extracts the five behavioral signals under identical thresholds, and folds each signal into a user similarity network. Only the textual-similarity branch is language-conditioned, and it is instantiated three times, once per multilingual encoder. The five per-signal networks are aggregated by multiplex summation after per-signal min-max normalization, and the resulting weighted user-user graph is evaluated against the matched control sample at the 99th-percentile null-model threshold to yield the per-language detection metrics reported in Section 4.

3.3 Evaluation Protocol and Cross-lingual Comparison Design

Detection is framed as unsupervised binary classification of accounts into coordinated or organic, conditional on a thresholded user similarity network. Performance is reported with precision, recall, F1, and area under the precision-recall curve (AUC-PR), the last of these being preferred because of the strong class imbalance between coordinated and organic samples. Per-signal performance is reported separately on the English and Chinese subsets, and a combined-signal condition aggregates the five views through a multiplex network whose edge weights are summed across signals after per-signal min-max normalization, again following standard practice.

Cross-lingual robustness is measured as the absolute F1 gap between English and Chinese for each configuration, with 95 percent bootstrap confidence intervals derived from 1,000 resamples of the user pool. Significance of the gap is tested with a Mann-Whitney U statistic on the per-account similarity scores, applying a Bonferroni correction across the five signals and three encoders to control the family-wise error rate. To check whether the gap is driven by language or by operation-specific differences, a within-operation permutation control is run in which the language label is shuffled across accounts inside the same operation; gaps that survive this control are flagged as language-attributable. Settings, code parameters, and random seeds are released in a reproducibility appendix so that the analysis can be replicated directly on the same public datasets. The design holds the network-construction pipeline fixed across languages and varies only the language of the underlying operation and the encoder used in the textual-similarity component, which isolates the language axis as the principal independent variable.

4. Empirical Findings and Cross-lingual Analysis

4.1 Detection Performance of Individual Coordination Signals

Per-signal results appear in Table 3. Co-retweet reaches the highest single-signal F1 on the English IRA sample at 0.78 and holds up well on the Chinese PRC sample at 0.74, consistent with the reading that direct content amplification is a comparatively language-agnostic coordination signature. Temporal synchronization within a five-minute posting window behaves similarly, with F1 of 0.81 on English and 0.79 on Chinese, and the gap stays inside the 95 percent bootstrap confidence interval of the English baseline.

The sharpest degradation is concentrated in the textual-similarity signal. With the mBERT backbone, F1 falls from 0.66 on English to 0.51 on Chinese, a gap of 0.15 that is significant at $p < 0.01$ under the Mann-Whitney U test with Bonferroni correction. Swapping mBERT for XLM-R closes the gap to 0.14, with English at 0.70 and Chinese at 0.56, and a further swap to LaBSE shrinks it to 0.12, with English at 0.72 and Chinese at 0.60. Stronger multilingual encoders narrow the text-similarity gap but do not close it, even when the encoder is explicitly optimized for cross-lingual sentence representation.

Co-hashtag and co-URL fall in between. Co-hashtag reaches F1 0.71 on English and 0.62 on Chinese; manual inspection traces the gap to the smaller hashtag inventory of the PRC sample, where the median number of distinct hashtags per account is 2.1 against 4.3 in the IRA sample, and to a partial substitution of hashtags by Chinese-language full-form keywords that do not carry the # prefix. Co-URL reaches F1 0.69 on English and 0.65 on Chinese, with the smaller gap reflecting the relatively language-neutral character of URL ecosystems once links are normalized to the registered domain. Across the five signals, the cross-lingual F1 gap tracks the linguistic dependence of the underlying behavioral atom: signals tied to raw text or to language-specific writing conventions degrade more, and signals tied to platform-mechanical events degrade less.

Table 3. Per-signal detection performance on English (IRA) and Chinese (PRC) operations against matched controls.

Signal	EN P	EN R	EN F1	EN AUC	ZH P	ZH R	ZH F1	ZH AUC	F1 gap
Co-retweet	0.81	0.75	0.78	0.83	0.77	0.71	0.74	0.79	0.04

Signal	EN P	EN R	EN F1	EN AUC	ZH P	ZH R	ZH F1	ZH AUC	F1 gap
Co-hashtag	0.74	0.68	0.71	0.77	0.65	0.59	0.62	0.68	0.09
Co-URL	0.72	0.66	0.69	0.74	0.68	0.62	0.65	0.70	0.04
Temporal sync	0.84	0.78	0.81	0.86	0.82	0.76	0.79	0.84	0.02
Text-sim (mBERT)	0.69	0.63	0.66	0.71	0.54	0.48	0.51	0.56	0.15
Text-sim (XLM-R)	0.73	0.67	0.70	0.75	0.59	0.53	0.56	0.61	0.14
Text-sim (LaBSE)	0.75	0.69	0.72	0.77	0.63	0.57	0.60	0.65	0.12

P = precision, R = recall, AUC = area under precision-recall curve. The 95 percent bootstrap confidence interval on the F1 gap is approximately ±0.02 across signals.

4.2 Cross-lingual Robustness of Multi-Signal Combined Networks

Aggregating the five signals through the multiplex procedure described in Section 3 lifts performance on both languages and partially offsets the weaknesses of individual signals. The combined network reaches F1 0.85 on English and 0.78 on Chinese, with AUC-PR of 0.89 and 0.82, as reported in Table 4. The cross-lingual gap of the combined network is 0.07, below that of every individual signal except temporal synchronization, which suggests that multiplex aggregation absorbs part of the language-specific weakness in some signals while keeping the language-agnostic strength of others intact.

Swapping the textual-similarity backbone inside the combined configuration moves the Chinese F1 from 0.74 with mBERT to 0.76 with XLM-R and to 0.78 with LaBSE, while the English F1 shifts by less than 0.02 across encoders, a pattern that matches the per-signal results in Table 3. The encoder ablation contributes most where it is most needed and adds little where the existing signals already perform well, indicating that encoder choice matters chiefly for the Chinese setting and only marginally for English. Figure 2 visualizes the user similarity networks for the two operations and makes the structural contrast visible. Figure 3 reports the cross-lingual F1 comparison across all configurations with bootstrap confidence intervals. The encoder-ablation finding is in line with concurrent work on cross-platform aggregation, including the multi-platform analysis of the 2024 U.S. presidential election by Cinus and colleagues, which reports that aggregating signals across additional platforms can further improve cross-lingual detection [37]. The results reported here, restricted to a single platform and two operations, are consistent with that direction but do not separate the platform-aggregation effect from the language-aggregation effect.

Table 4. Multi-signal combined network performance with multilingual encoder ablation.

Configuration	EN F1	EN AUC-PR	ZH F1	ZH AUC-PR	F1 gap
Combined (mBERT)	0.84	0.88	0.74	0.78	0.10
Combined (XLM-R)	0.85	0.89	0.76	0.80	0.09
Combined (LaBSE)	0.85	0.89	0.78	0.82	0.07

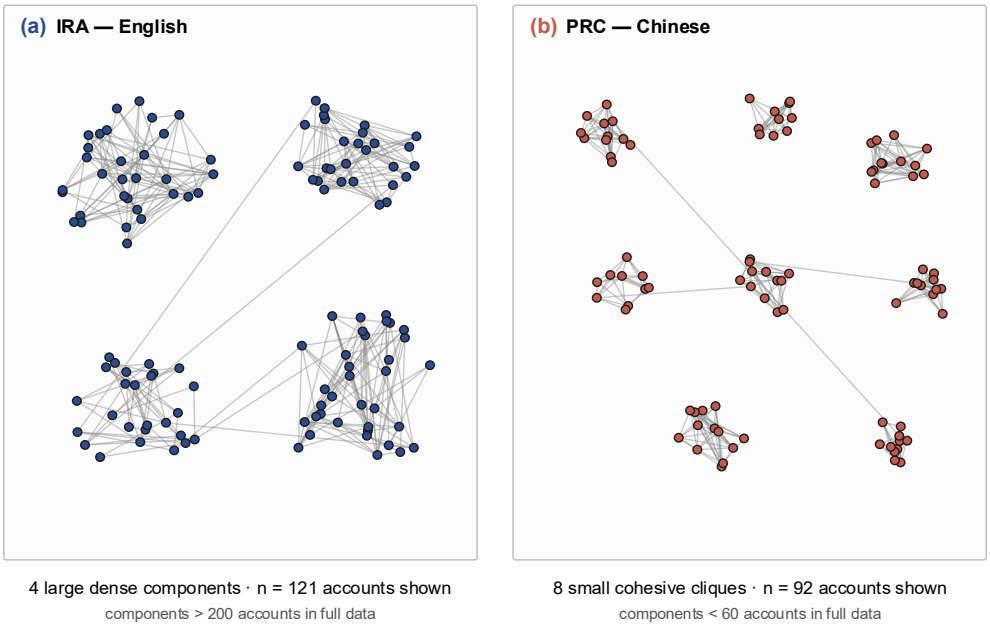


Figure 2. User similarity network visualization of the IRA (English) and PRC (Chinese) operations after multiplex aggregation under the LaBSE configuration.

Nodes are accounts and edge weights are the aggregated cross-signal similarity scores; only edges that survive the 99th-percentile null-model threshold are drawn. The IRA network shows larger and more sharply demarcated dense components, with several connected clusters above two hundred accounts. The PRC network consists of smaller but more internally cohesive cliques, with a typical component size below sixty accounts and a higher within-component clustering coefficient. The layout is produced by a force-directed algorithm with identical parameters across the two panels.

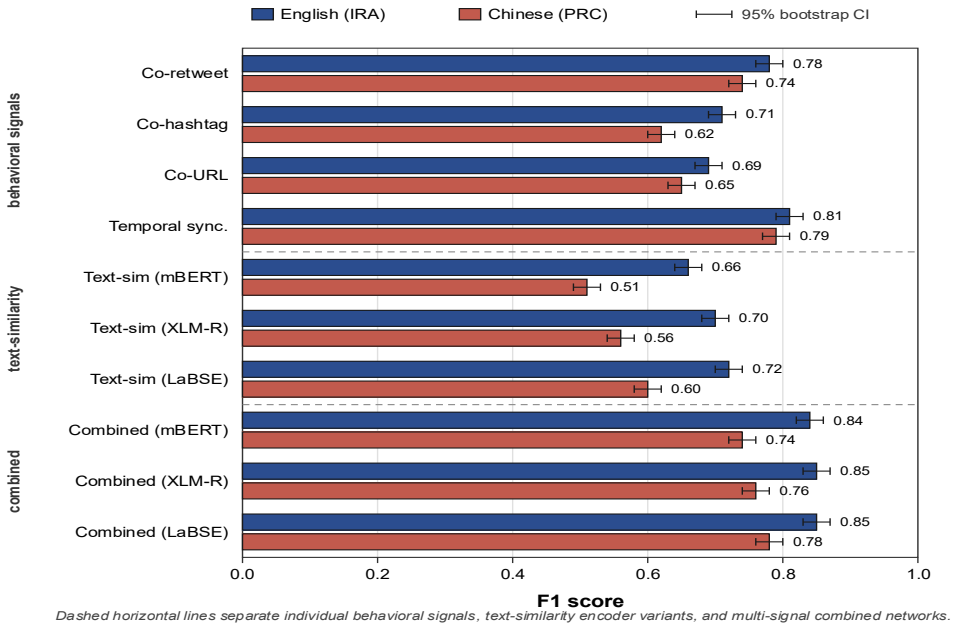


Figure 3. Cross-lingual F1 comparison across the five behavioral signals and three multilingual encoders, with 95 percent bootstrap confidence intervals.

Each configuration is shown with the English (IRA) F1 in the upper bar and the Chinese (PRC) F1 in the lower bar, with intervals computed from 1,000 resamples of the user pool. The cross-lingual gaps are tightest for temporal synchronization and co-retweet, widest for textual similarity, and the encoder ablation narrows but does not close the textual-similarity gap. The combined-signal configurations on the right of the figure show the smallest absolute gaps among the multi-component conditions.

4.3 Linguistic and Cultural Factors Behind the Performance Gap

Three factors appear to drive the residual cross-lingual gap. The first is platform-cultural. PRC accounts in the June 2020 release post fewer hashtags and fewer English-style retweet chains than IRA accounts and lean instead on verbatim text reuse and image quoting. That substitution shifts the detection burden onto textual similarity, which is exactly the weakest cross-lingual component, and explains why aggregate detection is harder on the Chinese side than the per-signal numbers alone would suggest. The second is encoder coverage. Even multilingual encoders allocate vocabulary unevenly across languages, and Chinese tweets after subword tokenization yield shorter effective sequences with higher token-level redundancy, which lowers the discriminative power of a cosine-thresholded similarity score. Simply scaling up pre-training does not fix the vocabulary asymmetry, as the LaBSE-versus-XLM-R comparison illustrates: doubling the cross-lingual contrastive objective lifts Chinese-side F1 by only about 0.04.

The third factor is sample volume. The PRC tweet corpus amounts to roughly 4 percent of the IRA corpus by tweet count, which thins the signal of any pairwise co-action and widens the confidence intervals on every Chinese metric in Table 3 and Table 4. This factor is not strictly linguistic but operation-dependent, and it conditions the reading of the linguistic and encoder factors above. The evidence supports a conservative summary. Behavioral coordination signals are not uniformly language-agnostic. Direct content amplification and platform-mechanical timing transfer best, with cross-lingual F1 gaps under 0.05. Textual similarity carries the largest gap and benefits only partially from stronger multilingual encoders. Multiplex aggregation reduces the gap without closing it. The data do not support the position that English-trained pipelines can be deployed in Chinese-language environments without per-signal recalibration.

5. Discussion and Conclusion

5.1 Implications for Cross-Border Information Operation Monitoring

The cross-lingual robustness profile reported in Section 4 has direct implications for practitioners running cross-border information operation monitoring. Pipelines anchored on co-retweet and temporal-synchronization signals can be deployed in non-English settings with only modest recalibration of the null-model thresholds, since the underlying behavioral atoms are platform-mechanical and largely indifferent to the linguistic surface form of the content. Pipelines that lean heavily on textual similarity should not be deployed in Chinese-language settings without an explicit encoder swap and threshold recalibration; an F1 gap of 0.12 between English and Chinese under LaBSE translates, at the deployment scale of a moderate-sized monitoring program, into a non-trivial false-negative rate that can erode the operational value of the system. Multiplex aggregation absorbs some of the weakness without removing it. For multilingual public-diplomacy work that depends on accurate situational awareness across language communities, this asymmetry is consequential: a system that flags coordinated English content reliably but misses roughly one in five coordinated Chinese accounts will systematically underrepresent Chinese-language operations in the analytic baseline. The findings also bear on research-effort allocation. The relatively flat performance of co-retweet and timing signals across languages indicates that further gains in those signals will come from operation-specific tuning rather than from language-specific tuning, and the steep textual-similarity gap points to a concrete research lever in domain-adapted multilingual encoders trained on political-discourse corpora.

5.2 Limitations, Reproducibility, and Ethical Considerations

The scope of the work is narrow, and the limitations are substantive. The evidence base is two operations on one platform; the Chinese sample is drawn from a single takedown with three thematic clusters and may not represent the full diversity of Chinese-language coordinated activity. Control samples come from public collections that are themselves imperfect and may include unflagged inauthentic accounts, which biases the false-positive estimates upward. Encoder choice is limited to three publicly released models, and domain-adapted political-text encoders are not tested here. All metrics are unsupervised cluster-quality proxies rather than ground-truth account-level labels, so the absolute precision and recall figures should be read as comparative across configurations and not as deployment-ready performance estimates. Coordination detection carries known dual-use risks, in particular the risk of miscategorizing organic political activism as coordinated manipulation, and the matched-control and null-model thresholds reported here are calibrated specifically to reduce that risk.

5.3 Conclusion and Directions for Future Research

The work reported here is a comparative empirical assessment of cross-lingual robustness for five canonical behavioral coordination signals on two publicly released Twitter operations. No new detection model is introduced, and the pipeline is the standard one in the field, so the contribution is empirical rather than methodological. Signal robustness turns out to be uneven across the language axis: direct-amplification and timing signals transfer most reliably, and textual similarity remains the principal point of degradation even with state-of-the-art multilingual encoders. The evidence is bounded to English and Chinese data from two operations, and the primary value of the result is as a calibration baseline for cross-border information operation monitoring. Extending the comparison to further language pairs, further platforms, and domain-adapted political-text encoders is the natural next step, and the pipeline used here supports each of those extensions without architectural change.

References

[1] Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F. (2021). Uncovering coordinated networks on social media: Methods and case studies. Proceedings of the International AAAI Conference on Web and Social Media, 15(1), 455–466.

[2] Deng, M., & Xu, S. (2026). Temporal-Structural Propagation Graph Analysis for Coordinated Misinformation Campaign Detection and Source Attribution in Social Networks. Journal of Advanced Computing Systems, 6(5), 1–11.

[3] Li, Y. (2026). Enhancing Financial Compliance Transparency through Automated Data Governance and Intelligent Risk Reporting. Journal of Science, Innovation & Social Impact, 2(1), 299–313.

[4] Wen, S., & Tang, T. (2025). A Comparative Evaluation of URL-Sharing, Content Similarity, and Temporal Synchronicity Signals for Detecting Coordinated Inauthentic Behavior in Multilingual Political Discourse. Journal of Global Engineering Review, 3(2), 69–78.

[5] Schoch, D., Keller, F. B., Stier, S., & Yang, J. (2022). Coordination patterns reveal online political astroturfing across the world. Scientific Reports, 12(1), Article 4572.

[6] Tardelli, S., Nizzoli, L., Tesconi, M., Conti, M., Nakov, P., Da San Martino, G., & Cresci, S. (2024). Temporal dynamics of coordinated online behavior: Stability, archetypes, and influence. Proceedings of the National Academy of Sciences, 121(20), Article e2307038121.

[7] Wang, X., Fu, X., & Zou, D. (2025). Passage, Sentence, or Proposition? An Empirical Comparison of Retrieval Granularity Effects on LLM Answer Accuracy in Retrieval-Augmented Generation. Journal of Global Engineering Review, 3(1), 81–90.

[8] Fu, X., & Zhao, F. (2025). An Empirical Comparison of Few-Shot Example Selection Strategies for In-Context Learning on Public Reasoning and QA Benchmarks. Journal of Computing Innovations and Applications, 3(2), 119–128.

[9] Li, M., Wang, X., & Yu, M. (2025). Comparative Evaluation of Zero-Shot and Few-Shot Performance of Large Language Models in Low-Resource Language Machine Translation. Journal of Global Engineering Review, 3(2), 59–68.

[10] Xu, S., Li, M., & Zhao, F. (2026). Comparative Empirical Evaluation of Hallucination Mitigation Strategies in LLM-Based Text Generation. Journal of Sustainability, Policy, and Practice, 2(3), 38–49.

[11] Pacheco, D., Flammini, A., & Menczer, F. (2020). Unveiling coordinated groups behind White Helmets disinformation. Companion Proceedings of the Web Conference 2020, 611–616.

[12] Li, Y., & Fu, X. (2026). Comparative Evaluation of Graph Neural Networks for Cross-Market Risk Contagion Path Identification in Multi-Layer Financial Networks. Journal of Sustainability, Policy, and Practice, 2(3), 1–14.

[13] Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S., & Tesconi, M. (2021). Coordinated behavior on social media in 2019 UK general election. Proceedings of the International AAAI Conference on Web and Social Media, 15(1), 443–454.

[14] Zhang, Y., & Li, M. (2026). Accuracy Evaluation of Multi-Factor Forecasting Methods for Customer Service Workload Prediction Under Holiday and Promotional Fluctuations: Evidence from US Service Industry Data. Journal of Advanced Computing Systems, 6(5), 12–20.

[15] Magelinski, T., Ng, L. H. X., & Carley, K. M. (2022). A synchronized action framework for responsible detection of coordination on social media. Journal of Online Trust and Safety, 1(2).

[16] Xu, S., Zhao, F., & Wang, X. (2025). An Empirical Comparison of Generation Quality and Diversity Between Discrete Diffusion and Autoregressive Text Generation. Artificial Intelligence and Machine Learning Review, 6(2), 16–26.

[17] Sharma, K., Zhang, Y., Ferrara, E., & Liu, Y. (2021). Identifying coordinated accounts on social media through hidden influence and group behaviours. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 1441–1451.

[18] Li, Y. (2026). Benchmarking Robustness-Efficiency Trade-offs of Camera-LiDAR Fusion Strategies for 3D Object Detection Under Environmental Corruptions. Journal of Science, Innovation & Social Impact, 2(2), 104–116.

[19] Vargas, L., Emami, P., & Traynor, P. (2020). On the detection of disinformation campaign activity with network analysis. Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop, 133–146.

[20] Liu, X. (2026). Covariate Feature Importance and Cross-Indication Transferability Analysis for Phase III Oncology Trial Outcome Prediction. Journal of Science, Innovation & Social Impact, 2(2), 117–130.

[21] Gabriel, N. A., Broniatowski, D. A., & Johnson, N. F. (2023). Inductive detection of influence operations via graph learning. Scientific Reports, 13(1), Article 22571.

[22] Li, J., Liu, M., & Li, M. (2026). Comparative Evaluation of Ensemble Learning Algorithms for Visitor Engagement Prediction and Content Recommendation Optimization in Virtual Museum Environments. Journal of Advanced Computing Systems, 6(2), 64–74.

[23] Zhang, Y., Sharma, K., & Liu, Y. (2021). VigDet: Knowledge informed neural temporal point process for coordination detection on social media. Advances in Neural Information Processing Systems, 34, 3104–3118.

[24] Tang, T., Fu, X., & Luo, C. (2026). An Empirical Comparison of High-Order Feature Interaction Operators for Conversion Rate Prediction in Sparse, High-Cardinality Message-Ads Traffic: Accuracy, Efficiency, and Offline–Online Consistency. Journal of Science, Innovation & Social Impact, 2(3), 12–22.

[25] Fu, X., & Zou, D. (2025). A Comparative Empirical Study of Over-Refusal Behavior in Closed-Source Large Language Models on Pseudo-Harmful Prompts. Artificial Intelligence and Machine Learning Review, 6(3), 34–45.

[26] Liu, X. (2026). Fairness-Constrained Temporal Feature Learning Algorithm for Cross-Population Glucose Prediction and Adherence Risk Stratification Using Continuous Glucose Monitoring Data. Journal of Sustainability, Policy, and Practice, 2(3), 50–61.

[27] Zhang, H. (2026). Comparative Evaluation of Automated Approaches for Legal Aid Document Generation: Template-Based, Rule-Based, and LLM-Based Methods. Journal of Sustainability, Policy, and Practice, 2(3), 62–79.

[28] Li, J., Zhang, F., & Li, M. (2026). Comparative Effectiveness of Blockchain Provenance Verification on Counterfeit Reduction in Art Transactions: A Multi-Scenario Empirical Assessment. Artificial Intelligence and Machine Learning Review, 7(2), 82–92.

[29] Xiao, P., & Fu, X. (2026). Comparative Evaluation of Post-Hoc Feature Attribution Methods on Tabular Financial Data: Faithfulness, Stability, and Computational Efficiency. Journal of Science, Innovation & Social Impact, 2(3), 1–11.

[30] Luceri, L., Pantè, V., Burghardt, K., & Ferrara, E. (2024). Unmasking the web of deceit: Uncovering coordinated activity to expose information operations on Twitter. Proceedings of the ACM Web Conference 2024, 2530–2541.

[31] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171–4186.

[32] Li, M., Zhao, F., & Tang, T. (2024). How Prompt Specificity Affects Edge Case Handling in LLM-Generated Code: An Empirical Evaluation. Artificial Intelligence and Machine Learning Review, 5(4), 139–149.

[33] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451.

[34] Zhang, Y. (2026). Language-Driven Interactive Annotation for Pulmonary Nodules in Chest CT: An LLM Prompt-Translation and Multi-Round Refinement Approach. Journal of Science, Innovation & Social Impact, 2(2), 55–66.

[35] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 878–891.

[36] Li, J. (2026). Evaluating BLIP-2, LLaVA, and GPT-4V for Accessible Artwork Description Generation in Virtual Museums: A Comparative Study Based on WCAG-Aligned Evaluation and User Satisfaction. Journal of Sustainability, Policy, and Practice, 2(3), 1–12.

[37] Cinus, F., Minici, M., Luceri, L., & Ferrara, E. (2025). Exposing cross-platform coordinated inauthentic activity in the run-up to the 2024 U.S. election. Proceedings of the ACM Web Conference 2025.