International Journal of Computational Research and Development (IJCRD)
International Peer Reviewed - Refereed Research Journal, Website: www.dvpublication.com
Impact Factor: 5.015, ISSN (Online): 2456 - 3137, Volume 11, Issue 2, July - December, 2026

DETECTION METHOD LANGUAGE AND PLATFORM CHARACTERISTICS PREDICT MACHINE GENERATED TEXT IDENTIFICATION ACROSS MULTILINGUAL SOCIAL PLATFORMS

Divya Vishwanath* & Paul Johnson Asamoah**
* Janardhan Rai Nagar Rajasthan Vidyapeeth, Udaipur, Rajasthan, India
** Power Life University College, Accra, Ghana, West Africa

Cite This Article: Divya Vishwanath & Paul Johnson Asamoah, "Detection Method Language and Platform Characteristics Predict Machine Generated Text Identification across Multilingual Social Platforms", International Journal of Computational Research and Development, Volume 11, Issue 2, July - December, Page Number 12-20, 2026.
Copy Right: © DV Publication, 2026 (All Rights Reserved). This is an Open Access Article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.
Type of Review: Peer Reviewed as per |C|O|P|E| Guidance.
Disclaimer: The scholarly papers reviewed and published by DV Publication, India, reflect the views and opinions of their respective authors and do not necessarily represent the views or opinions of DV Publication. The publisher disclaims any responsibility for any harm, loss, or damage resulting from the use of the published content by any party.
DOI: https://doi.org/10.5281/zenodo.24041537

Abstract:

This analysis examines whether detector rankings remain consistent across multilingual social media environments when evaluation coverage is standardized. A descriptive secondary analysis uses 1,032 published Multi Social evaluations from 2025, covering twelve detectors, twenty-two languages and five platforms. Comparisons summarize AUROC distributions and match detector, language and platform groups. Configurations adapted to Multi Social achieve mean AUROC of 0.9622, compared with 0.5470 for previously trained alternatives, with higher family averages in all 86 common evaluation combinations. Across nine languages with complete platform coverage, the spread in language means narrows from 0.2116 among previously trained configurations to 0.0171 among adapted configurations. Platform ordering also changes: Telegram leads unrestricted comparisons, whereas Discord leads after matching language coverage. This reversal demonstrates how evaluation composition shapes aggregate rankings without changing underlying scores. The analysis refines the descriptive application of domain adaptation reasoning by distinguishing consistent configuration advantages from composition sensitive comparisons. Detector procurement and governance should require matched evaluations and documented linguistic coverage before translating pooled benchmark rankings into operational decisions.

Key Words: Machine Generated Text Detection; Multilingual Evaluation; Social Media; Domain Adaptation; Matched Descriptive Analysis.

1. Introduction:

Selecting a detector for machine generated text requires more than identifying the highest benchmark score. A score summarizes performance within a particular evaluation sample, while institutions must decide whether that performance is relevant to the languages and communication environments they encounter. This distinction matters for multilingual social platforms, where comparisons can combine different linguistic populations and sources of text. When coverage differs, an apparent performance advantage may reflect both detector capability and evaluation composition. The substantive question is therefore how detector comparisons change when languages and platforms are made comparable. Existing research establishes that detection performance depends on the material being evaluated. Liang et al., 2023, document misclassification of English writing by nonnative speakers, showing that linguistic differences can affect detector outputs within a single language. Weber Wulff et al., 2023, identify limitations in detection tools and examine their sensitivity to machine translation and content obfuscation. These findings caution against interpreting performance as a stable property of a detector across all textual environments. They also distinguish the ability to discriminate authorship under specified conditions from the reliability required for decisions affecting individual writers.

Domain adaptation theory provides a basis for examining this dependence. Ben David et al., 2010, relate classification performance in a target domain to performance in the source domain, divergence between the domains and the existence of a classifier that performs adequately in both. Applied to multilingual social media, this reasoning suggests that learned distinctions between human and generated writing may remain informative in some evaluation environments while transferring less effectively to others. Detector configuration, evaluation language and platform consequently describe related dimensions of the conditions under which discrimination is observed. Their associations with performance do not, by themselves, identify the separate causal contributions of architecture, training data or adaptation procedures. The closest empirical foundation is Multi Social. Macko et al., 2025, introduce a benchmark containing 472,097 human and machine generated texts across twenty-two languages and five social platforms. Their evaluation compares existing detection methods and demonstrates the relevance of adaptation to social media texts and platform selection during training. This evidence already establishes multilingual and platform variation as substantive evaluation concerns. The present analysis therefore does not introduce a new benchmark or claim the discovery of platform dependence. It undertakes a secondary analysis of published results to examine the stability of comparative conclusions under explicitly matched evaluation coverage.

The analytical question concerns the relationship between performance rankings and the composition of the observations used to construct them. An unrestricted platform average combines available language evaluations, whereas a matched average compares platforms using the same languages and detectors. These summaries answer different questions. Similarly, a language average can obscure whether variation is concentrated among particular detector configurations. Making these distinctions explicit allows the analysis to separate differences across evaluated systems from differences arising when the composition of the comparison changes. The outcome is the published area under the receiver operating characteristic curve, abbreviated AUROC.

This measure describes discrimination across classification thresholds, allowing the analysis to compare reported performance on a common scale. It does not measure classification accuracy, deployment specific error costs or the proportion of correctly identified messages. Saito and Rehmsmeier, 2015, demonstrate that receiver operating characteristic and precision recall assessments convey different information under class imbalance. Accordingly, the present analysis interprets AUROC as benchmark discrimination and confines its conclusions to that measured outcome.

This analysis aims to determine how published AUROC varies across detector configurations, evaluation languages and platforms when comparisons use common observed coverage. It examines 1,032 complete evaluation records extracted from the Multi Social publication issued in 2025. Each record represents one detector evaluated within one language and platform combination. Twelve detectors share 86 retained combinations, enabling direct descriptive comparisons across configurations. Additional comparisons restrict language and platform coverage to common groups. The design supports transparent assessment of performance distributions and ranking sensitivity without treating shared benchmark evaluations as independent experiments. The completed comparisons identify two distinct patterns. Configurations adapted to Multi Social outperform previously trained alternatives across every retained language and platform combination when performance is averaged within each detector family. At the same time, the platform with the highest unrestricted mean differs from the platform with the highest mean under matched language coverage. Language variation is also substantially narrower among adapted configurations. Together, these results show that a strong configuration contrast can coexist with platform rankings that depend on evaluation composition.

The contribution is a bounded descriptive reassessment of published benchmark evidence. It clarifies which comparisons remain consistent across observed settings and which change when coverage is standardized. Its international relevance follows from the multilingual evaluation structure rather than geographical representativeness. For detector selection and reporting, the findings support presenting common coverage comparisons alongside pooled scores. This makes the population of evaluations underlying a performance claim visible and helps readers distinguish broad benchmark averages from evidence applicable to particular language and platform settings.

2. Theory, Conceptual Framework and Research Questions Development:

This analysis draws on domain adaptation theory to explain variation in machine generated text detection across methods, languages and social platforms. Detection depends on whether the distinctions learned from training data remain informative in the evaluation environment. Changes in language or platform can alter the relationship between textual features and authorship labels, limiting the transferability of a detector. Ben David et al., 2010, establish that performance in a target domain depends partly on divergence from the source domain and the availability of a classifier that performs adequately in both. This provides a theoretical basis for examining detection performance as conditional on the evaluation setting rather than treating it as a fixed property of a model. The conceptual framework connects three observed dimensions of detection evaluation configuration with published detection performance. Detection method identifies the evaluated detector configuration, evaluation language identifies the language of the benchmark texts, and evaluation platform identifies their social media source. The outcome is the area under the receiver operating characteristic curve, abbreviated AUROC. Higher AUROC indicates stronger discrimination between human and machine generated texts across classification thresholds. These relationships represent descriptive associations among observed categories and performance scores. The framework does not treat configuration as a latent construct or combine its dimensions into a numerical index.

Detection method captures differences in how evaluated systems distinguish human writing from generated text. Gehrmann et al., 2019, demonstrate that statistical regularities in generated language can support detection, providing evidence that generation processes leave identifiable textual patterns. The theoretical implication is that methods differ according to the signals they capture and the conditions under which those signals remain informative. However, sensitivity to one generation pattern does not establish superiority across languages and platforms. This analysis therefore compares the available detector configurations within common language and platform combinations rather than imposing a universal ranking. Published benchmark evidence also supports treating detector performance as dependent on its training context. Macko et al., 2025, report that detectors adapted to social media texts perform strongly on the Multi Social benchmark and that platform selection during training matters. These findings motivate comparisons between the evaluated methods, but the available aggregate records do not isolate architecture from training data or adaptation procedures. Accordingly, differences between configurations describe their combined performance rather than the causal contribution of any single design choice. Research question 1 asks: How does published AUROC vary across detector configurations within the same evaluation language and platform?

Evaluation language provides a second source of variation because language technologies operate within an uneven distribution of linguistic resources. Joshi et al., 2020, document substantial disparities in language representation and question assumptions that existing systems operate independently of language. This supports examining whether detector performance remains consistent across linguistic settings. Nevertheless, resource inequality does not establish the direction of every language comparison. Multilingual learning may support transferable distinctions, while language specific variation may weaken their consistency. This analysis therefore uses language as an observed categorical dimension without assigning unmeasured resource scores or assuming that English must outperform every alternative. Research question 2 asks: How does published AUROC vary across evaluation languages within the same detector and platform?

Evaluation platform captures the source environment represented by the benchmark texts. Bender and Friedman, 2018, explain why the populations and contexts represented in language datasets constrain the generalization of technological claims. Applied here, platform comparisons assess whether a detector maintains similar discrimination across different sources of social communication. Comparable results would indicate descriptive consistency across the observed environments, whereas divergent results would identify limits to that consistency. Platform labels do not independently measure communication practices, user characteristics or institutional rules. Research question 3 therefore asks: How does published AUROC vary across evaluation platforms within the same detector and language?

The available dataset supports these questions through 1,032 complete detector, language and platform evaluation records. Descriptive comparisons can report performance distributions, matched differences and changes in detector rankings

across observed settings. Comparisons use available common combinations because some source cells lack reported AUROC. Published scores are rounded to two decimal places, and multiple detectors evaluate shared benchmark material. Consequently, the framework supports comparisons of reported performance without treating every record as an independent experiment or interpreting small numerical differences as established statistical differences.

3. Data:

This section establishes the provenance of the published evaluation records, documents sample construction and defines the observed variables. It also reports data checks and descriptive statistics supporting comparisons across detectors, languages and platforms.

3.1 Data Sources, Population, and Sample Construction:

This analysis uses published aggregate evaluation results from Multi Social, provided by Macko and colleagues through the Association for Computational Linguistics. Tables 24 and 25, on printed pages 749 and 750, report AUROC by detector, evaluation language and platform. Table 24 covers five previously trained detectors, while Table 25 covers seven detectors adapted to Multi Social. The analytical workbook consolidates these published results rather than individual social media messages. Macko et al., 2025, describe the underlying benchmark and its multilingual evaluation design. The associated Multi Social repository identifies version v1 under DOI 10.5281/zenodo.13846152. Its message files require approved research access and prohibit redistribution beyond authorized recipients. This analysis uses publicly reported aggregate scores without accessing or redistributing those restricted files. The repository record establishes benchmark provenance but does not constitute a second analytical dataset. Macko et al., 2024, document the access conditions.

The unit of analysis is one detector evaluated for one language and one platform. The empirical setting is automated text detection across Discord, Gab, Telegram, Twitter and WhatsApp. Coverage is multilingual and platform based rather than geographically representative. Language identifiers do not establish the countries or locations of message authors. The data form a cross-sectional benchmark published in 2025. They contain neither repeated annual observations nor message timestamps. Accordingly, the stated 2010-2025 scope identifies a publication eligibility window rather than a longitudinal observation period. Individual message dates cannot be established from the analytical records. The extraction date is documented separately in the accompanying workbook and does not define the period represented by the benchmark. The target population comprises detector performance evaluations across multilingual social platforms. The accessible sampling frame is narrower: the detector, language and platform cells reported in Tables 24 and 25. Selection includes every language specific numerical AUROC in those tables. Pooled results across all languages are excluded because they overlap the disaggregated observations. Cells reported as unavailable are retained in the source worksheet but excluded from the analytical sample.

Table 1: Sample construction

Stage	Detector Units	Evaluation Observations	Implemented Rule
Published cells in the two selected tables	12	1,380	Five platforms and 23 language columns, including the pooled column
Exclusion of pooled language results	12 retained	60 excluded	Remove the overlapping all languages column
Language specific sampling frame	12	1,320	Twelve detectors across 22 languages and five platforms
Exclusion of unavailable AUROC cells	12 retained	288 excluded	Exclude source cells marked N/A
Final descriptive sample	12	1,032	Retain complete numerical evaluation records

The retained sample represents 78.18 percent of the language specific frame. Each detector contributes 86 observations covering the same available language and platform combinations. Coverage is therefore balanced across detectors within the retained combinations, but incomplete across the full language and platform grid. The construction rules are recorded in the workbook; the supplied materials contain no separate preregistered selection protocol.

3.2 Data Integration, Cleaning, and Quality Assurance:

Integration consisted of appending the records from the two source tables into a common structure. Detector identity, language code and platform jointly define the analytical key. Detector names preserve the distinction between previously trained configurations and configurations adapted to Multi Social. Source table number, printed page and publication DOI provide record level provenance. Verification identified 1,032 unique analytical keys and no duplicate keys. Every retained AUROC matched its corresponding value in the source worksheet. All analytical records contain complete detector, language, platform and outcome fields. The integration introduced no additional observations beyond the eligible source cells.

The 288 unavailable source cells arise from combinations for which the published tables do not report a meaningful AUROC because of insufficient observations in at least one authorship class. These values were excluded without interpolation, imputation or replacement by zero. Missing cells number 84 for Discord, 12 for Gab, 12 for Telegram, 48 for Twitter and 132 for WhatsApp. Retained observations consequently number 180, 252, 252, 216 and 132, respectively. Availability also differs across languages. Nine languages have complete coverage across all five platforms: Catalan, Czech, German, English, Spanish, Estonian, Dutch, Portuguese and Romanian. Scottish Gaelic appears for only one platform, while Slovenian and Ukrainian appear for two. Language comparisons therefore require attention to the platform combinations actually represented. AUROC values were checked against the admissible interval from zero to one. All retained scores satisfy this condition. Scores below 0.50 were preserved rather than reversed, and scores reported as 1.00 were retained. No trimming, winsorization or rescaling was applied. These checks distinguish valid extreme scores from extraction errors without changing the source evidence. The resulting sample describes available benchmark evaluations rather than representative platform populations. Unequal language and platform

coverage limits unrestricted comparisons of pooled averages. This interpretation respects the distinction between dataset coverage and generalization emphasized by Bender and Friedman, 2018.

3.3 Variable Measurement, Construction, and Validation:

The framework uses three observed categorical dimensions of evaluation configuration and one numerical outcome. Detector method identifies the evaluated system, language identifies the linguistic evaluation category, and platform identifies the source environment. These dimensions remain separate categories rather than components of a weighted index. Detection performance is measured by published AUROC. The scores were extracted at their reported precision of two decimal places and were not recalculated from individual predictions. Higher values indicate stronger discrimination between human and machine generated texts across classification thresholds. AUROC is neither classification accuracy nor a percentage of correctly classified messages. Its interpretation also differs from precision and recall under class imbalance, as Saito and Rehmsmeier, 2015, demonstrate.

Table 2: Variable dictionary

Variable or Indicator	Analytical Role	Definition and Construction	Source and Identifier	Unit and Frequency	Transformation and Direction	Missing Value Rule	Coverage and Limitation
Detector method	Evaluation dimension	Published detector configuration	Multi Social, Tables 24 and 25	Nominal category; one benchmark	No numerical ordering	Retain complete identifiers	Twelve configurations; architecture and training differences are bundled
Evaluation language	Evaluation dimension	Published ISO language code	Multi Social, language columns	Nominal category; one benchmark	No numerical ordering	Retain reported categories	Twenty two languages; platform coverage varies
Evaluation platform	Evaluation dimension	Source platform of evaluation texts	Multi Social, platform rows	Nominal category; one benchmark	No numerical ordering	Retain reported categories	Five platforms; does not identify training platform
AUROC	Outcome	Published area under the receiver operating characteristic curve	Multi Social, numerical cells	Ratio scale from zero to one; one benchmark	No transformation; higher indicates stronger discrimination	Exclude N/A cells	Two decimal precision; individual prediction uncertainty unavailable
Method family	Descriptive classification	Previously trained or adapted to Multi Social, determined by source table	Tables 24 and 25	Nominal category; one benchmark	No numerical ordering	Complete for retained records	Nested within detector identity
Publication year	Provenance	Year of the benchmark publication	Source publication	Calendar year	No transformation	Complete	Not a message observation year
Source table, page and DOI	Traceability	Location of each extracted result	Publication metadata	Record identifiers	No transformation	Complete	Establish source location rather than independent measurement

Validation concerned observed values: key uniqueness, completeness, admissible score ranges and agreement between analytical and source worksheets. These checks establish consistency of extraction without claiming independent validation of the underlying authorship labels.

Table 3: Descriptive statistics for published AUROC

Measure	Verified Value
Valid observations	1,032
Mean	0.7892
Sample standard deviation	0.2486
Median	0.9400
Minimum	0.0500
Maximum	1.0000

The previously trained detector category contributes 430 observations, while detectors adapted to Multi Social contribute 602. Each detector contributes equally to the overall summary, but platforms and languages contribute according to their available evaluation cells. The pooled mean therefore describes the retained benchmark records rather than performance weighted by platform usage or message prevalence.

3.4 Analytical Framework, Identification, and Reproducibility:

The implemented analysis is descriptive. Its quantities of interest are the distribution of published AUROC and its variation across observed detector, language and platform categories. The completed procedures comprise sample accounting, measurement verification and calculation of valid counts, mean, sample standard deviation, median and range. No regression, causal estimator, forecasting model or significance test was fitted. The structure supports matched descriptive comparisons. Detector comparisons can use identical language and platform combinations because all twelve detectors share the same retained coverage. Language and platform comparisons require common observed combinations to avoid conflating differences in performance with differences in sample composition. The aggregate scores do not identify the separate causal contributions of architecture, training resources or adaptation procedures.

Evaluations share benchmark material, so the 1,032 records cannot be treated as independent experiments. Published rounding further limits interpretation of small differences. The data provide no message level predictions or uncertainty estimates from which to reconstruct confidence intervals for individual AUROC scores. Conclusions therefore concern observed benchmark performance rather than statistical significance or deployment effectiveness. Reproducibility rests on separating published source values from the retained analytical records and documenting each record through its detector, language, platform and source location. Peng, 2011, identifies access to analytical data and computational procedures as central to reproducible research. The

workbook supplies the source cells, analytical dataset, variable dictionary and selection documentation needed to inspect the reported descriptive calculations. The preserved source cells also make exclusions inspectable rather than concealing unavailable combinations. Sandve et al., 2013, emphasize retaining the information needed to trace computational results. Here, consistent identifiers, unchanged AUROC scales and explicit missing value rules align the analytical records with the published evidence and bound the conclusions to the evaluations actually observed.

4. Methods:

4.1 Research Design and Comparative Framework:

This analysis uses a descriptive secondary analysis of published benchmark evaluations to examine differences in detection performance across configurations, languages and social platforms. The unit of analysis is one detector evaluated within one language and platform combination. The quantities of interest are group means, dispersion, matched performance differences and changes in comparative rankings after standardizing evaluation coverage. The design distinguishes comparisons across detector configurations from comparisons affected by unequal language or platform coverage. Detector comparisons use the same 86 observed language and platform combinations for all twelve configurations. Language comparisons restrict the sample to languages represented across all five platforms. Platform comparisons use identical languages and detectors. These restrictions make the composition of each comparison explicit.

The sample construction rules and measurement definitions appear in Table 1 and Table 2, respectively. The analysis uses the final retained sample without additional exclusions based on performance. Its comparative variation comes from published AUROC differences across observed evaluation settings. The benchmark is cross sectional; publication year does not represent repeated annual measurement. The design identifies descriptive differences rather than causal effects. Detector configurations combine architectural and training characteristics that are not independently assigned. Matching evaluation categories removes differences in their observed composition but does not isolate the contribution of individual training procedures. The framework therefore supports conclusions about evaluated configurations within specified benchmark settings.

4.2 Descriptive Measures and Matched Comparisons:

Published AUROC is retained on its original scale from zero to one. Higher values indicate stronger discrimination between human and machine generated texts across classification thresholds. No score is converted into classification accuracy or interpreted as the proportion of correctly classified messages. This distinction matters because discrimination and precision under class imbalance describe different properties of a classifier, as Saito and Rehmsmeier, 2015, demonstrate. For each analytical group, the arithmetic mean is calculated as:

$$\bar{A}_g = \frac{1}{n_g} \sum_{i \in g} A_i$$

Here, A_i is the published AUROC for evaluation record i, g identifies the selected group of records, n_g is its observation count, and $\bar{A}_g$ is its mean AUROC. Groups comprise individual detectors, detector families, languages or platforms. Each retained evaluation receives equal weight within its group. Standard deviations use the sample convention with a denominator equal to the observation count less one. Medians, minima and maxima supplement means to describe dispersion and asymmetry. Detector family comparisons first average scores separately within each common language and platform combination. The matched difference is then calculated as:

$$D_c = \bar{A}_{F,c} - \bar{A}_{P,c}, \qquad \bar{D} = \frac{1}{C} \sum_{c=1}^{C} D_c$$

Here, cidentifies a common language and platform combination, Fdenotes the seven configurations adapted to Multi Social, P denotes the five previously trained configurations, D_c is their difference in mean AUROC, C is the number of common combinations, and $\bar{D}$ is the average matched difference. Equation 2 gives each combination equal weight and prevents the unequal number of detectors in the two families from determining the family contrast. Platform comparisons also use paired differences:

$$D_{d,l}^{p,q} = A_{d,l,p} - A_{d,l,q}$$

Here, d identifies a detector, l identifies an evaluation language, pand q identify two platforms, and $D_{d,l}^{p,q}$is their AUROC difference for the same detector and language. Positive differences favour platform \(p\), negative differences favour platform q, and zero differences indicate equality at the published precision. The analysis summarizes their mean, range and counts by direction.

4.3 Analytical Sequence and Coverage Sensitivity:

Processing begins with the cleaned analytical worksheet and its corresponding source cells. Records are indexed by detector identifier, language code and platform. The first stage calculates overall descriptive statistics and detector specific summaries using all 1,032 retained evaluations. The second stage groups detectors by their published training category and applies Equation 2 across the 86 common evaluation combinations. The third stage identifies languages represented across all five platforms for every detector. Nine languages meet this condition: Catalan, Czech, German, English, Spanish, Estonian, Dutch, Portuguese and Romanian. Each language contributes sixty evaluations, comprising twelve detectors across five platforms. Language means are calculated for the full detector set and separately for each detector family.

The fourth stage uses these same nine languages to construct a common platform sample. Each platform contributes 108 evaluations, comprising twelve detectors across nine languages. Unrestricted platform means are compared with means from this matched sample to assess sensitivity to language coverage. Changes between these summaries reflect a deliberate change in sample composition while preserving the outcome definition and calculation method. Selected paired platform comparisons use Equation 3 across the 108 common detector and language pairs. The analysis additionally compares the lowest adapted detector score with the highest previously trained detector score within each of the 86 available combinations. This evaluates whether differences between family averages coexist with overlap among individual configurations. These matched comparisons are

descriptive extensions of the published evaluations. They were not established through a preregistered protocol and are treated as exploratory. Calculations use Python standard library procedures for grouping, arithmetic summaries, medians and sample standard deviations. No random sampling, stochastic optimization or simulation enters the analysis.

4.4 Verification, Reproducibility and Interpretive Limits:

Verification assesses whether the analytical records reproduce the source information consistently. The implemented checks cover uniqueness of the detector, language and platform key, completeness of retained fields, agreement between analytical AUROC values and source worksheet values, and compliance with the admissible interval from zero to one. Observation counts are reconciled across the sample construction stages and analytical groups. Unavailable source cells remain excluded without imputation. Valid extreme scores, including values below 0.50 and values reported as 1.00, are preserved. No inversion, trimming, winsorization or normalization is applied. All calculations use the published two decimal scores without introducing artificial precision into the underlying observations. The analysis reports descriptive dispersion rather than standard errors, confidence intervals or significance tests. Evaluations share benchmark material, and individual prediction records are unavailable for reconstructing sampling uncertainty. Small differences between averages are therefore interpreted in relation to source rounding rather than presented as established superiority. Reproducibility depends on preserving the distinction between source evidence, sample restrictions and calculated summaries. Sandve et al., 2013, emphasize maintaining the information needed to trace computational results. The analytical workbook supports this traceability through separate source cells, retained records, variable definitions and publication identifiers. The methodological contribution is the explicit use of common evaluation coverage to clarify descriptive comparisons, rather than a new estimator or causal identification strategy.

5. Findings:

This analysis examines variation in published detection performance across detector configurations, evaluation languages and social platforms. The findings resolve the three research questions through descriptive comparisons of AUROC, including comparisons restricted to common evaluation groups. The results establish patterns within the benchmark without interpreting those patterns as causal effects or statistically significant differences.

5.1 Sample Composition and Performance Distribution:

The analytical sample contains 1,032 evaluations covering twelve detectors, twenty-two languages and five platforms. Each observation represents one detector evaluated within one language and platform combination. All twelve detectors share the same 86 retained combinations, permitting comparisons of detector performance without differences in language and platform coverage. The source tables contain 288 unavailable language specific cells, which were excluded without imputation. These observations constitute a benchmark published in 2025 rather than repeated observations throughout the stated publication window. Across the retained evaluations, mean AUROC is 0.7892, the median is 0.9400 and the sample standard deviation is 0.2486. Scores range from 0.0500 to 1.0000. The difference between the mean and median reflects substantial lower scoring observations alongside a concentration of higher scores. Consequently, the pooled mean conceals differences between detector configurations that require separate examination.

5.2 Detection Method and Performance Differences:

Table 4 reports detector performance over identical evaluation combinations. The seven configurations adapted to Multi Social have mean AUROC values between 0.9505 and 0.9747. The five previously trained configurations span a wider interval, from 0.3302 to 0.7562. This separation answers research question 1: published performance differs substantially across the evaluated detector configurations under common language and platform coverage.

Table 4: Detection performance across common evaluation combinations

Detector Configuration	Evaluations	Mean AUROC	Standard Deviation	Median	Minimum	Maximum
Llama 3 8b adapted to Multi Social	86	0.9747	0.0233	0.9800	0.8800	1.0000
Mistral 7b v0.1 adapted to Multi Social	86	0.9740	0.0230	0.9800	0.8900	1.0000
Aya 101 adapted to Multi Social	86	0.9721	0.0227	0.9800	0.8900	1.0000
Falcon rw 1b adapted to Multi Social	86	0.9570	0.0355	0.9650	0.8000	1.0000
BLOOMZ 3b adapted to Multi Social	86	0.9560	0.0407	0.9700	0.8100	1.0000
XLM RoBERTa large adapted to Multi Social	86	0.9512	0.0399	0.9600	0.7200	1.0000
mDeBERTa v3 base adapted to Multi Social	86	0.9505	0.0411	0.9600	0.7500	1.0000
BLOOMZ 3b mixed detector	86	0.7562	0.0947	0.7650	0.4600	0.9600
ChatGPT detector RoBERTa Chinese	86	0.7421	0.1025	0.7450	0.3600	0.9500
ruRoBERTaruatd binary	86	0.4578	0.1077	0.4500	0.2300	0.7400
Long former detector	86	0.4490	0.1157	0.4400	0.1600	0.7800
RoBERTa large OpenAI detector	86	0.3302	0.1998	0.2950	0.0500	0.8200

The adapted configurations collectively attain mean AUROC of 0.9622 across 602 evaluations, compared with 0.5470 across 430 evaluations for previously trained configurations. Within each of the 86 common language and platform combinations, the mean across adapted detectors exceeds the mean across previously trained detectors. The average matched difference is 0.4151 AUROC units, with individual combination differences ranging from 0.2683 to 0.5540. The separation extends beyond differences between group averages. In 85 of the 86 combinations, the lowest score among adapted detectors exceeds the highest score among previously trained detectors. The remaining combination is Catalan on Discord, where the two boundaries equal 0.9600. Thus, the group separation is consistently observed across the available evaluation settings. Adapted configurations also exhibit less variation across settings. Their individual standard deviations range from 0.0227 to 0.0411, compared with 0.0947 to 0.1998 among previously trained configurations. However, the difference of 0.0007 between the highest two detector means is smaller

than the reporting increment of the underlying scores. The descriptive ordering does not establish a practically meaningful advantage between those leading configurations. All 188 scores below 0.5000 occur among previously trained configurations. These scores remain unchanged from the source. Their interpretation is limited to the reported scoring orientation because individual predictions were not available for independent verification.

5.3 Language Variation Under Common Platform Coverage:

Language comparisons use the nine languages represented across all five platforms. Each language contributes sixty evaluations comprising the same twelve detectors and five platforms. This restriction removes differences in platform coverage from the comparison while retaining the observed detector composition.

Table 5: Language performance under identical detector and platform coverage

Evaluation Language	Evaluations	Mean AUROC	Mean For Previously Trained Detectors	Mean For Adapted Detectors
English	60	0.8557	0.6860	0.9769
Spanish	60	0.7973	0.5524	0.9723
Catalan	60	0.7960	0.5608	0.9640
Estonian	60	0.7912	0.5368	0.9729
German	60	0.7843	0.5356	0.9620
Portuguese	60	0.7787	0.5160	0.9663
Dutch	60	0.7722	0.5096	0.9597
Czech	60	0.7657	0.4772	0.9717
Romanian	60	0.7635	0.4744	0.9700

English has the highest mean AUROC within the matched language sample, exceeding Romanian by 0.0922 AUROC units. However, the language spread differs markedly between detector families. Previously trained detectors range from 0.4744 for Romanian to 0.6860 for English, a spread of 0.2116. Adapted detectors range from 0.9597 for Dutch to 0.9769 for English, a spread of 0.0171. The smaller spread among adapted configurations indicates greater descriptive consistency across these languages. Czech and Romanian illustrate this distinction: their previously trained detector means remain below 0.5000, while their adapted detector means reach 0.9717 and 0.9700. Research question 2 is therefore resolved as observed language variation whose magnitude depends on the detector configurations being summarized. These comparisons do not identify language resources or linguistic structure as independently measured explanations.

5.4 Platform Variation and Sensitivity To Sample Composition:

Platform comparisons use the same nine languages and twelve detectors, producing 108 evaluations per platform. The matched sample contains 540 evaluations in total. Table 6 presents both unrestricted and matched means to show how coverage affects the observed platform ordering.

Table 6: Platform performance before and after matching language coverage

Platform	Available Evaluations	Unrestricted Mean AUROC	Matched Evaluations	Matched Mean AUROC	Matched Mean For Previously Trained Detectors	Matched Mean for Adapted Detectors
Discord	180	0.7873	108	0.7981	0.5342	0.9865
Gab	252	0.7718	108	0.7786	0.5429	0.9470
Telegram	252	0.8009	108	0.7964	0.5556	0.9684
Twitter	216	0.7920	108	0.7841	0.5264	0.9681
WhatsApp	132	0.7982	108	0.7898	0.5347	0.9721

Telegram has the highest unrestricted mean, whereas Discord has the highest matched mean. This reversal demonstrates that platform rankings depend partly on which languages enter the comparison. Within the matched sample, platform means span 0.7786 to 0.7981, a difference of 0.0194 AUROC units using unrounded calculations. The matched Discord and Gab comparison contains 79 detector and language pairs favouring Discord, twenty sixfavouring Gab and three ties. The mean difference is 0.0194, but individual paired differences range from negative 0.2800 to positive 0.2500. The aggregate advantage therefore coexists with opposing results across individual evaluations. Platform patterns also differ between detector families. Telegram has the highest matched mean among previously trained configurations, while Discord leads among adapted configurations. Research question 3 is resolved as platform variation that remains conditional on detector configuration and language coverage. The results support reporting performance for explicit evaluation settings rather than presenting a pooled platform ranking as universally applicable.

6. Discussion:

A detector ranking changes meaning when the evaluation setting changes. Table 4 shows that configurations adapted to Multi Social consistently outperform previously trained configurations across common language and platform combinations, while Table 6 shows that platform rankings change after matching language coverage. Together, these findings distinguish two properties often compressed into a single performance score: discrimination within observed settings and consistency across settings. Moreno Torres et al., 2012, explain why changes in data distributions complicate classifier comparisons. The contribution of this analysis is a matched descriptive interpretation of published benchmark results that separates configuration differences from differences in evaluation composition. The evidence supports configuration specific performance claims rather than an unconditional hierarchy of detection capability. The separation between detector families in Table 4 is accompanied by narrower language variation among adapted configurations in Table 5. This combination matters because stronger average discrimination does not necessarily imply greater consistency across linguistic settings. Here, both properties occur together within the observed

benchmark. Liang et al., 2023, identify misclassification risks affecting English texts written by nonnative speakers, demonstrating that linguistic variation can influence detector outputs even within one language. The present comparison addresses a different level of variation: performance across languages under common detector and platform coverage. It indicates that language disparities should be evaluated conditionally on detector configuration rather than treated as uniform characteristics of automated detection. The observed convergence among adapted configurations concerns benchmark discrimination and does not establish equivalence in error rates across user populations.

The configuration contrast is more pronounced than the platform contrast within the matched comparisons, but these differences should not be interpreted as decomposed causal contributions. Tables 4 and 6 show that detector selection and platform composition describe distinct aspects of performance variation. Previously trained configurations perform best on Telegram within the matched platform sample, whereas adapted configurations perform best on Discord. Thus, no platform supplies a universally favourable evaluation environment across the observed detector families. Weber Wulff et al., 2023, document limitations in detection tools and sensitivity to the texts being evaluated. This analysis adds a conditional comparison within multilingual social media evaluations: platform suitability changes with the configuration being assessed. The theoretical implication is that transferability is relational, involving both the detector and its evaluation environment, rather than residing in either alone. The ranking reversal in Table 6 exposes evaluation composition as a substantive determinant of comparative conclusions. Telegram leads when all available evaluations are pooled, while Discord leads when languages and detectors are held compositionally comparable. The interpretation therefore changes even though the underlying published scores remain unchanged. Gebru et al., 2021, advocate documentation that makes dataset composition and appropriate uses visible. The present results demonstrate why that information matters analytically: unequal coverage can alter the apparent ordering of platforms. Moreover, the opposing paired differences between Discord and Gab show that an aggregate advantage can coexist with disadvantages for particular detector and language combinations. Reporting coverage alongside performance is therefore necessary for interpreting comparative rankings, rather than merely an administrative feature of dataset documentation.

The international relevance of these findings lies in multilingual evaluation rather than country level comparison. Tables 5 and 6 show that conclusions depend on which linguistic and platform environments are represented, making a single pooled score an incomplete basis for selecting detectors for diverse communication settings. Mitchell et al., 2019, propose model reporting that documents performance under relevant conditions. The evidence here supports a specific implementation of that principle: report detector performance within common language and platform groups, distinguish adapted configurations from previously trained alternatives, and present matched comparisons alongside unrestricted summaries. Decision makers should assess the configuration against its intended evaluation environment before relying on an overall ranking. This redirects comparative assessment from selecting the highest pooled score toward identifying where a detector demonstrates consistent discrimination within documented coverage.

7. Conclusion and Implications:

A detector that leads one benchmark summary may lose that position when the comparison becomes compositionally equivalent. This analysis shows that comparative performance depends jointly on the evaluated configuration and the linguistic and communication settings represented in the evidence. Adapted systems display stronger discrimination and greater consistency across observed settings, while pooled rankings remain sensitive to coverage. The contribution is a matched reinterpretation of published evaluations that distinguishes consistent comparative advantages from conclusions shaped by sample composition. This refines the application of domain adaptation reasoning by showing why transferability assessments must specify the environments being compared. The evidence establishes conditional descriptive patterns rather than causal pathways. For managers selecting detection systems, the implication is to assess relevant operating environments before relying on aggregate rankings. Procurement and governance procedures should require documented coverage, comparable evaluation groups and disclosure of unavailable combinations. Operational reporting should present both unrestricted summaries and matched comparisons, making changes in comparative conclusions visible. These practices can improve the evidential basis for selecting systems across multilingual communication environments. Their wider social relevance lies in supporting more accountable decisions about automated authorship assessment, while keeping claims proportionate to the populations, settings and performance measures actually evaluated.

The aggregate benchmark structure creates specific opportunities for further investigation. Access to individual predictions would permit uncertainty estimation, assessment of error types and evaluation at operational thresholds. Published rounding limits distinctions between closely ranked systems, motivating replication with more precise scores. Shared evaluation material calls for designs that explicitly address dependence across comparisons. Broader coverage could establish whether the observed ranking changes persist across additional languages, platforms and generators. Controlled comparisons separating architecture from adaptation procedures would clarify the sources of performance differences. Repeated evaluations across technological releases could then assess whether the observed comparative consistency persists as generation and detection systems evolve.

References:

1. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine Learning, 79(1-2), 151-175. https://doi.org/10.1007/s10994-009-5152-4
2. Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587-604. https://doi.org/10.1162/tacl_a_00041
3. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. https://doi.org/10.1145/3458723
4. Gehrmann, S., Strobelt, H., & Rush, A. (2019). GLTR: Statistical detection and visualization of generated text. In M. R. Costa-jussà& E. Alfonseca (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational

Linguistics: System Demonstrations (pp. 111-116). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-3019

5. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 6282-6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560
6. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779
7. Macko, D., Kopal, J., Moro, R., & Srba, I. (2024). Multi Social (Version v1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13846152
8. Macko, D., Kopál, J., Moro, R., & Srba, I. (2025). Multi Social: Multilingual benchmark of machine-generated text detection of social-media texts. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 727-752). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.36
9. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596
10. Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. Pattern Recognition, 45(1), 521-530. https://doi.org/10.1016/j.patcog.2011.06.019
11. Peng, R. D. (2011). Reproducible research in computational science. Science, 334(6060), 1226-1227. https://doi.org/10.1126/science.1213847
12. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432
13. Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. PLOS Computational Biology, 9(10), Article e1003285. https://doi.org/10.1371/journal.pcbi.1003285
14. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. International Journal for Educational Integrity, 19, Article 26. https://doi.org/10.1007/s40979-023-00146-z