Peer Review

Sobrier, Maxime. 2026. “Efficient Language Identification in Monolingual and Multilingual Documents.” *Journal of High School Science* 10 (2): 1–23. https://doi.org/10.64336/001c.160011

A very good effort. Congratulations. I do have some comments and concerns that will need to be addressed.
1.Your main novelty claim is multi-label document-level LID. However, you report only overall accuracy for the top two languages in Table 8. This is not sufficient. I want to see you evaluate the Multilingual Webpages dataset using Precision, Recall, F1 (micro & macro). Treat each document as a multi-label classification problem. Count a language as correct if:it appears in the ground truth and its score ≥ 0.8 × max score. For comparison, add fastText: take top-k predictions where probability ≥ threshold, CLD3: same (if available), GlotLID: top-k probabilities.
2.You have used (multiple) heuristics in your algorithm. To assess which ones actually matter, you will need to perform ablation studies. At a minimum, Run WiLI-2018 and Multilingual Webpages with no differential weighting, no English down-weighting, single-label normalization (force max language only) and keep everything else fixed. Report metrics drop (accuracy, F1) relative to full model. This will provide evidence (or not) that the multi-label design, the differential weighting are not cosmetic, but essential.
3.Perform a threshold sensitivity analysis (say from 0.6 to 0.9) and tabulate precision, recall and F1 on multilingual performance. This will demonstrate whether your chosen threshold of 0.8 adequately captures the Tradeoff between over-detection and under-detection, that the method is not brittle and that it is stable to small perturbations.
4.No confidence intervals, variance estimates, or hypothesis tests are reported; classification comparisons lack paired tests (e.g., McNemar’s) and per-language macro metrics, limiting inferential strength.
5.For benchmarking, clarify whether metrics are macro- or micro-averaged and whether per-language sample counts were normalized; consider stratified, balanced evaluations to avoid dominance by high-resource languages.
6.Experimental design is generally sound with broad datasets and strong baselines; to improve replicability, report hardware/software details for timing, provide the exact GitHub link, document training data splits and preprocessing scripts, and [add ablations and threshold sensitivity (as commented on previously)].

_____

Journal of High School Science Comments
Efficient Language Detection in Multilingual Documents
Maxime Sobrier

1. Your main novelty claim is multi-label document-level LID. However, you report only overall accuracy for the top two languages in Table 8. This is not sufficient. I want to see you evaluate the Multilingual Webpages dataset using Precision, Recall, F1 (micro & macro). Treat each document as a multi-label classification problem. Count a language as correct if:it appears in the ground truth and its score ≥ 0.8 × max score. For comparison, add fastText: take top-k predictions where probability ≥ threshold, CLD3: same (if available), GlotLID: top-k probabilities.

Table 8 was fully revised to evaluate the Multilingual Webpages dataset as a true multi-label document-level classification task. Each document was treated as a set of language labels. A language was counted as correctly predicted if and only if it appeared in the ground truth and its predicted score was ≥ 0.8 × the maximum score for that document, exactly following the reviewer’s proposed criterion.

Precision, Recall, F1-score, Accuracy, and Specificity were reported for this dataset. Micro and macro averages were reported for each metric.
Other libraries cannot detect multiple languages; this was clarified in section 4.2.2.

2. You have used (multiple) heuristics in your algorithm. To assess which ones actually matter, you will need to perform ablation studies. At a minimum, Run WiLI-2018 and Multilingual Webpages with no differential weighting, no English down-weighting, single-label normalization (force max language only) and keep everything else fixed. Report metrics drop (accuracy, F1) relative to full model. This will provide evidence (or not) that the multi-label design, the differential weighting are not cosmetic, but essential.

A new ablation study was added as Section 5.6 to quantify the contribution of each major design choice. The full model was evaluated against the following controlled variants, with all other components held constant:

- No differential weighting
- Differential weighting
- Differential weighting with English bias

These ablations were evaluated on WiLI-2018, FLORES+ and Multilingual Webpages with single-label normalization.

3. Perform a threshold sensitivity analysis (say from 0.6 to 0.9) and tabulate precision, recall and F1 on multilingual performance. This will demonstrate whether your chosen threshold of 0.8 adequately captures the Tradeoff between over-detection and under-detection, that the method is not brittle and that it is stable to small perturbations.

A threshold sensitivity analysis was added in Section 5.5 and summarized in Table 9. Multilingual document performance on the Multilingual Webpages dataset using thresholds ranging from 0.6 to 0.9 was evaluated.

4. No confidence intervals, variance estimates, or hypothesis tests are reported; classification comparisons lack paired tests (e.g., McNemar's) and per-language macro metrics, limiting inferential strength.

Statistical significance testing was incorporated throughout the evaluation. Paired bootstrap resampling with 2,000 iterations was used to compare libraries.
For each iteration, 100 languages were sampled with replacement, and the mean difference in macro accuracy between methods was computed. 95% confidence intervals were derived from the 2.5th and 97.5th percentiles, and intervals excluding zero are interpreted as statistically significant differences.
These results were reported in Tables 3–7.
Due to space constraints, full per-language precision, recall, and F1 metrics were not included in the paper. However, all per-language results for all datasets and libraries are publicly released in a GitHub repository:
https://github.com/MaximeSobrier/language-detector-research
This repository is referenced in Section 5, ensuring transparency and enabling independent verification.

5. For benchmarking, clarify whether metrics are macro- or micro-averaged and whether per-language sample counts were normalized; consider stratified, balanced evaluations to avoid dominance by high-resource languages.

All evaluation metrics were explicitly reported as both macro- and micro-averaged throughout the paper.
Bootstrap resampling with 95% confidence intervals was applied to all reported metrics, further reducing sensitivity to dataset imbalance.

6. Experimental design is generally sound with broad datasets and strong baselines; to improve replicability, report hardware/software details for timing, provide the exact GitHub link, document training data splits and preprocessing scripts, and [add ablations and threshold sensitivity (as commented on previously)].

Table 1 caption now includes the hardware used to compare the libraries. All datasets, benchmarks, and evaluation scripts are publicly available at https://github.com/MaximeSobrier/language-detector-research, enabling full reproducibility. The raw results used for all tables in this research is also included. The repository is mentioned in section 5.
Replicability was significantly strengthened. The caption of Table 1 now reports full hardware details used for timing experiments, including CPU model and memory.
All datasets, preprocessing scripts, evaluation code, raw results, and ablation configurations are publicly available at:
https://github.com/MaximeSobrier/language-detector-research
The repository is explicitly referenced in Section 5, and includes the raw outputs used to generate every reported result.

_____
Thank you for addressing my comments. However, some comments remain un-addressed and some inconsistency remains. Please address.
1.Add multi-label baselines (thresholded fastText / CLD3): The core claim is still not properly validated. The paper claims:“multi-label document-level LID”, But: Multilingual dataset, No baseline comparison, No adapted baselines (top-k / thresholded probabilities). The ablation uses top-1 prediction only. This is contradictory. The core novelty (multi-label detection) is not rigorously benchmarked.
2.methodological inconsistency:You cannot validate a multi-label system using single-label evaluation. You have 2 evaluation protocols: two different evaluation protocols:Section 5.5: Multi-label evaluation (correct) Section 5.6 (ablation): “used to identify the top language”
3.De-emphasize accuracy metric “up to 98% accuracy surpassing existing libraries” and others, because it is dominated by negatives, inflated by small label sets and not meaningful for multi-label tasks. F1 is the real metric. The specificity metric is expected and non-informative, please replace with Hamming loss, subset accuracy or equivalent.
4.Scoring method seems heuristic rather than based on first-principles. For example, (word freq × length², etc.) is used without justification. Why squared length? Why only first 2 occurrences?
5.Claim about baselines is misleading. You state,“These models cannot detect multiple languages” .This is incorrect:fastText outputs probability distribution and can probably execute multi-label via thresholding. Please check all over-reach claims.
6.Explicitly state that you do not test low-resource languages, or subtle mixing or code switching.

_____
Journal of High School Science Comments 2
Efficient Language Detection in Monolingual and Multilingual Documents
Maxime Sobrier

Thank you for addressing my comments. However, some comments remain un-addressed and some inconsistency remains. Please address.

1. Add multi-label baselines (thresholded fastText / CLD3): The core claim is still not properly validated. The paper claims: “multi-label document-level LID”, But: Multilingual dataset, No

baseline comparison, No adapted baselines (top-k / thresholded probabilities). The ablation uses top-1 prediction only. This is contradictory. The core novelty (multi-label detection) is not rigorously benchmarked.

Thank you for this comment. I agree that the earlier wording was too strong. My intention is not to suggest that fastText, CLD3, GlotLID, or the BERT-based baseline cannot, in principle, be adapted for multi-label prediction. Rather, the point is that the published models evaluated in this study were trained for single-label language identification on monolingual documents. Their training objective assumes a single target language per input, and their standard prediction behavior is therefore optimized to recover the dominant language of a document rather than all languages present in multilingual documents.

To verify this point, I evaluated thresholded versions of these baselines on multilingual documents. These additional experiments (Tables 7.1–7.4) show that the models typically recover the most prevalent language but do not reliably identify secondary languages. This limitation is particularly evident in their lower recall on multilingual documents. In other words, the issue is not that these models cannot produce multiple labels after post-processing, but rather that they were not trained for multilingual document detection, and their predictions remain dominated by the primary language in the text.

To verify this point, I evaluated thresholded versions of these baselines on multilingual documents. These additional experiments (Tables 7.1 - 7.4) show that the models typically recover the most prevalent language but do not reliably identify secondary languages. This limitation is particularly evident in their lower recall of multilingual documents. In other words, the issue is not that these models cannot produce multiple labels after post-processing, but rather that they were not trained for multilingual document detection, and their predictions remain dominated by the primary language in the text.

I have revised the paper to avoid overstatement in Section 4.2.2. Instead of stating that these baselines “cannot detect multiple languages,” it now states that they are single-label LID models by design and training, and that, when applied to multilingual documents, they tend to emphasize the dominant language rather than recover the full set of languages present.

Table 7.1. Effect of decision threshold on multilingual language detection with fastText.

Threshold	Averaging	Precision	Recall	Accuracy	F1 score
0.6	Macro	47.35	37.02	98.49	37.46
	Micro	96.77	65.94	99.59	78.43
0.7	Macro	47.73	34.49	98.32	35.99
	Micro	97.32	61.26	99.54	75.19
0.8	Macro	44.28	30.04	98.15	32.78
	Micro	97.48	56.83	99.49	71.80
0.9	Macro	42.05	28.33	97.99	31.41
	Micro	97.64	52.80	99.45	68.54

Table 7.2. Effect of decision threshold on multilingual language detection with GlotLID.

Threshold	Averaging	Precision	Recall	Accuracy	F1 score
0.6	Macro	53.33	42.42	97.79	39.47
	Micro	95.10	49.99	98.99	65.53
0.7	Macro	53.94	41.30	99.74	39.23

	Micro	95.48	45.84	98.97	64.36
0.8	Macro	55.27	39.99	97.69	39.18
	Micro	95.94	47.17	98.95	63.24
0.9	Macro	55.42	39.12	97.65	38.61
	Micro	96.43	45.88	98.93	62.18

Table 7.3. No effect of decision threshold on multilingual language detection with BERT.

Threshold	Averaging	Precision	Recall	Accuracy	F1 score
0.6	Macro	77.87	63.80	97.58	66.81
	Micro	92.73	46.58	98.86	62.01
0.7	Macro	77.87	63.80	97.58	66.81
	Micro	92.73	46.58	98.86	62.01
0.8	Macro	77.87	63.80	97.58	66.81
	Micro	92.73	46.58	98.86	62.01
0.9	Macro	77.87	63.80	97.58	66.81
	Micro	92.73	46.58	98.86	62.01

Table 7.4. Effect of decision threshold on multilingual language detection with CLD3.

Threshold	Averaging	Precision	Recall	Accuracy	F1 score
0.6	Macro	39.06	69.17	97.42	42.84
	Micro	69.68	59.12	98.69	63.96
0.7	Macro	43.42	69.11	97.53	46.58
	Micro	73.46	58.63	98.77	65.21
0.8	Macro	47.31	68.72	97.61	49.09
	Micro	78.17	57.67	98.85	66.37
0.9	Macro	51.22	65.20	97.57	50.57
	Micro	80.84	54.73	98.85	65.27

In the previous round of review (Comments #2), you requested to create the ablation study that includes “Multilingual Webpages with no differential weighting [...] single-label normalization (force max language only)”. I understood this as using the Multilingual Webpages dataset to detect the top language only. Table 10 has been updated to include all detected languages.

2. methodological inconsistency: You cannot validate a multi-label system using single-label evaluation. You have 2 evaluation protocols: two different evaluation protocols: Section 5.5: Multi-label evaluation (correct) Section 5.6 (ablation): “used to identify the top language”.

In the revised version, I updated the ablation study to use the same multi-label evaluation protocol as the main multilingual experiments. Table 10 now reports all detected languages instead of only the top predicted language.

3. De-emphasize accuracy metric “up to 98% accuracy surpassing existing libraries” and others, because it is dominated by negatives, inflated by small label sets and not meaningful for multi-label tasks. F1 is the real metric. The specificity metric is expected and non-informative, please replace with Hamming loss, subset accuracy or equivalent.

I have updated the paper to use F1 score in place of accuracy as the overall metric.

For monolingual documents, subset accuracy reduces to exact-match accuracy, but Hamming loss is not identical to standard accuracy: under the usual binary-indicator definition, an incorrect single-label prediction produces one false positive and one false negative across the label vector. For multilingual documents, Hamming loss can remain numerically small even for poor predictions because the error is normalized by the total number of supported languages. For example, with approximately 100 candidate languages, a bilingual document for which both predicted languages are incorrect yields a Hamming loss of 4 / 100 = 0.04. This illustrates why Hamming loss is useful as a complementary metric, but less discriminative than F1 score in our setting.

I computed the cost-sensitive loss with different benchmarks and models. Results is available at https://github.com/MaximeSobrier/language-detector-research. The loss is consistently below 0.1 across all tests.

I removed Section 5.4 (Accuracy and Specificity), as well as the Specificity results from Tables 6 and 7.

4. Scoring method seems heuristic rather than based on first-principles. For example, (word freq × length², etc.) is used without justification. Why squared length? Why only first 2 occurrences?

I agree that the scoring function is heuristic rather than derived from a first-principles probabilistic formulation, and we have revised the manuscript to make that clearer. As noted at the end of Section 3.1, parameters such as the length exponent and the maximum number of counted occurrences per word were selected empirically rather than analytically.

To address this concern, we expanded the ablation study to explicitly evaluate the impact of these parameters (Tables 10 and 11). The revised paper reports the effects of varying the word-length weighting and the occurrence cap on detection performance. This revision clarifies that choices such as using a squared word-length term and counting only the first two occurrences were not intended as theoretically optimal constants, but as empirically selected settings that balance performance and robustness.

5. Claim about baselines is misleading. You state,“These models cannot detect multiple languages” .This is incorrect:fastText outputs probability distribution and can probably execute multi-label via thresholding. Please check all over-reach claims.

Thank you for this comment. I agree that the original wording was too strong and could be misleading. In the revised manuscript, I removed statements such as “these models cannot detect multiple languages” and replaced them with a more precise formulation in Section 4.2.2.

The baseline models considered in our comparison, including fastText, CLD3, GlotLID, and the BERT-based baseline, were primarily designed and trained as single-label document-level language identification systems, typically using monolingual training data and a single target label per input. As a result, their standard use is to return the most likely language for a document rather than a set of languages.

The results show that, although thresholding allows these single-label baselines to produce multiple candidate languages, their multilingual document-level performance remains substantially below that of LDW (Tables 7.1–7.4 in Comment 1). This suggests that the limitation is not that they are impossible to adapt, but rather that their original training objective and score semantics are not well aligned with true multi-label document-level language identification.

I also clarified an important distinction in the manuscript: for fastText and similar classifiers trained with a single-label objective, the output probabilities are normalized under the assumption of one target class per input. Consequently, these scores should not be interpreted as language proportions within a multilingual document. In contrast, LDW is designed to estimate the contribution of multiple languages at the document level, making it better suited to multilingual and mixed-language inputs.

6. Explicitly state that you do not test low-resource languages, or subtle mixing or code switching.

I have updated in Section 7 Limitations (3rd paragraph) to clarify that our experiments do not evaluate low-resource languages, do not target subtle intra-document language mixing or code-switching, and therefore should not be interpreted as covering those settings.

Thank you for addressing my comments in such a thorough manner. Nicely done.
I have several lingering minor concerns before I can accept.
1.your algorithm will achieve a high F1 even if (say) it identifies a document as 90% English and 10% French, whereas the actual document may contain 10% English and 90% French - correct? In other words, you are not evaluating proportions/distributions, only sets.
That being the case, please check the manuscript for over-reaching claims such as “relative amount”; and replace with “…..Scores are relative indicators (not calibrated proportions)…..”; and better yet, also include - a blanket caveat such as "provides relative scoring that correlates with language prominence, but is not calibrated as a proportion estimate”.
The following concerns involve ADDING CONTENT TO THE LIMITATIONS SECTION. Expand the “limitations” section of the manuscript to include:
2.Add this content " The scoring mechanism is not calibrated to represent true language proportions. While relative scores correlate with language prominence, they should not be interpreted as precise estimates of content distribution without further calibration." or equivalent.
3.Append the sentence “Many of the benchmarks employed in this study are unbalanced and do not encompass all languages supported by the models.” with “…, therefore improvements against these benchmarks cannot be construed as being fair comparisons…” Also add the following sentence before the above appended sentence "Comparisons with baseline models are constrained by differences in training data, architecture, and task formulation, particularly between single-label and multi-label language identification approaches. "
4.Since you have no clear separation of tuning versus evaluation data, add this sentence 'Extensive parameter tuning was conducted across multiple datasets, which may introduce a risk of overfitting. Future work should include evaluation on strictly held-out datasets and cross-domain validation."
5.You already state “Frequency tables come from CC-100 (web crawl)”. Make this further explicit by adding this sentence “The model relies on frequency distributions derived from web-based corpora. Its performance under domain shift (e.g., technical, medical, or legal text) has not been systematically evaluated.”
6.Add the following sentence
Since script does not necessarily equal language, add “Character-based scoring may bias detection toward writing systems rather than linguistic identity, potentially limiting performance for languages sharing scripts.”
7.Add, “The robustness of the model to noisy input, extremely short text, or OCR artifacts has not been systematically evaluated.”
8.Add, “The model relies on empirically derived heuristics, including weighting schemes and scoring functions, which may not generalize optimally across all languages or datasets.”

Maxime Sobrier

Thank you for addressing my comments in such a thorough manner. Nicely done.
I have several lingering minor concerns before I can accept.

7. your algorithm will achieve a high F1 even if (say) it identifies a document as 90% English and 10% French, whereas the actual document may contain 10% English and 90% French - correct? In other words, you are not evaluating proportions/distributions, only sets.
That being the case, please check the manuscript for over-reaching claims such as “relative amount”; and replace with “…..Scores are relative indicators (not calibrated proportions)…..”; and better yet, also include - a blanket caveat such as "provides relative scoring that correlates with language prominence, but is not calibrated as a proportion estimate”.

Thank you for the kind words. I see the latest version of the paper is much improved, thanks to the thorough reviews.
You are correct that I have not measured the proportion of text. Some languages are more verbose than others, and some languages have longer words than others, so it is hard to define what 10% in one language really means. I have updated the introduction in Section 1 using the wording you suggested.

The following concerns involve ADDING CONTENT TO THE LIMITATIONS SECTION. Expand the “limitations” section of the manuscript to include:
8. Add this content " The scoring mechanism is not calibrated to represent true language proportions. While relative scores correlate with language prominence, they should not be interpreted as precise estimates of content distribution without further calibration." or equivalent.

I have added this sentence to the 4th paragraph of Section 7, Limitations.

9. Append the sentence “Many of the benchmarks employed in this study are unbalanced and do not encompass all languages supported by the models.” with “…, therefore improvements against these benchmarks cannot be construed as being fair comparisons…” Also add the following sentence before the above appended sentence "Comparisons with baseline models are constrained by differences in training data, architecture, and task formulation, particularly between single-label and multi-label language identification approaches. "

I have added the two sentences to the 2nd paragraph of Section 7, Limitations.

10. Since you have no clear separation of tuning versus evaluation data, add this sentence 'Extensive parameter tuning was conducted across multiple datasets, which may introduce a risk of overfitting. Future work should include evaluation on strictly held-out datasets and cross-domain validation."

I have added the two sentences to the 5th paragraph of Section 7, Limitations.
Before running the benchmarks for this research, I used this library in a different project (https://ieeexplore.ieee.org/document/10874070). I tuned the original algorithm against a couple of million web pages, and compared the results with a couple of other libraries. But the process was not as rigorous as using a well-established ground truth from different benchmarks.

11. You already state “Frequency tables come from CC-100 (web crawl)”. Make this further explicit by adding this sentence “The model relies on frequency distributions derived from web-based corpora. Its performance under domain shift (e.g., technical, medical, or legal text) has not been systematically evaluated.”

I have added these sentences to the 3rd paragraph about the CC-100 dataset in Section 7, Limitations.

12. Add the following sentence
Since script does not necessarily equal language, add “Character-based scoring may bias detection toward writing systems rather than linguistic identity, potentially limiting performance for languages sharing scripts.”.

I have added this statement to the 6th paragraph of Section 7, Limitations.

13. Add, “The robustness of the model to noisy input, extremely short text, or OCR artifacts has not been systematically evaluated.”

I have added this statement alongside the limitations on low-resource languages and other domains, at the end of the 3rd paragraph in Section 7, Limitations.

14. Add, “The model relies on empirically derived heuristics, including weighting schemes and scoring functions, which may not generalize optimally across all languages or datasets.”

I have added this statement alongside Comment 4, in the 5th paragraph of Section 7, Limitations.

_____

Thank you for addressing my comments. Accepted.