Language Identification using Character N-grams

A Reimplementation of Cavnar & Trenkle's 1994 Method

Shivam Singh

April 21, 2025

Abstract

This report presents a reimplementation of the character n-gram ranking approach for language identification, based on the Cavnar and Trenkle (1994) method. The performance of this implementation is compared to the modern `langdetect` library on a multilingual dataset. The effect of varying n-gram sizes is evaluated, and common error cases such as confusion between Italian and Spanish are analyzed. The experiments show that the approach remains competitive and highly interpretable.

1 Introduction

Language identification is a foundational task in natural language processing. It is commonly used in preprocessing pipelines, search engines, and language-specific NLP models. In this project, the classic n-gram based approach proposed by Cavnar and Trenkle was reimplemented. This method ranks languages based on the "out-of-place" distance between n-gram frequency profiles.

2 Implementation

2.1 Method Overview

The algorithm builds a character n-gram profile for each language and compares it with the profile of a test text. The distance is computed based on rank differences of matching n-grams and a penalty for missing ones.

2.2 Code Snippet: LanguageIdentifier Class

Listing 1: Core Implementation

```
class LanguageIdentifier:
    def __init__(self, n=3, top_k=400, penalty=400):
        ...
```

2.3 Preprocessing

Text is normalized by lowercasing and collapsing whitespace.

3 Dataset

The `papluca/language-identification` dataset available on the HuggingFace Datasets Hub was used. It contains labeled text samples in 21 languages for supervised language identification.

For this project, five languages were selected for evaluation:

```
['en', 'de', 'fr', 'es', 'it']
```

The dataset was used in its entirety for evaluation only. No part of it was used to train the language profiles. Instead, language profiles were built using external text corpora consisting of approximately 10,000 words per language. The text lengths used to build each profile were:

- English (en): 9294 words
- German (de): 10701 words
- French (fr): 10724 words
- Italian (it): 10017 words
- Spanish (es): 10308 words

All evaluation was conducted on the full dataset including train, validation, and test splits combined, to maximize statistical robustness.

4 Evaluation Methodology

- Used all dataset splits merged for evaluation
- Compared against the `langdetect` library
- Metrics included overall accuracy, per-language accuracy, and confusion matrix
- Tested effect of varying n-gram sizes (n = 3, 4, 5)

5 Results

5.1 N-gram Size Comparison

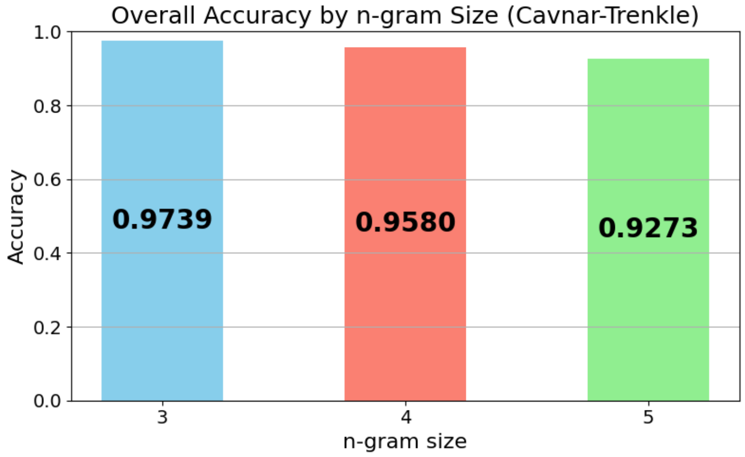


Figure 1: Overall Accuracy by n-gram Size (Cavnar-Trenkle)

The best performance was achieved using 3-grams, with an accuracy of 0.9739. Therefore, the following results focus on the implementation using n=3.

5.2 Overall Accuracy

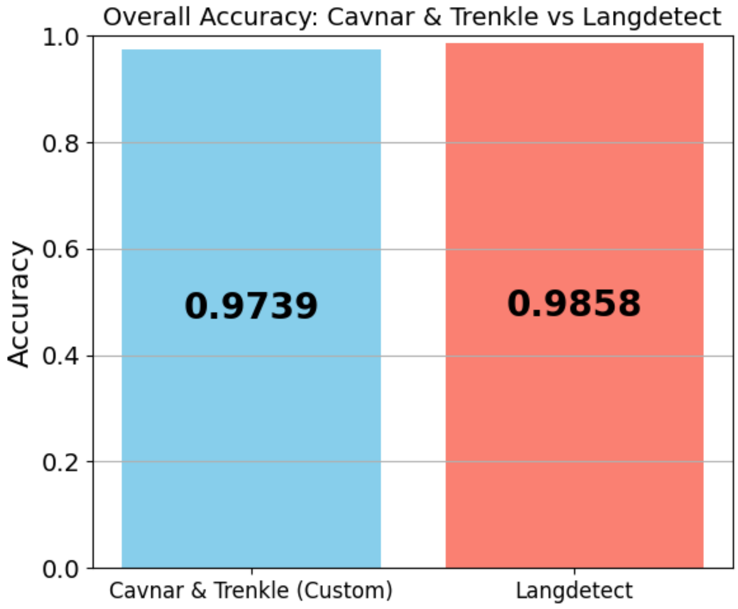


Figure 2: Overall Accuracy: Cavnar & Trenkle vs Langdetect (n=3)

5.3 Per-Language Accuracy

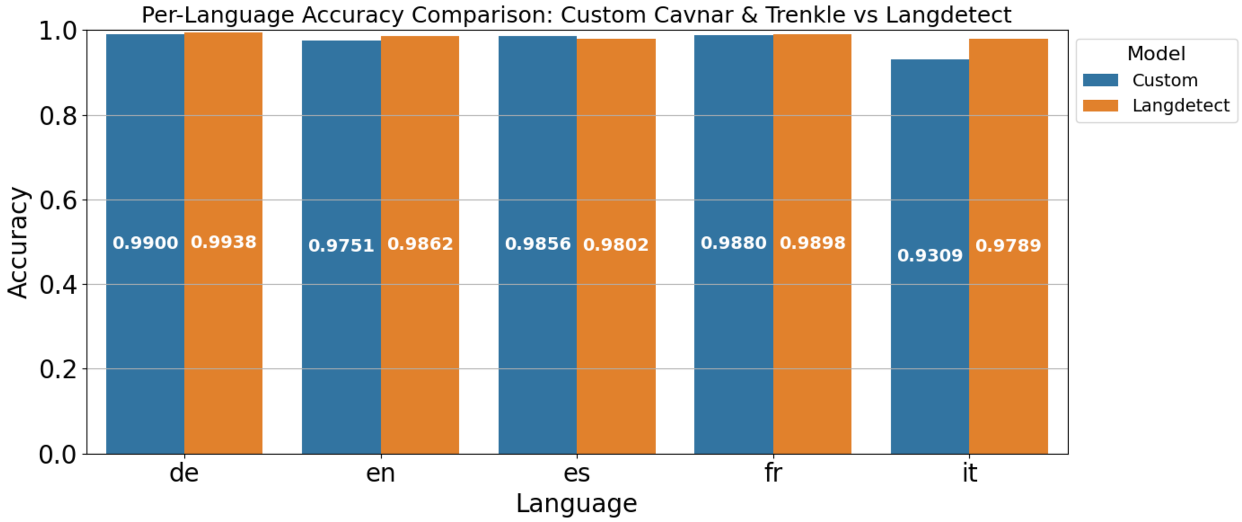
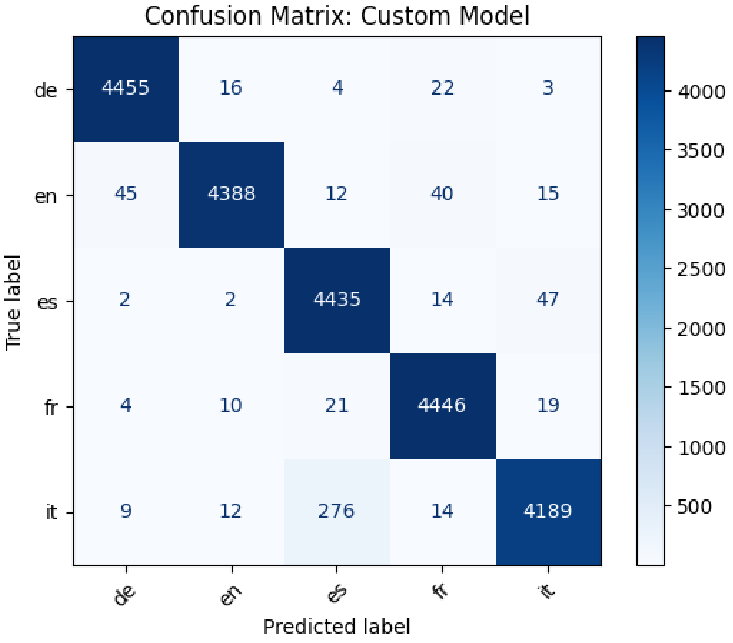


Figure 3: Accuracy Breakdown by Language (n=3)

5.4 Confusion Matrix



	de	en	es	fr	it
de	4455	16	4	22	3
en	45	4388	12	40	15
es	2	2	4435	14	47
fr	4	10	21	4446	19
it	9	12	276	14	4189

Figure 4: Confusion Matrix for Custom Model (n=3)

5.5 Sentence Length Analysis for Italian Misclassifications

To further analyze the confusion between Italian and Spanish, a sentence length distribution was examined. The histogram below compares the lengths of correctly classified Italian sentences with those misclassified as Spanish.

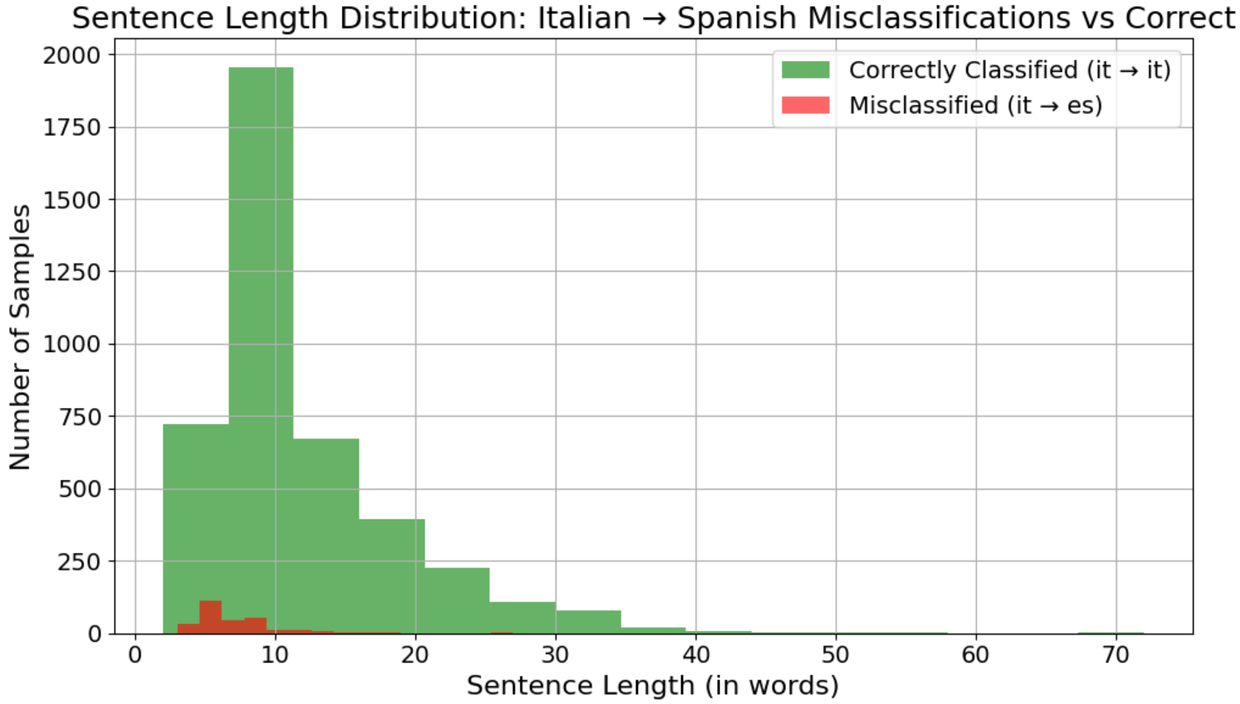


Figure 5: Sentence Length Distribution: Correctly Classified vs Misclassified Italian Samples

The analysis reveals that the majority of misclassified Italian samples were short sentences, typically between 3 to 7 words in length. In contrast, correctly classified Italian samples tended to be longer, often exceeding 10 words. This pattern indicates that the Cavnar & Trenkle model relies on sufficient n-gram context to make accurate predictions.

- **Average sentence length (Correctly classified)**: 11.77 words
- **Average sentence length (Misclassified as Spanish)**: 7.09 words

This suggests that short sentences do not provide enough unique character n-gram patterns to confidently distinguish between similar Romance languages such as Italian and Spanish. Providing more context through longer input appears to significantly reduce misclassification errors.

6 Conclusion

The reimplemented Cavnar and Trenkle model performs well in a modern setting. Frequent misclassification of short Italian sentences as Spanish was observed, likely due to shared vocabulary between Romance languages and lack of contextual n-gram information in short texts.

"Un uomo sta correndo." (Italian) → Predicted: Spanish

Future improvements could include support for variable-length n-grams (e.g., 1–5 grams), introducing confidence thresholds for ambiguous cases, applying discriminative weighting such as TF-IDF for rare n-grams, and better handling of short texts through heuristic approaches.

Overall, the method shows strengths in transparency and interpretability, and with some refinements, it can serve as a solid baseline for low-resource or explainable NLP tasks.

References

- Cavnar, W. B., & Trenkle, J. M. (1994). N-Gram-Based Text Categorization.
- `https://huggingface.co/datasets/papluca/language-identification`
- `https://pypi.org/project/langdetect/`