Language Detection Using NLP & Machine Learning

Identifying languages from text using Multilingual Models

Why Language Detection? & Where It's Used

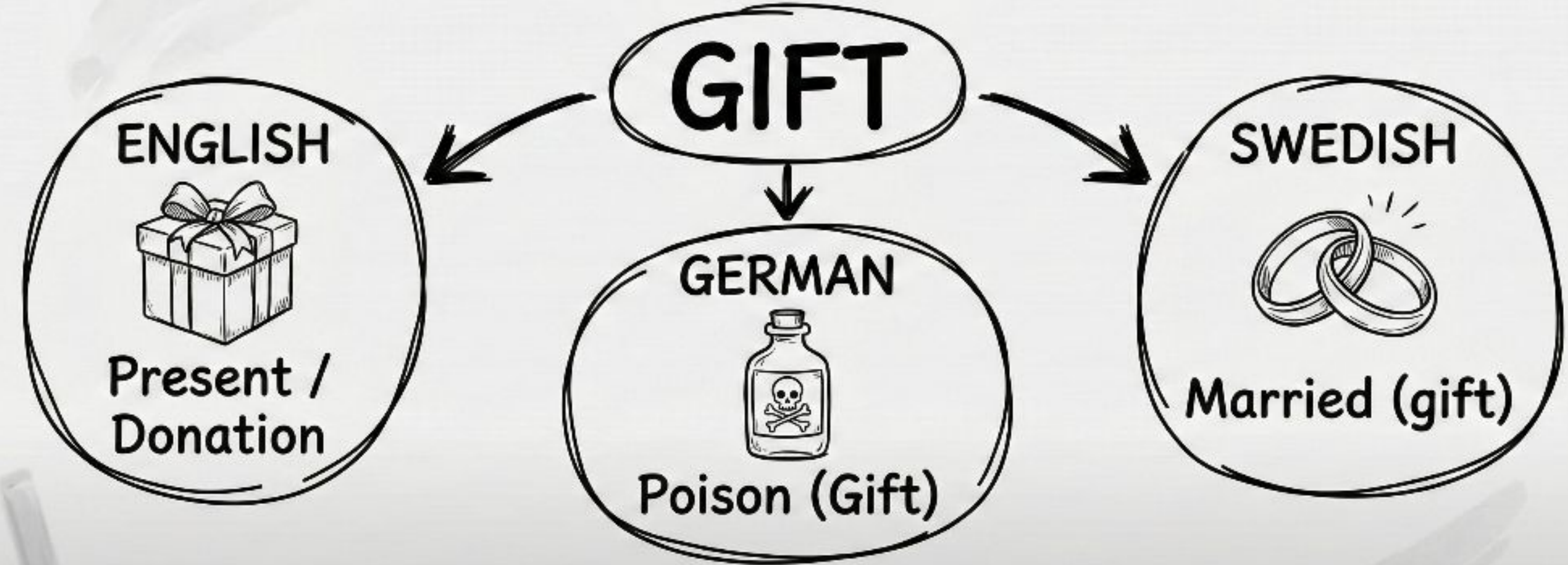
WHY IT'S IMPORTANT

- Multilingual Search & Relevance
- Content Moderation & Safety
- Seamless Machine Translation

COMMON APPLICATIONS

- Search Engines (e.g., Google)
- Social Media Platforms
- Chatbots & Virtual Assistants
- Document Processing & Analytics

The “False Friend” Problem: Same Word, Different Meaning



GIFT

ENGLISH

Present / Donation

GERMAN

Poison (Gift)

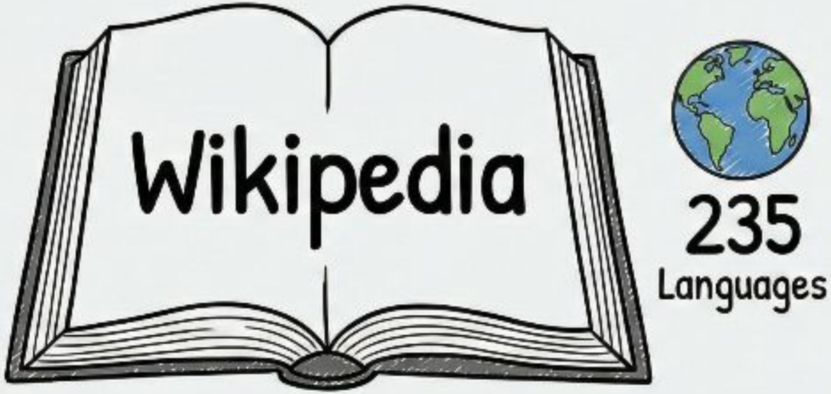
SWEDISH

Married (gift)

Why Rules Fail: Context is Key for Accurate Detection. Machine Learning Helps Discern Meaning.

Dataset Used: WiLI-2018 (Wikipedia Language Identification)

235 Languages, Wikipedia Sentences, ISO-639 Codes



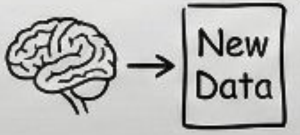
- Source: Wikipedia Extracts
- Format: Text + Language Label
- Covers 235 Languages (ISO-639)
- Used for Training & Evaluation

Q: Why Train a Model If We Already Have a Labeled Dataset?

A: Why Training Is Necessary

Dataset provides **EXAMPLES**, not **RULES**. Model must learn **PATTERNS**.

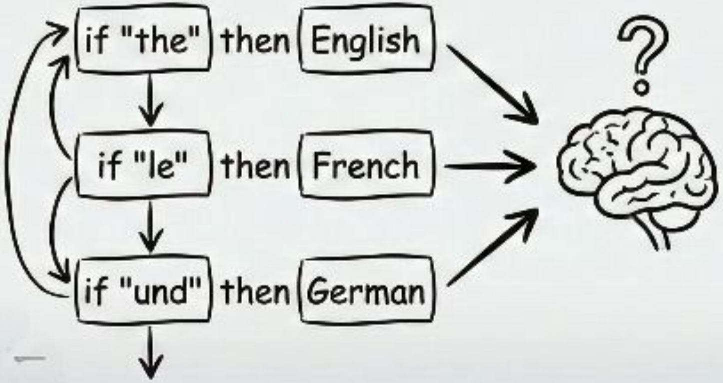
Many languages share similar letters/words. Model learns to **DIFFERENTIATE**.



Training allows the model to **GENERALIZE** to new, unseen sentences.

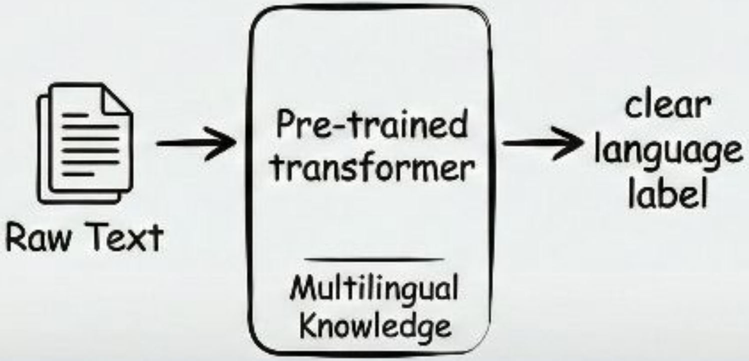
Traditional Models vs. DistilBERT

Traditional Rule-Based (Fails)



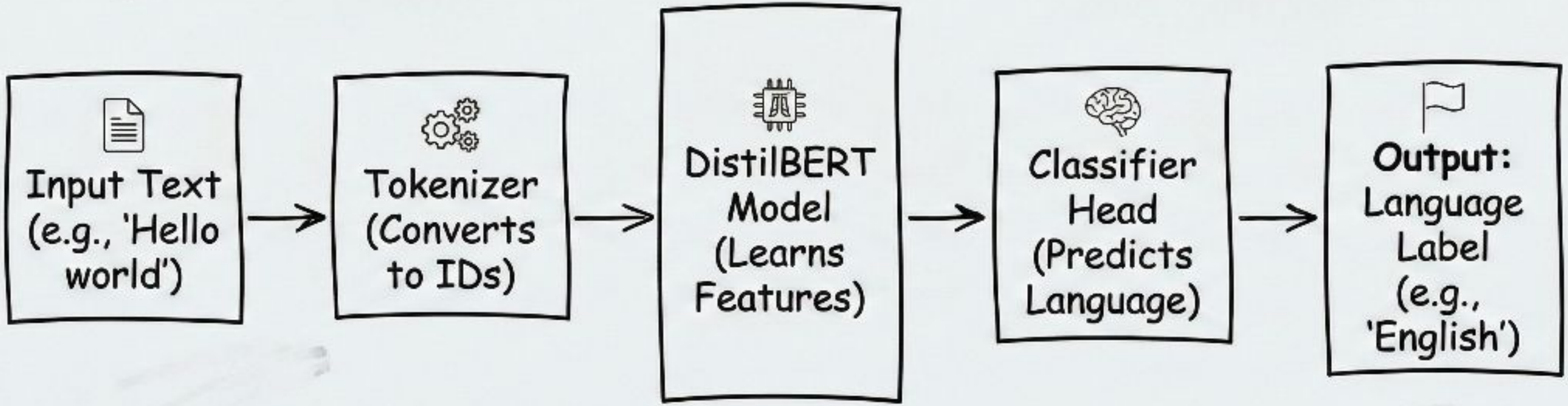
Complex rules, struggles with nuances & shared words.

DistilBERT (Multilingual Transformer)



Learns context & features, generalizes well across languages.

Language Detection Pipeline (Training Flow)



Input Text (e.g., 'Hello world')

Tokenizer (Converts to IDs)

DistilBERT Model (Learns Features)

Classifier Head (Predicts Language)

Output: Language Label (e.g., 'English')

End-to-end process for training the model to identify languages.

Training & Test Observations

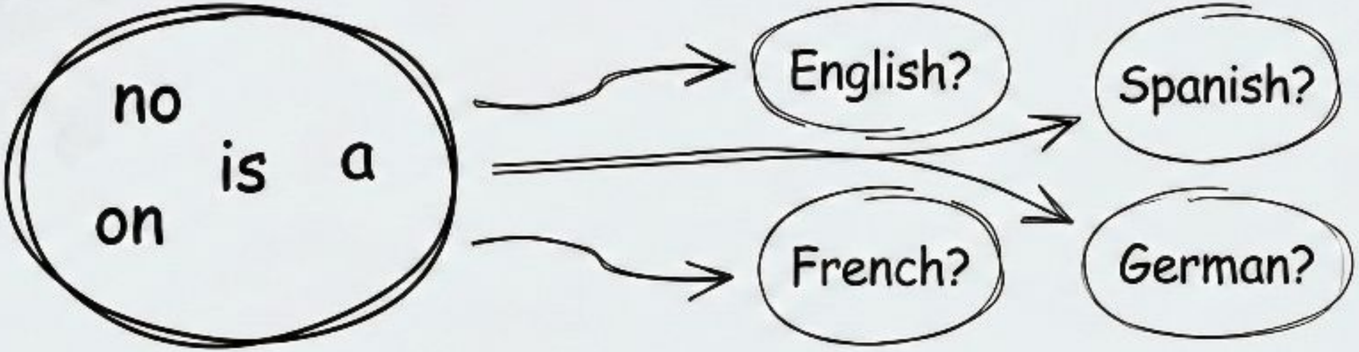
✅ What Worked Well

- High accuracy on long, full sentences.
- Clear distinction between linguistically distant languages (e.g., English vs. Chinese).
- Multilingual BERT's pre-training provided a strong foundation.

❌ Challenges & Confusion

- Short texts (1-2 words) are highly ambiguous.
- Similar languages (e.g., Spanish/Portuguese, Hindi/Urdu) can be misclassified.
- Rare languages with limited training data showed lower performance.

The “Short Word” Challenge: Ambiguity in Action



no

is

a

on

English?

Spanish?

French?

German?

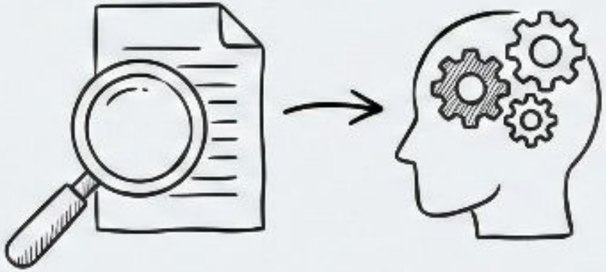
Short, common words lack sufficient context for reliable detection.

The weather is very nice today.

English ✅

Longer sentences provide more linguistic features, leading to correct classification.

Understanding Model Performance through Test Cases



Why Show Examples?
To evaluate accuracy & confidence.

✅ Model Strengths	🤔 Potential Challenges
• High confidence on distinct scripts (e.g., Dravidian languages).	• Lower confidence on closely related languages (e.g., Hindi/Bhojpuri).
• Clear separation of top prediction.	• Ambiguity in short or common English sentences.

Sample Test Case Results (Multilingual Model)

Input Text Snippet	Top Prediction (Confidence)	Other Predictions
आज सुबह जब मैं पार्क...	Hindi (hin) – 0.97	Bhojpuri (0.02), Bengali (0.00)
ఈరోజు మా ఇంట్లో ఒక...	Telugu (tel) – 1.00	Vietnamese (0.00), Chavacano (0.00)
ഇന്നലെ രാത്രി ഞാൻ...	Malayalam (mal) – 1.00	Gilaki (0.00), Western Panjabi (0.00)
இன்று நண்பர்களுடன்...	Tamil (tam) – 1.00	Korean (0.00), Panjabi (0.00)
I woke up early this...	English (eng) – 0.01	Old English (0.01), Pangasinan (0.00)

Real-world examples showing model predictions and confidence scores.

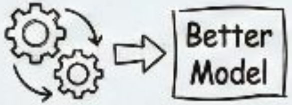
Interpreting the Results: Successes & Key Takeaways

✅ Success Stories: High Confidence

- ✓ Telugu, Malayalam, Tamil (1.00): Perfect detection.
- ✓ Distinct scripts and unique linguistic features lead to near-perfect detection.

🧠 Key Takeaways & Future Improvements

- 🧠 Hindi (0.97): Correct, but shows slight confusion with closely related Bhojpuri.
- 🧠 English (0.01): Problematic; short, common sentences are highly ambiguous.
- 🧠 Future work: Fine-tuning on domain-specific data and handling short texts.



Conclusion: Key Takeaways

- Language detection is essential for global, multilingual AI systems.
- ML models (like DistilBERT) learn deeper linguistic patterns than simple rules.
- Diverse datasets (e.g., WiLi-2018) are critical for building robust and accurate models.
- The project demonstrates a complete NLP pipeline from data to real-world prediction.

Thank You!
Gracias
Merci
धन्यवाद
谢谢
ขอบคุณ
شكراً
?
Questions?