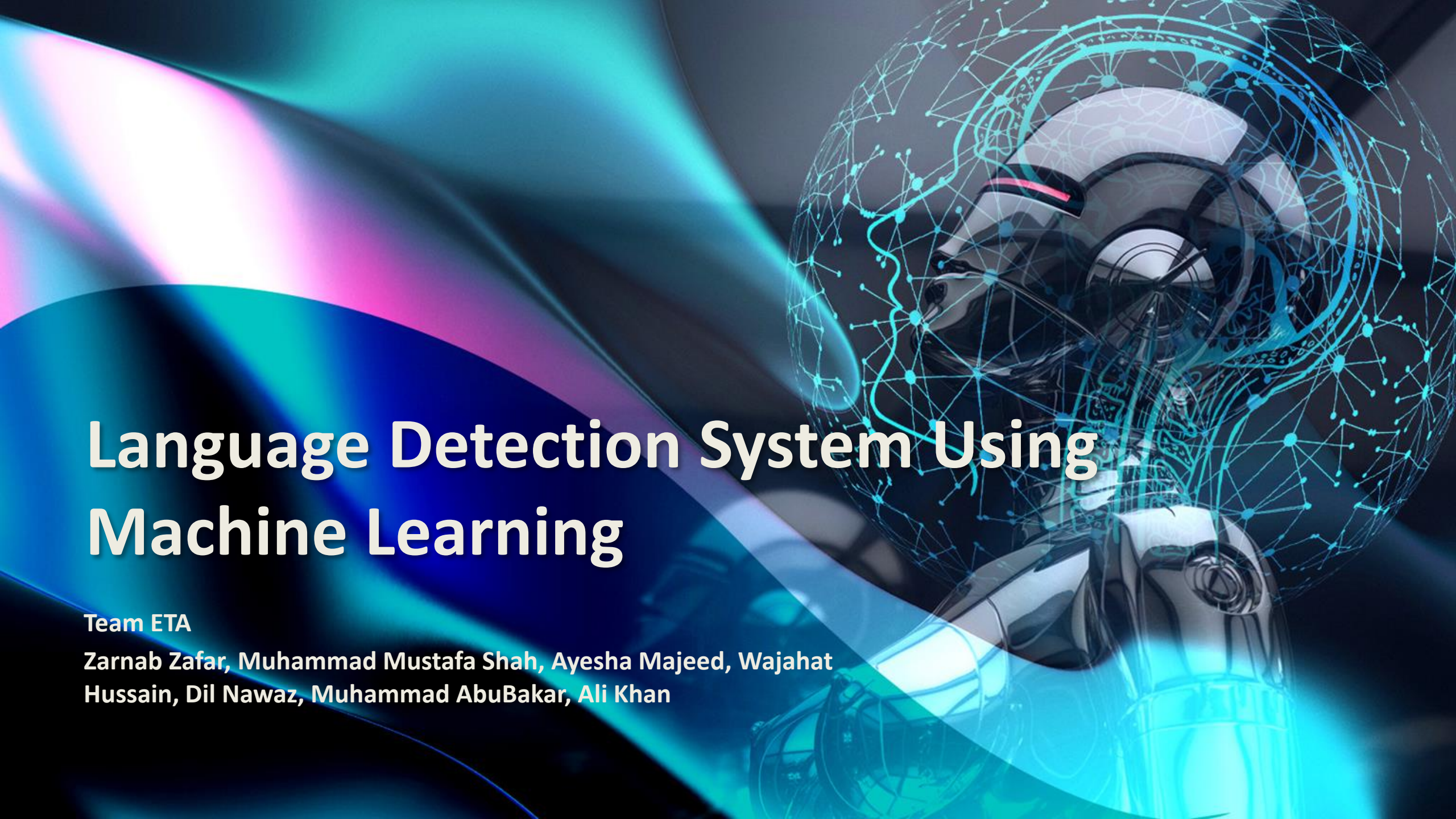


Language Detection System Using Machine Learning

Team ETA

Zarnab Zafar, Muhammad Mustafa Shah, Ayesha Majeed, Wajahat Hussain, Dil Nawaz, Muhammad AbuBakar, Ali Khan

Introduction

- Language detection is a critical task in Natural Language Processing (NLP), enabling applications like:
  - Machine translation
  - Search engines
  - Content analysis
- Traditional methods face limitations with short or mixed texts, making Machine Learning a powerful alternative.

Traditional Approaches

- Rule-based methods are limited in handling:
  - Short texts
  - Mixed-language content
  - Noisy data
- Machine Learning approaches help overcome these challenges by learning from large datasets.

Machine Learning for Language Detection

- Machine Learning models, such as Naive Bayes and Support Vector Machines (SVM), offer flexibility in handling:
  - Complex patterns
  - Short texts
  - Mixed languages in noisy environments.

Problem Statement

- Traditional approaches fail in scenarios involving short, mixed-language content. A robust Machine Learning-based solution is required to ensure high accuracy.

Objectives

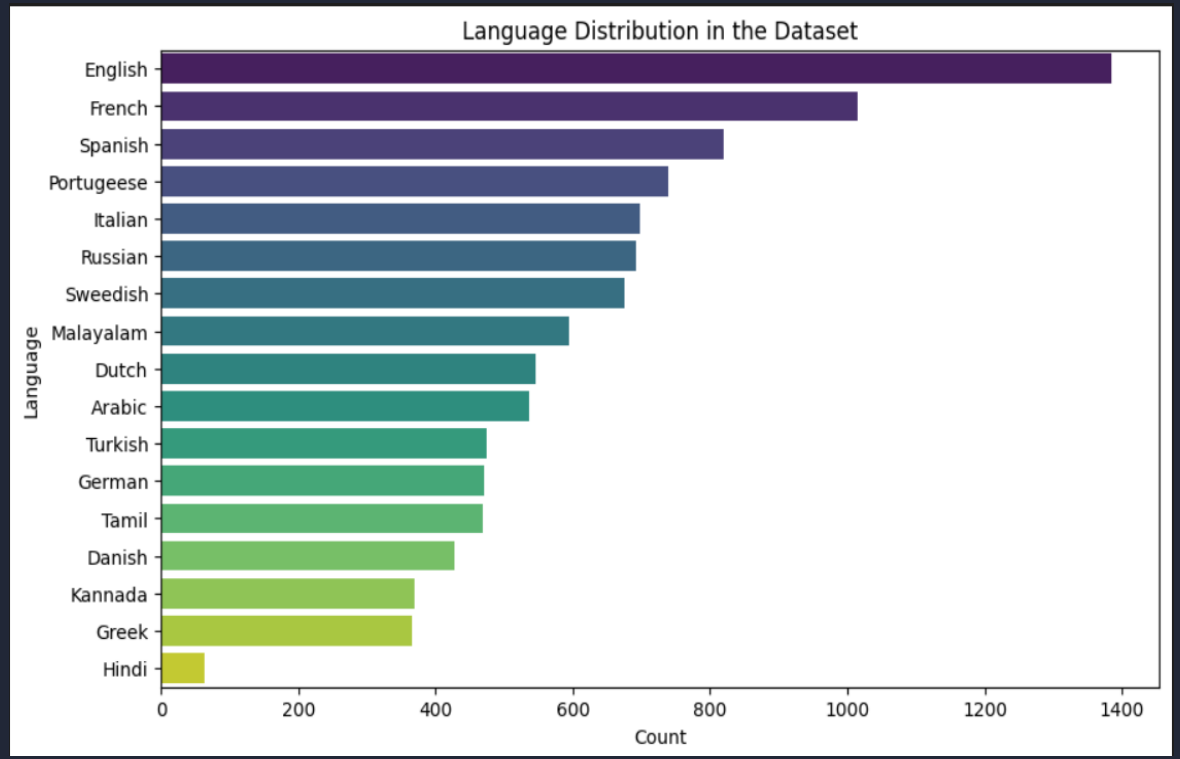
- Develop a machine learning system to detect languages with high accuracy.
- Evaluate Naive Bayes and SVM models.
- Provide recommendations for model optimization.
- Explore real-world applicability.

Dataset Description

- The dataset consists of text samples from 17 languages, including English, French, Spanish, Portuguese, Russian, and Arabic.
- Each text sample is labeled with its respective language, allowing for supervised learning.

Data Analysis

- Exploratory Data Analysis was performed to identify imbalances in the dataset and to understand the distribution of text samples across the 17 languages.



Data Preprocessing

- Key steps in preprocessing included:
  - Stop word Removal: Eliminating common words like 'the', 'and', 'in'.
  - TF-IDF Vectorization: Converting text into numerical features.

Model Selection

- Two machine learning models were selected:
- **Naive Bayes:** A simple, efficient model for text classification.
- **SVM:** Suitable for high-dimensional data like text.

Training and Testing

- The dataset was split into training and testing sets. Both models were trained on the training set and tested to assess performance.

Evaluation Metrics

- Metrics used to evaluate the models included:
  - Accuracy
  - Confusion Matrix
  - Precision, Recall, F1-score

Naive Bayes Model Results

- Naive Bayes achieved 95.6% accuracy, excelling in identifying language patterns and handling noisy data.

Naive Bayes Results:

	precision	recall	f1-score	support
Arabic	1.00	0.93	0.97	106
Danish	1.00	0.92	0.96	73
Dutch	0.99	0.95	0.97	111
English	0.81	1.00	0.89	291
French	0.99	0.97	0.98	219
German	0.99	0.95	0.97	93
Greek	1.00	0.90	0.95	68
Hindi	1.00	0.70	0.82	10
Italian	0.99	0.96	0.97	145
Kannada	1.00	1.00	1.00	66
Malayalam	1.00	0.98	0.99	121
Portugeese	1.00	0.94	0.97	144
Russian	1.00	0.92	0.96	136
Spanish	0.96	0.97	0.97	160
Sweedish	0.96	0.98	0.97	133
Tamil	1.00	0.99	0.99	87
Turkish	1.00	0.89	0.94	105
accuracy			0.96	2068
macro avg	0.98	0.94	0.96	2068
weighted avg	0.96	0.96	0.96	2068

Accuracy: 0.956963249516441

SVM Model Results

- SVM achieved 94.7% accuracy, performing well with short texts but slightly underperforming compared to Naive Bayes.

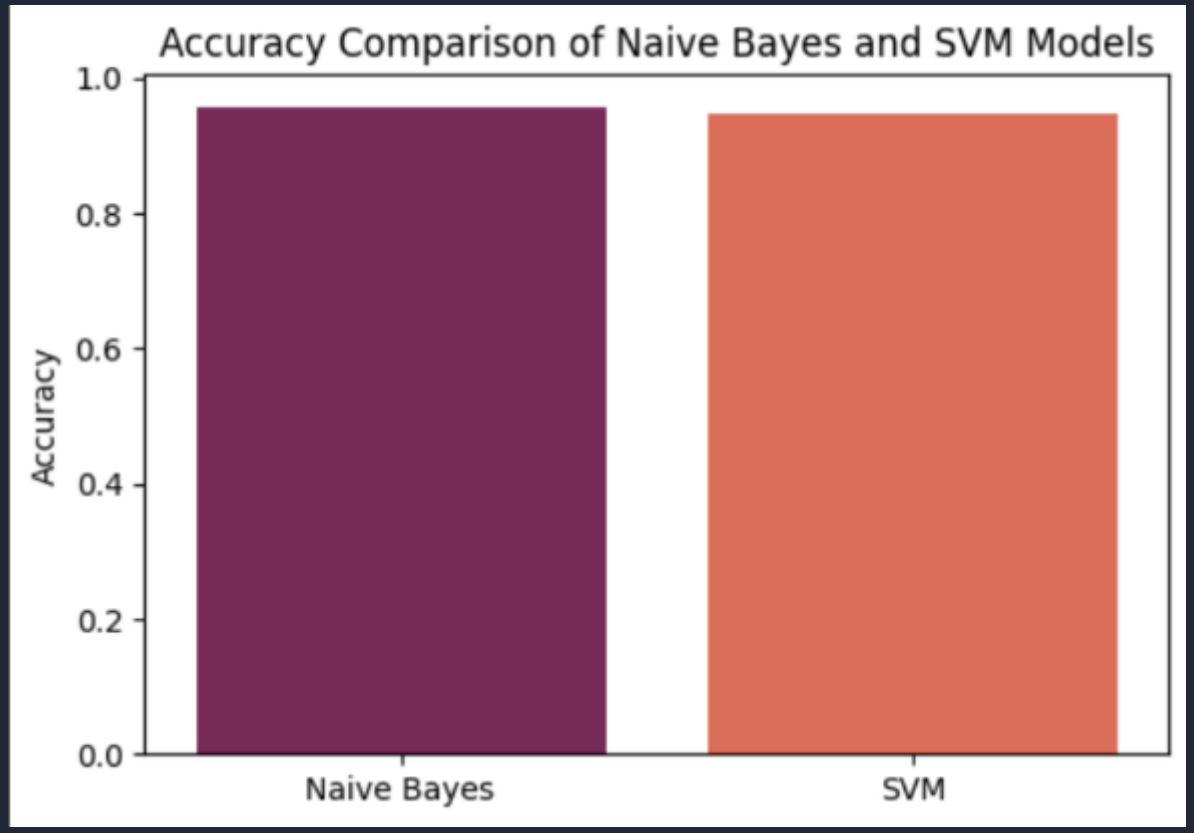
SVM Results:

	precision	recall	f1-score	support
Arabic	1.00	0.93	0.97	106
Danish	0.93	0.90	0.92	73
Dutch	0.99	0.93	0.96	111
English	0.78	1.00	0.87	291
French	1.00	0.97	0.98	219
German	0.99	0.97	0.98	93
Greek	1.00	0.87	0.93	68
Hindi	1.00	0.80	0.89	10
Italian	0.99	0.92	0.95	145
Kannada	1.00	0.98	0.99	66
Malayalam	1.00	0.97	0.98	121
Portugeese	1.00	0.92	0.96	144
Russian	0.98	0.93	0.95	136
Spanish	0.93	0.96	0.95	160
Sweedish	1.00	0.94	0.97	133
Tamil	1.00	0.97	0.98	87
Turkish	0.98	0.90	0.94	105
accuracy			0.95	2068
macro avg	0.97	0.93	0.95	2068
weighted avg	0.96	0.95	0.95	2068

Accuracy: 0.9472920696324951

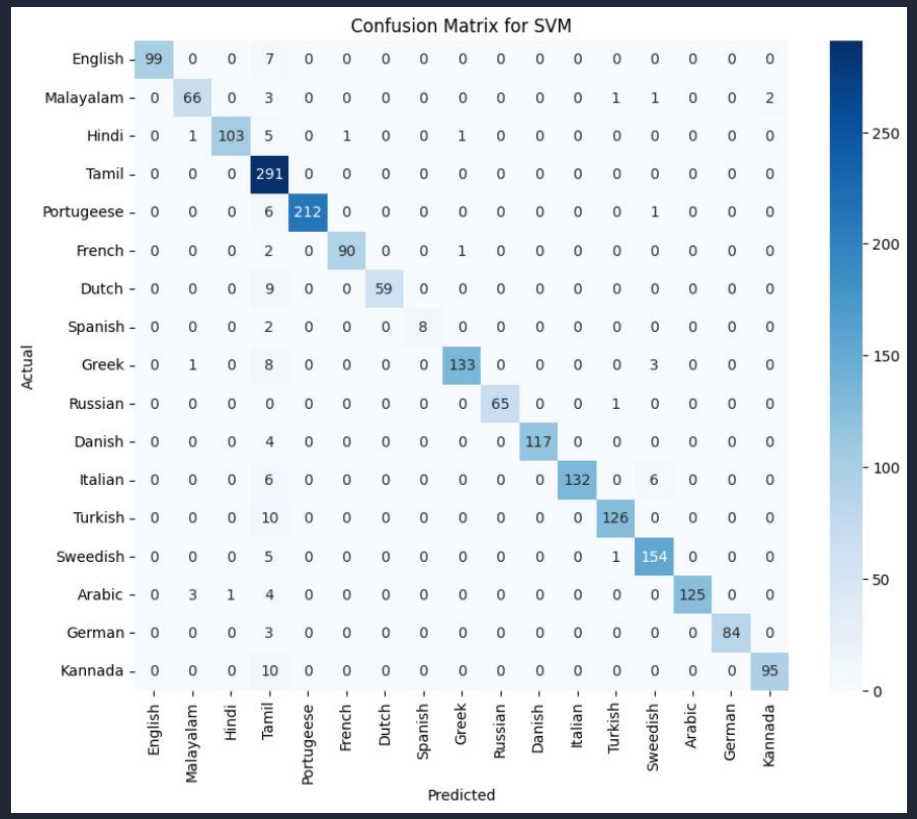
Model Comparison

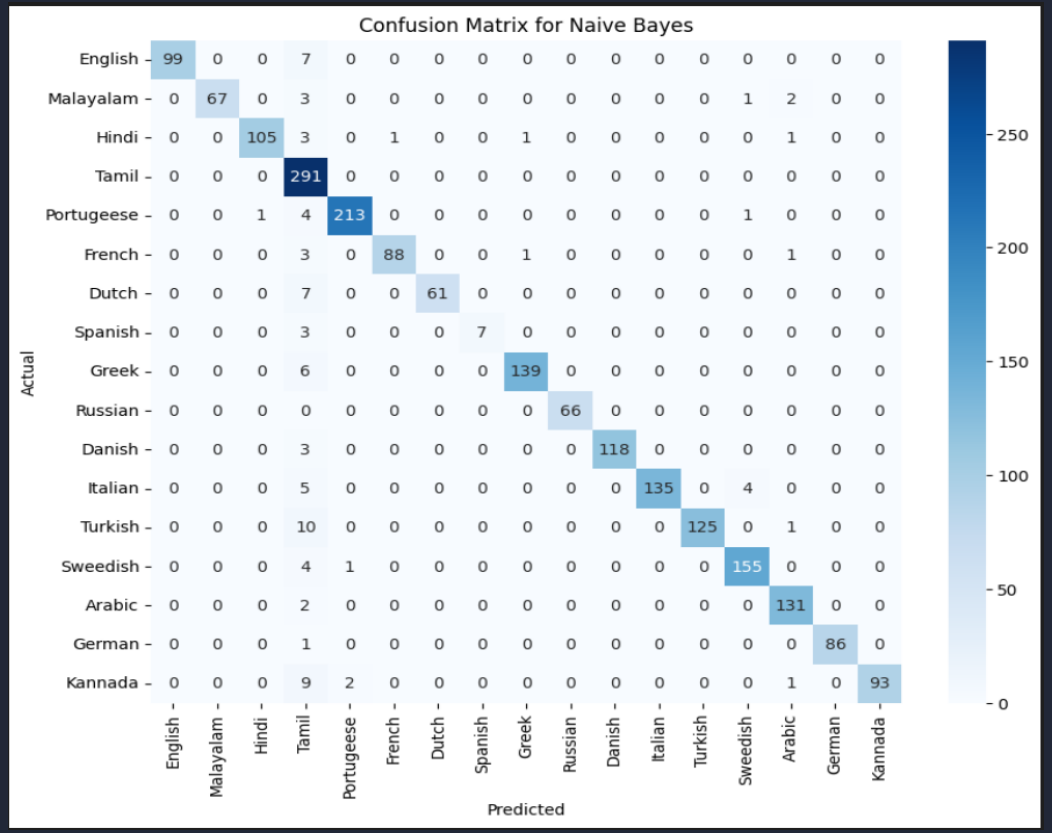
- Comparison between Naive Bayes and SVM showed that Naive Bayes had slightly better accuracy, particularly with noisy language data.



Confusion Matrix

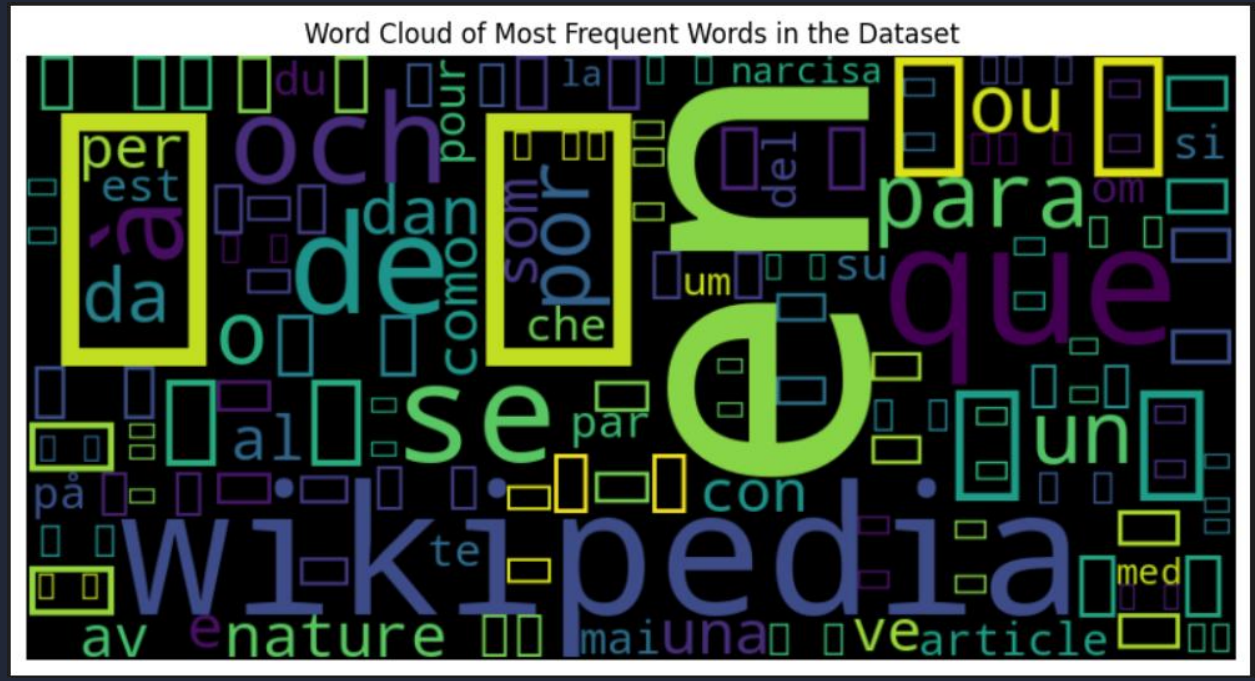
- The confusion matrix for both models provided insight into where each model struggled. Some languages were misclassified more often, particularly those with similar structures.


Confusion Matrix for SVM
Actual
Predicted
English 99 0 0 7 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Malayalam 0 66 0 3 0 0 0 0 0 0 0 0 0 1 1 0 0 2
Hindi 0 1 103 5 0 1 0 0 1 0 0 0 0 0 0 0 0 0
Tamil 0 0 0 291 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Portugeese 0 0 0 6 212 0 0 0 0 0 0 0 0 1 0 0 0
French 0 0 0 2 0 90 0 0 1 0 0 0 0 0 0 0 0 0
Dutch 0 0 0 9 0 0 59 0 0 0 0 0 0 0 0 0 0 0
Spanish 0 0 0 2 0 0 0 8 0 0 0 0 0 0 0 0 0 0
Greek 0 1 0 8 0 0 0 0 133 0 0 0 0 3 0 0 0
Russian 0 0 0 0 0 0 0 0 0 65 0 0 1 0 0 0 0
Danish 0 0 0 4 0 0 0 0 0 0 117 0 0 0 0 0 0
Italian 0 0 0 6 0 0 0 0 0 0 0 132 0 6 0 0 0
Turkish 0 0 0 10 0 0 0 0 0 0 0 0 126 0 0 0 0
Sweedish 0 0 0 5 0 0 0 0 0 0 0 0 1 154 0 0 0
Arabic 0 3 1 4 0 0 0 0 0 0 0 0 0 0 125 0 0
German 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 84 0
Kannada 0 0 0 10 0 0 0 0 0 0 0 0 0 0 0 0 95
English Malayalam Hindi Tamil Portugeese French Dutch Spanish Greek Russian Danish Italian Turkish Sweedish Arabic German Kannada
250 200 150 100 50 0


Confusion Matrix for Naive Bayes
Actual
Predicted
English 99 0 0 7 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Malayalam 0 67 0 3 0 0 0 0 0 0 0 0 0 0 1 2 0 0
Hindi 0 0 105 3 0 1 0 0 1 0 0 0 0 0 1 0 0
Tamil 0 0 0 291 0 0 0 0 0 0 0 0 0 0 0 0 0
Portugeese 0 0 1 4 213 0 0 0 0 0 0 0 0 1 0 0 0
French 0 0 0 3 0 88 0 0 1 0 0 0 0 0 1 0 0
Dutch 0 0 0 7 0 0 61 0 0 0 0 0 0 0 0 0 0
Spanish 0 0 0 3 0 0 0 7 0 0 0 0 0 0 0 0 0
Greek 0 0 0 6 0 0 0 0 139 0 0 0 0 0 0 0 0
Russian 0 0 0 0 0 0 0 0 0 66 0 0 0 0 0 0 0
Danish 0 0 0 3 0 0 0 0 0 0 118 0 0 0 0 0 0
Italian 0 0 0 5 0 0 0 0 0 0 0 135 0 4 0 0 0
Turkish 0 0 0 10 0 0 0 0 0 0 0 0 125 0 1 0 0
Sweedish 0 0 0 4 1 0 0 0 0 0 0 0 0 155 0 0 0
Arabic 0 0 0 2 0 0 0 0 0 0 0 0 0 0 131 0 0
German 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 86 0
Kannada 0 0 0 9 2 0 0 0 0 0 0 0 0 0 1 0 93
English Malayalam Hindi Tamil Portugeese French Dutch Spanish Greek Russian Danish Italian Turkish Sweedish Arabic German Kannada
250 200 150 100 50 0

TF-IDF Word Cloud

- The word cloud generated using TF-IDF scores highlights the most important words across the dataset, revealing which terms contributed most to the model's predictions.



Conclusion

- Both Naive Bayes and SVM models proved effective for language detection, with Naive Bayes slightly outperforming SVM in terms of accuracy and handling complex data.

Future Work

- Future improvements could include:
  - Hyperparameter tuning for both models.
  - Expanding the dataset with more diverse language samples.
  - Integration with real-world translation services.

Real-World Application

- This language detection system can be integrated into applications such as:
  - Automatic translation pipelines.
  - Search engines and content categorization systems.

Q&A

- Questions?