16:954:577:01 STATISTICAL SOFTWARE

NATURAL LANGUAGE PROCESSING

TITLE: LANGUAGE IDENTIFICATION

FINAL PROJECT REPORT

Jayanth Dasamantharao (dj480)

ABSTRACT

This report centers on the task of Language Identification (LI), aiming to categorize text snippets into various languages. It outlines a structured methodology covering data collection, preprocessing, feature extraction, model development, training, and evaluation. The project is categorized into three distinct cases: languages with minimal similarities (e.g., English, Chinese, Hindi, etc.), languages sharing alphabets, linguistic roots and grammatical structures (mainly European), and languages with shared dialects (such as American vs. British English or different versions of Portuguese). With practical applications spanning business analytics, search engines, and education, this study aims to contribute to the field of automated language recognition systems, acknowledging its pivotal role in diverse industries and linguistic landscapes.

TABLE OF CONTENTS

INTRODUCTION

Identifying languages using NLP is crucial in today's global digital landscape. It's fundamental for various NLP applications like translation and sentiment analysis, enabling efficient handling of multilingual data in the internet and social media era.

Initially, the aim was to work on a dataset containing varied languages, achieving good accuracy. Exploring Language Translation and Identification seemed interesting, but complexities arose due to time constraints. Thus, Language Identification focused on similar languages to add complexity, yielding positive outcomes. Later, considering Identifying Languages in a different script faced data availability challenges, leading to a shift towards languages with akin dialects, presenting difficulties due to minute differences. Consequently, this project was categorized into three distinct language cases.

The first case, identifying diverse languages, is crucial for global communication. This aspect is exemplified in research such as Baldwin and Lui's (2010) work on language identification for short and noisy texts, which underscores the challenges and solutions in distinguishing between unrelated languages. This is important for applications like machine translation and content localization where accurate language identification is foundational.

The second case deals with the subtleties of differentiating languages that share the Roman script i.e., European Languages. The work of Zampieri et al. (2014), which explores automatic language identification of similar languages using machine learning, provides valuable insights into the challenges and techniques applicable in this scenario. Identifying these languages accurately is vital for tailored content delivery and effective communication strategies in a multilingual environment.

The third case centers on identifying languages with similar dialects, a nuanced task given the subtle linguistic variations. Research such as the study by Scannell (2007) on language identification for closely related languages offers a foundation for understanding and tackling these challenges. This case has significant implications for dialect-sensitive technologies and sociolinguistic research, contributing to the understanding and preservation of linguistic diversity. The baseline models, Naïve Bayes, and Logistic Regression, served as our foundational models, with Naïve Bayes' simplicity and effectiveness complementing Logistic Regression's capacity for capturing nuanced relationships in language identification.

Additionally, when the exploration was expanded, advanced models like DistilBERT were integrated, aiming to achieve a balance between efficiency, performance, and speed. Transformer-based architecture was leveraged to unravel complex language patterns and capture subtle linguistic nuances.

Overall, these cases offer a comprehensive platform to address language identification challenges. The project fosters learning in both theoretical and practical aspects of NLP model implementation. Essential methodologies like vectorization enhance the ability to develop accurate, efficient models adaptable to diverse linguistic scenarios. In summary, the project aims to create and deploy NLP models for precise language identification while comprehending the complexities inherent in distinguishing languages with diverse traits.

METHODOLOGY

The methodology employs a systematic process to preprocess and analyze language data for precise identification. The steps include:

Data Preprocessing:

- Noise Removal: Eliminating irrelevant special characters and numbers that may not contribute to language identification.
- Lowercasing: Converting all text to lowercase for uniformity, as letter case is often not informative for language identification.
- Tokenization: Dividing the text into individual words or tokens is a fundamental procedure, transforming unprocessed text into a structured format.

Text Vectorization:

- N-Grams: First, N-grams are generated from the text. N-grams are contiguous sequences of 'n' items (words or characters) from the text. For example, in a bigram (2-gram) approach, every two adjacent words or characters are paired together. This process helps in capturing the context and linguistic patterns specific to a language, which might be lost in single word (unigram) analysis.

- TF-IDF: After the N-grams are generated, the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm is applied. TF-IDF transforms the n-grams into a numerical representation, considering not just the frequency of these n-grams in a single document (or text snippet), but also their frequency across all documents in the corpus. This way, common n-grams that might not be very informative (appearing frequently across many languages) are given less weight, while unique or rare n-grams in certain languages are emphasized.

Combining N-grams with TF-IDF provides a more nuanced feature representation of the text. It allows a machine learning model to capture both the local context (through N-grams) and the importance of n-grams in a broader corpus context (through TF-IDF), which can significantly enhance the accuracy of language identification models, especially in distinguishing languages or dialects with similar vocabulary and structure.

Model Selection:

For the task of language identification, two foundational models were selected: the Multinomial Naive Bayes Classifier, which is apt for categorical data such as word and character counts in text, and the Logistic Regression Classifier, recognized for its predictive strength in classification tasks and trained with a high iteration count to ensure model convergence. Both models were trained on text data transformed by the TF-IDF technique, which provides a numerical representation that emphasizes important words while maintaining context. Additionally, the advanced DistilBERT model, a streamlined transformer architecture, was integrated to capture more nuanced linguistic features, utilizing its 'distilbert-base-multilingual-cased' version, optimized with AdamW, and an 80:20 train-test split. To cover a broad spectrum of languages, three datasets representing diverse language families were included in the study, ensuring a comprehensive approach to language identification across varied linguistic attributes.

CASE 1: Languages with no similarities

This case involves vastly different languages with very few or almost no similarities like English, Chinese, Hindi, etc.

Dataset Overview: The dataset from HuggingFace included 20,000 text samples across 20 distinct languages, uniformly distributed. It consisted of two columns: one with the text and the other with language labels corresponding to the text.

labels	text
Arabic	لا يزال هناك ما زال هناك لا يزال هناك الكثير م...
Bulgarian	затова слагаме бял кедър в къщата и го оставим...
German	tolles designe aber die gummis sind zu schnell...
Greek	το aib παραδίδει πληροφορίες σε πραγματικό χρό...
English	this lamp is ok but not what i expected for th...
Spanish	una estrella porque no puedo dejarla sin estre...
French	pour en avoir acheté deux supplémentaires pour...
Hindi	नीचे के कमरे में स्नान करने के लिए नीचे दिए ...
Italian	le azioni mondiali diminuiscono prima della sc...
Japanese	ハイサイクルでばらまくのにちょうど良いと思い購入。 が、他の弾で
Dutch	meer franse soldaten naar centraalafrikaanse r...
Polish	pies skacze z doku i do wody
Portuguese	um homem está a cortar relva
Russian	богатые источники питания могли бы сделать мал...
Swahili	presha ya rufaa hizi kwa zawadi imekuwa kubwa ...
Thai	หิน ดำ สร้าง แห วนรอบ แพ ลด ฟอร์ม ของ แกะสลัก หิน
Turkish	tamam suça karşı hislerin nedir ve bu konuda n...
Urdu	سب جہانوں کی کہانیاں .
Vietnamese	chúng tôi đã bắt được một người chúng tôi đang...
Chinese	cm, cm胸围 kg 因为常健身所以胸围比这个身高体重等级的人稍大点,

Fig 1: Dataset Overview for Case 1

Model Performance:

To assess the effectiveness of the models in identifying languages with minimal similarities, employed key performance metrics, including accuracy, precision, recall, and F1-score. These metrics were crucial in evaluating the models' language identification capabilities.

Classification Report:

	precision	recall	f1-score
Arabic	0.9949	1.0000	0.9975
Bulgarian	0.9799	0.9949	0.9873
Chinese	1.0000	0.9947	0.9974
Dutch	1.0000	0.9704	0.9850
English	0.8958	1.0000	0.9451
French	0.9949	1.0000	0.9974
German	0.9814	1.0000	0.9906
Greek	1.0000	1.0000	1.0000
Hindi	1.0000	1.0000	1.0000
Italian	0.9674	0.9780	0.9727
Japanese	1.0000	1.0000	1.0000
Polish	1.0000	0.9905	0.9952
Portuguese	1.0000	0.9364	0.9672
Russian	0.9948	0.9797	0.9872
Spanish	0.9242	1.0000	0.9606
Swahili	1.0000	0.8957	0.9450
Thai	1.0000	1.0000	1.0000
Turkish	1.0000	0.9840	0.9920
Urdu	1.0000	0.9955	0.9977
Vietnamese	1.0000	1.0000	1.0000
accuracy			0.9858
macro avg	0.9867	0.9860	0.9859
weighted avg	0.9867	0.9858	0.9858

Fig 2: Naive Bayes (Case 1)

Classification Report:

	precision	recall	f1-score
Arabic	0.9949	1.0000	0.9975
Bulgarian	0.9745	0.9745	0.9745
Chinese	0.9948	1.0000	0.9974
Dutch	0.9710	0.9901	0.9805
English	0.9907	0.9907	0.9907
French	1.0000	1.0000	1.0000
German	0.9905	0.9905	0.9905
Greek	1.0000	1.0000	1.0000
Hindi	1.0000	1.0000	1.0000
Italian	0.9677	0.9890	0.9783
Japanese	1.0000	1.0000	1.0000
Polish	0.9953	0.9953	0.9953
Portuguese	0.9718	0.9942	0.9829
Russian	0.9745	0.9695	0.9720
Spanish	1.0000	0.9727	0.9861
Swahili	0.9912	0.9739	0.9825
Thai	1.0000	1.0000	1.0000
Turkish	1.0000	0.9840	0.9920
Urdu	1.0000	0.9955	0.9977
Vietnamese	1.0000	1.0000	1.0000
accuracy			0.9910
macro avg	0.9908	0.9910	0.9909
weighted avg	0.9911	0.9910	0.9910

Fig 3: Logistic Regression (Case 1)

The classification reports show that the Logistic Regression model generally performs better than the Naive Bayes model across all metrics for language identification. It has higher accuracy, as well as higher precision, recall, and F1-scores across the languages. Both models perform well, with high metrics in each category, but the Logistic Regression model has a slight edge in overall performance.

N-gram Analysis:

N-gram Type	Word Level	Character Level
unigrams	0.9085	0.9493
uni+bigrams	0.9135	0.9850
uni+bi+trigrams	**0.9173**	**0.9900**

Table 1: Accuracy for Naive Bayes (Case 1)

The table presents the accuracy improvements of a Naive Bayes classifier when using different N-gram types for language processing. Character-level N-grams consistently outperform word-level N-grams, with the accuracy increasing as more N-grams are combined. This suggests that including more granular details (like character combinations) allows the model to better capture and differentiate between the nuances of various languages, leading to a higher accuracy, especially when uni, bi, and trigrams are used together.

N-gram Type	Word Level	Character Level
unigrams	**0.9012**	0.9385
uni+bigrams	0.8982	0.9780
uni+bi+trigrams	0.8975	**0.9835**

Table 2: Accuracy for Logistic Regression (Case 1)

The table indicates that for a Logistic Regression classifier, accuracy improves when transitioning from word-level to character-level N-grams, with the highest accuracy seen at character-level uni+bi+trigrams. This highlights the effectiveness of character-level analysis in capturing linguistic features for language identification tasks.

DistilBert: To improve upon the already high accuracy achieved by the baseline models in language identification, DistilBERT was employed with the intention of refining the detection of subtle linguistic patterns and enhancing performance. The following key components were configured for this implementation:

- **Optimization Technique:** Utilized the AdamW optimizer, with a fine-tuned learning rate set at 5e-5 to balance speed and accuracy.
- **Model Architecture:** Adopted the 'distilbert-base-multilingual-cased' model, specifically designed to understand multiple languages and sensitive to the case of the input text, which is crucial for language differentiation.
- **Training Epochs:** The model was trained over three epochs to ensure it learned a robust representation of language features without overfitting.
- **Batch Size:** A batch size of 32 was selected, providing a good trade-off between computational efficiency and the model's ability to generalize.
- **Tokenizer:** The DistilBertTokenizer was used to convert language text into a format that DistilBERT can process, preserving the intricate characteristics of each language.

These measures aimed to create a sophisticated model capable of surpassing the baseline models by capturing complex linguistic nuances more effectively.

Model	**Precision**	**Recall**	**Accuracy**
Distil-BERT(multilingual-cased)	0.9962	0.9959	0.9960

Table 3: Results for Distil-BERT (Case 1)

The DistilBERT (multilingual-cased) model shows superior precision, recall, and accuracy compared to the baseline Naive Bayes and Logistic Regression models previously. Its scores are consistently high and close to perfection, indicating that it is exceptionally effective in correctly identifying the language of the given data. This suggests that DistilBERT is more complex architecture and pre-training on a diverse multilingual corpus allows it to capture the nuances of language better than the simpler, traditional machine learning approaches of the baseline models.

The confusion matrix below illustrates the model's strong language identification proficiency, evident in high diagonal values reflecting correct predictions. Despite a few off-diagonal cells indicating occasional confusion between languages, the model maintains overall high accuracy, effectively distinguishing between languages.

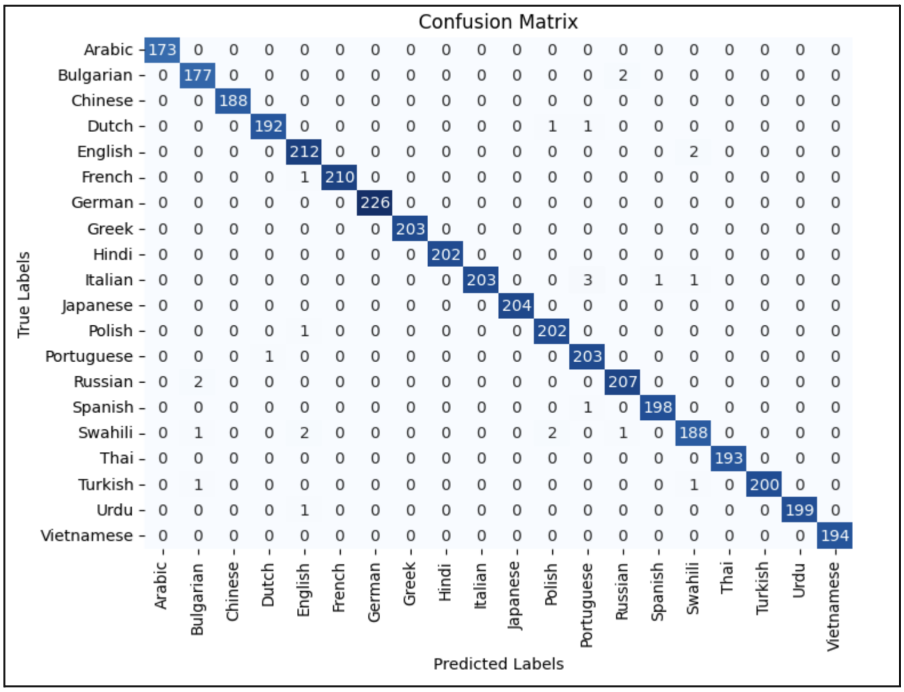

Fig 4: Confusion Matrix for Case 1

CASE 2: Languages with Common Linguistic Roots

This case focuses on European languages like French, Spanish, German, etc. due to their script and linguistic affinities.

Dataset Overview: The dataset for this case was obtained from Tatoeba.com, contains 12,000 text excerpts spanning 12 European languages, including Danish, Dutch, English, French, German, Swedish, Italian, Latin, Portuguese, Spanish, Irish, and Polish. It is structured with two columns indicating Language Labels and Text Samples.

label	text
Danish	Hvordan har du det?
Dutch	Wilt u een kopje koffie?
English	Let's try something.
French	Lorsqu'il a demandé qui avait cassé la fenêtre...
German	Lass uns etwas versuchen!
Swedish	Vi trodde att det var ett flygande tefat.
Italian	Devo andare a dormire.
Latin	Verba volant, scripta manent.
Portuguese	Uma menina chorando abriu a porta.
Spanish	¡Intentemos algo!
Irish	Cá bhfuil críochfort na mbus?
Polish	Piękne są chmury płynące po niebie.

Fig 5: Dataset Overview for Case 2

Model Performance:

Classification Report:

	precision	recall	f1-score
Danish	0.9709	0.9950	0.9828
Dutch	1.0000	0.9801	0.9899
English	0.9764	0.9857	0.9810
French	0.9862	0.9954	0.9908
German	0.9852	1.0000	0.9926
Irish	1.0000	0.9910	0.9955
Italian	0.9734	0.9839	0.9786
Latin	1.0000	0.9333	0.9655
Polish	1.0000	0.9947	0.9973
Portuguese	0.9254	0.9947	0.9588
Spanish	0.9677	0.9730	0.9704
Swedish	1.0000	0.9567	0.9779
accuracy			0.9821
macro avg	0.9821	0.9820	0.9818
weighted avg	0.9827	0.9821	0.9821

Fig6:Naive Bayes (Case 2)

Classification Report:

	precision	recall	f1-score
Danish	0.9850	0.9801	0.9825
Dutch	0.9950	0.9900	0.9925
English	0.9951	0.9762	0.9856
French	0.9906	0.9769	0.9837
German	1.0000	0.9950	0.9975
Irish	0.9955	0.9865	0.9910
Italian	0.9482	0.9839	0.9657
Latin	0.9461	0.9897	0.9674
Polish	0.9895	0.9947	0.9921
Portuguese	0.9784	0.9679	0.9731
Spanish	0.9838	0.9838	0.9838
Swedish	0.9902	0.9760	0.9831
accuracy			0.9833
macro avg	0.9831	0.9834	0.9832
weighted avg	0.9836	0.9833	0.9834

Fig 7: Logistic Regression (Case 2)

While both models are effective for language identification, the Logistic Regression model appears to have an edge in terms of consistent performance across different languages.

N-gram Analysis:

N-gram Type	Word Level	Character Level
unigrams	0.9679	0.7887
uni+bigrams	**0.9688**	0.9554
uni+bi+trigrams	0.9679	**0.9804**

Table 4: Accuracy for Naive Bayes (Case 2)

N-gram Type	Word Level	Character Level
unigrams	0.9537	0.8271
uni+bigrams	**0.9554**	0.9617
uni+bi+trigrams	0.9537	**0.9792**

Table 5: Accuracy for Logistic Regression (Case 2)

The accuracy tables for both models reveal that word-level n-grams consistently demonstrate high accuracy, indicating that word choice is a significant indicator of language in this dataset. Regarding the efficiency at the character level, it's evident that character-level n-grams, particularly bi-grams and trigrams, play a crucial role in language identification. This efficiency is likely attributed to their ability to capture language-specific morphological patterns effectively. Furthermore, this underscores the significance of letter combinations and sequences in the process of identifying languages.

DistilBERT: Similar as in Case 1, to enhance language identification accuracy beyond the capabilities of baseline models, the integration of DistilBERT into the workflow was pursued. The hyperparameters mirror those employed in the previous case.

Model	Precision	Recall	Accuracy
Distil-BERT(multilingual-cased)	0.9866	0.9867	0.9867

Table 6: Results for Distil-BERT (Case 2)

The confusion matrix suggests that the model is highly accurate in identifying languages, with the majority of predictions correctly aligned along the diagonal. However, it also highlights some confusion between closely related languages, particularly between Latin and Italian, where several instances of Latin are mistakenly identified as Italian. Despite these occasional misclassifications, the overall performance of the model is notably proficient.

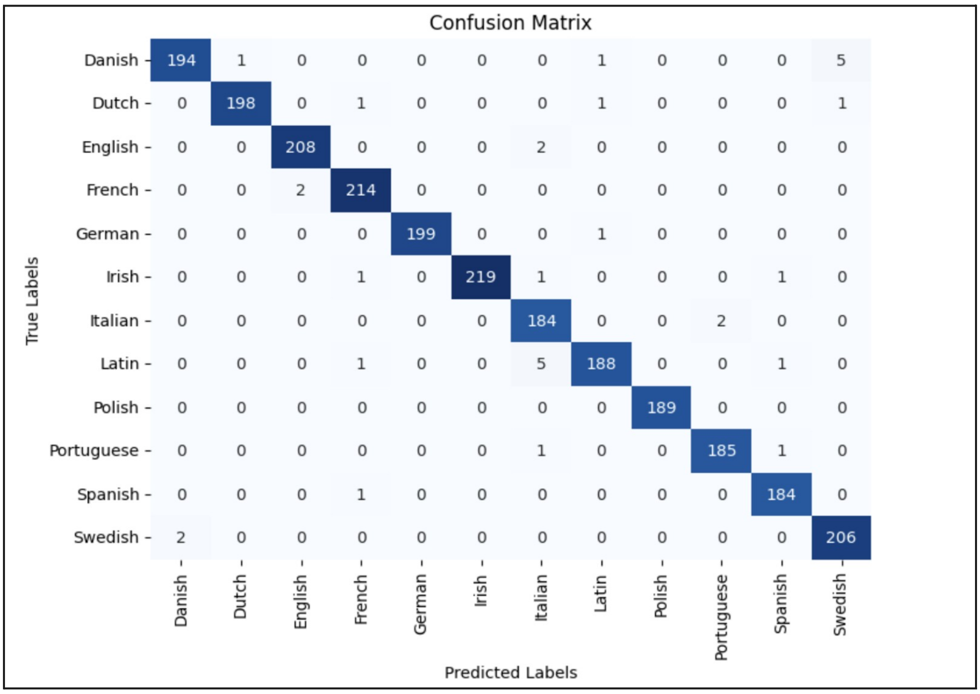

Fig 8: Confusion Matrix for Case 2

CASE 3: Languages with shared dialects

This case explores different language families such as British and American English, Brazilian and European Portuguese, etc., aiming to discern differences in pronunciation, vocabulary, idiomatic expressions, and occasional grammar nuances.

Dataset Overview: Utilized DSL Corpus dataset with 100,000 text samples for 10 Languages. These languages encompass 4 distinct dialects: English, Portuguese, Spanish and Slavic. This is the repository for the DSL Corpus Collection (DSLCC). The DSLCC is a multilingual collection of short excerpts of journalistic texts. It has been used as the main data set for the DSL shared tasks organized within the scope of the workshop on NLP for Similar languages, Varieties and Dialects.

label	text
American English	[analysts equally say that the feat of keeping...
Argentine Spanish	[la bolsa no se va desentender de los corredor...
Bosnian	[angie je predana novom hobiju i smatra to pri...
Brazilian Portuguese	[se você precisa divulgar produtos serviços ou...
British English	[feeling that they were ready they headed into...
Castilian Spanish	[igualmente informó sobre nuevas nominaciones ...
Croatian	[meblove kuhinje astra i sphera odlikuju se vr...
European Portuguese	[sinceramente acho que isto é uma espécie de s...
Peruvian Spanish	[en efecto cuando la economía internacional co...
Serbian	[u generalnom plasmanu vodi pedrosa sa poena m...

Fig 9: Dataset Overview for Case 3

Model Performance:

Classification Report:

	precision	recall	f1-score
American English	0.4975	0.6057	0.5463
Argentine Spanish	0.7119	0.8639	0.7805
Bosnian	0.7375	0.6287	0.6787
Brazilian Portuguese	0.8677	0.9078	0.8873
British English	0.5292	0.4227	0.4700
Castilian Spanish	0.6388	0.8962	0.7459
Croatian	0.8062	0.8318	0.8188
European Portuguese	0.9074	0.8630	0.8846
Peruvian Spanish	0.9220	0.3412	0.4981
Serbian	0.7824	0.8706	0.8241
accuracy			0.7246
macro avg	0.7400	0.7232	0.7134
weighted avg	0.7406	0.7246	0.7149

Fig 10: Naive Bayes (Case 3)

Classification Report:

	precision	recall	f1-score
American English	0.4934	0.4990	0.4962
Argentine Spanish	0.8407	0.7806	0.8095
Bosnian	0.7269	0.6958	0.7110
Brazilian Portuguese	0.8410	0.8773	0.8588
British English	0.5201	0.5151	0.5176
Castilian Spanish	0.8225	0.7955	0.8088
Croatian	0.8169	0.7826	0.7994
European Portuguese	0.8818	0.8285	0.8543
Peruvian Spanish	0.7861	0.8866	0.8333
Serbian	0.8181	0.8844	0.8500
accuracy			0.7548
macro avg	0.7548	0.7545	0.7539
weighted avg	0.7553	0.7548	0.7543

Fig 11: Logistic Regression (Case 3)

The performance of Naive Bayes and Logistic Regression models in dialect identification reveals that Logistic Regression consistently outshines Naive Bayes, offering higher precision, recall, and F1-scores. Despite some success with Naive Bayes, its performance is less stable across dialects. Logistic Regression not only provides greater accuracy but also maintains a more reliable performance in distinguishing between similar dialects.

N-gram Analysis:

N-gram Type	Word Level	Character Level
unigrams	**0.7311**	0.5009
uni+bigrams	0.7299	0.6362
uni+bi+trigrams	0.7249	**0.7059**

Table 7: Accuracy for Naive Bayes (Case 3)

N-gram Type	Word Level	Character Level
unigrams	**0.7393**	0.5460
uni+bigrams	0.7291	0.6669
uni+bi+trigrams	0.7071	**0.7641**

Table 8: Accuracy for Logistic Regression(Case 3)

Here, the accuracy word levels for both Naive Bayes and Logistic Regression are consistent for uni,bi and tri grams. But, the accuracy for character level is relatively higher in tri-grams for both Naive Bayes and Logistic Regression because character-level tri-grams might generalize well across different languages and dialects, capturing common patterns that extend beyond individual words.

DistilBert: For this case, DistilBERT was fine-tuned using the AdamW optimizer, trained for three epochs with a batch size of 8, and utilized gradient accumulation and a warm-up in the learning rate scheduler to effectively adapt to the nuances of language identification tasks.

Model	Precision	Recall	Accuracy
Distil-BERT(multilingual-cased)	0.7442	0.7343	0.7607

Table 9: Results for Distil-BERT (Case 3)

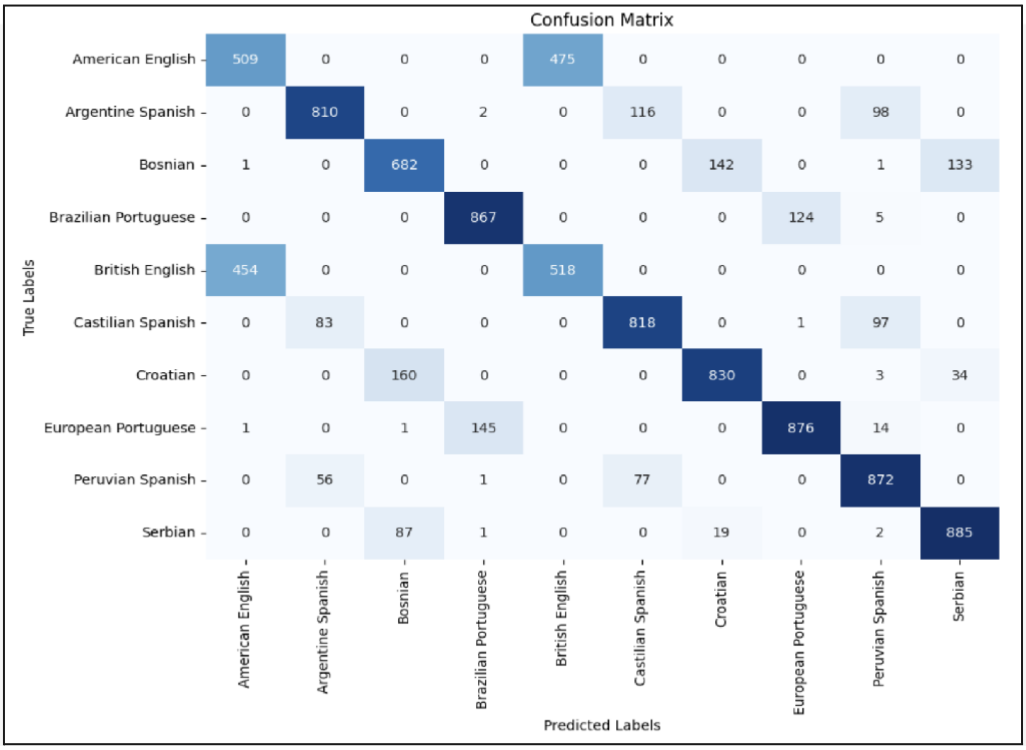

Fig 12: Confusion Matrix for Case 3

The confusion matrix reveals some misclassifications within language families, which is expected. For instance, the model may struggle to distinguish between British English and American English due to the high similarity in vocabulary and linguistic structures between these closely related language variants.

CONCLUSION

The performance of the baseline models, Naive Bayes and Logistic Regression, was critically evaluated. Both models proved to be effective tools for language identification. Logistic Regression, in particular, showed a consistent edge over Naive Bayes in terms of precision, recall, and accuracy, demonstrating its capability in handling language data with a higher degree of sophistication.

The study also highlighted the importance of N-gram analysis, especially the preference for character-level N-grams over word-level N-grams. It was observed that the accuracy improved significantly with the inclusion of bi- and tri-grams. This finding suggests that focusing on smaller units of language, like characters, is crucial for accurately differentiating between languages, especially those with subtle linguistic differences.

The integration of DistilBERT into the project represented a significant step forward in language identification accuracy. The performance of DistilBERT, in terms of precision, recall, and accuracy, surpassed that of the baseline models. This underscores the effectiveness of transformer-based architectures in capturing the intricate nuances of languages, offering a more refined approach to language processing.

The project was structured into three distinct cases, each with its unique challenges. The first case involved languages with minimal similarities, showcasing the fundamental effectiveness of the models. The second case focused on languages that use similar alphabets and common linguistic roots highlighting the complexities involved in differentiating languages with shared scripts and linguistic roots. The third case, arguably the most challenging, dealt with languages that have shared dialects, emphasizing the intricate task of discerning subtle differences between closely related dialects.

Overall, the project not only showcased the efficacy of various machine learning and NLP techniques in language identification but also illuminated the diverse challenges associated with different language sets. The addition of advanced models like DistilBERT to the study emphasized the dynamic nature of language processing and the potential for increasingly sophisticated language identification methods in the future

FUTURE SCOPE

- Leverage Extensive Corpus Data: Utilize larger and varied textual datasets for enhanced model performance and broader language coverage.
- Explore Diverse Transformer Architectures: Experiment with alternative transformer models like GPT-3, T5, or RoBERTa to leverage unique strengths for specific tasks.
- Enhance Transliteration Techniques: Focus on improving transliteration capabilities, especially in multilingual contexts, for applications like cross-lingual search engines.

- Expand Dataset Diversity: Include texts from various domains, languages, and dialects to improve the model's adaptability and accuracy.
- Incorporate Real-World Noisy Data: Embrace the challenge of handling real-world noisy text, contributing to the development of more robust and practical models.

REFERENCES:

1. https://www.yourdatateacher.com/2021/04/30/an-efficient-language-detection-model-using-naive-bayes/

2. https://www.analyticsvidhya.com/blog/2021/03/language-detection-using-natural-language-processing/

3. Baldwin, T., & Lui, M. (2010). Language Identification: The Long and the Short of the Matter. Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

4. Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J., & Nakov, P. (2014). Overview of the 6th Workshop on Building and Using Comparable Corpora. Proceedings of the Workshop on Building and Using Comparable Corpora.

5. Scannell, K. P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. Building and Exploring Web Corpora.

6. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics.