MICHIGAN STATE UNIVERSITY

Language detection

Yashashvini Rachamallu, Bhanu Kanamarlapudi, Neel Joshi

Introduction

- A method that classifies text into a set of accessible languages.
- Plays a critical role in numerous NLP applications, such as autocorrection, machine translation, information retrieval, summarization, and question answering.
- Two approaches to language detection : computational and non-computational.
- Plays crucial role in natural language processing for sentiment analysis
- Useful in the development of chatbots and virtual assistants that can interact with users in multiple languages
- Helps linguists and historians identify the language of ancient manuscripts and documents

Source : Google

Problem Statement

MICHIGAN STATE UNIVERSITY

Implementing different classification models to predict the languages

Motivation

MICHIGAN STATE UNIVERSITY

- Is an important tool for facilitating communication and ensuring that content is appropriate and accessible to diverse audiences.

- Plays a crucial role in analyzing language usage and trends to provide insights into customer behavior, sentiment analysis.

- Companies operating globally use language detection to provide users with localized content and services.

- Can also be used for security purposes.

Scope

- Implementing different embedding methods
- Implementing numerous Machine learning models
- Implementing deep learning models
- Fine tuning mBERT
- Analyze and Compare

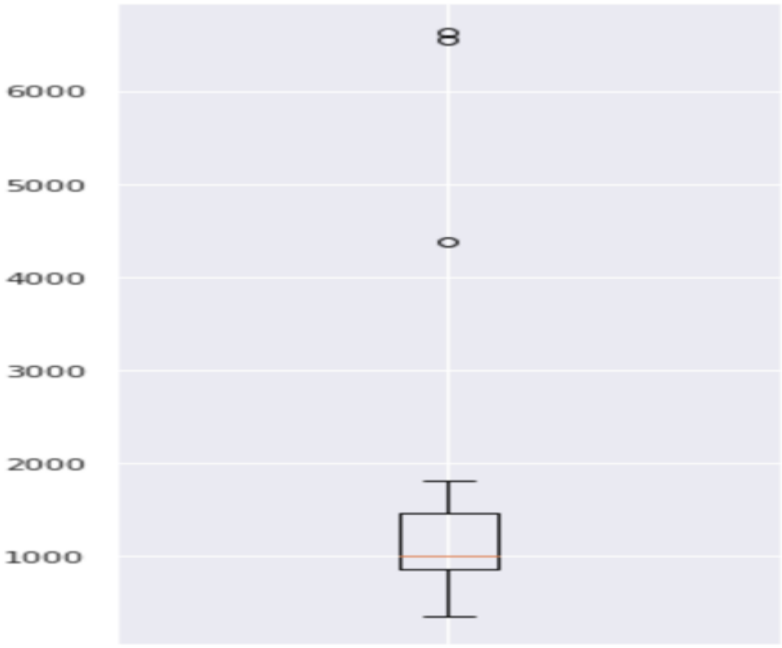
Datasets

MICHIGAN STATE UNIVERSITY

- Data-1(10267 records): https://www.kaggle.com/datasets/basilb2s/language-detection
- Data-2(21859 records): https://www.kaggle.com/code/martinkk5575/language-detection/data
- Data-3(12646 records): https://www.kaggle.com/datasets/lailaboullous/language-detection-dataset
- Data-4(10^7 records): https://www.kaggle.com/datasets/chazzer/big-language-detection-dataset?select=sentences.csv

Dataset Creation

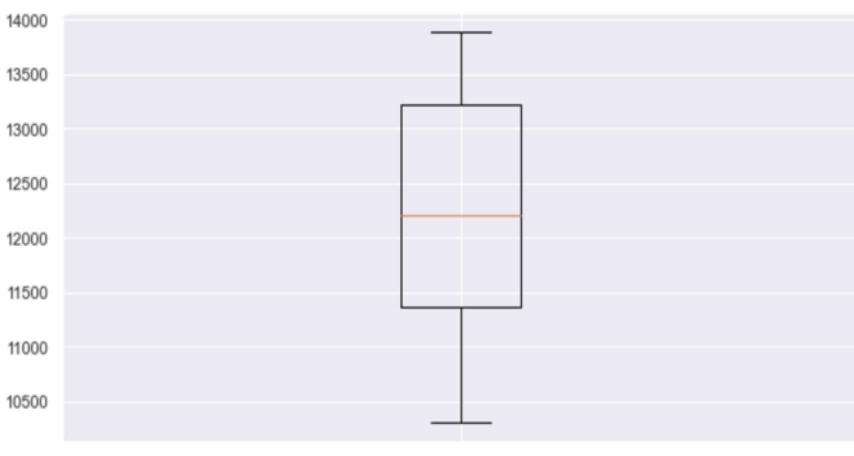
- We created two different datasets by merging the above four datasets.
- For Dataset-1, we started with combining Data-1, Data-2 and Data-3 resulting in approximately 45000 rows.
- To remove the outliers in above combined Dataset-1, we used the Data-4 and picked only languages whose sentences count is in between 4000-6000.
- We ended up generating Dataset-1 with around 2 lakh sentences.

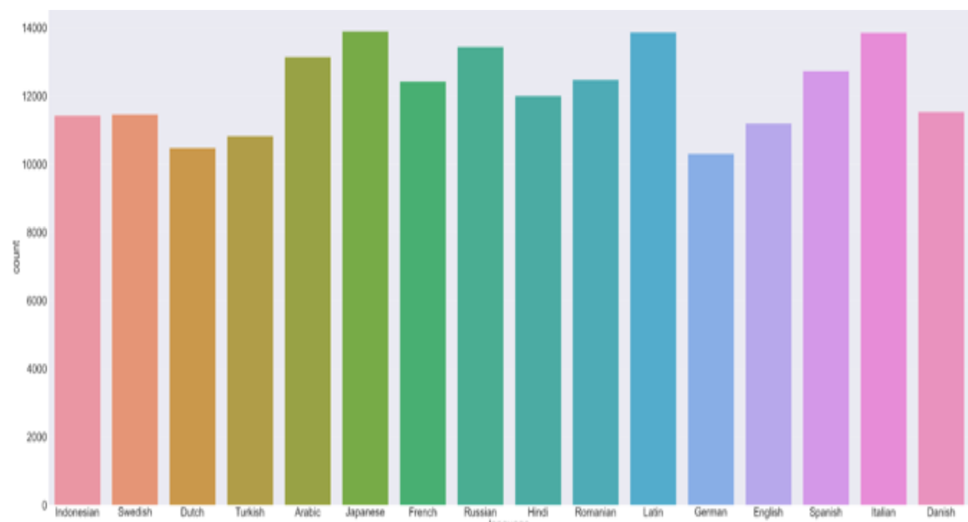


Dataset Creation

- The Dataset-2 is generated by combining Data-1, Data-2, Data-3 and few sentences from Data-4, to make each language in Dataset-2 to be between 10000 and 14000.

- Dataset-2 contains around 2 lakh rows.

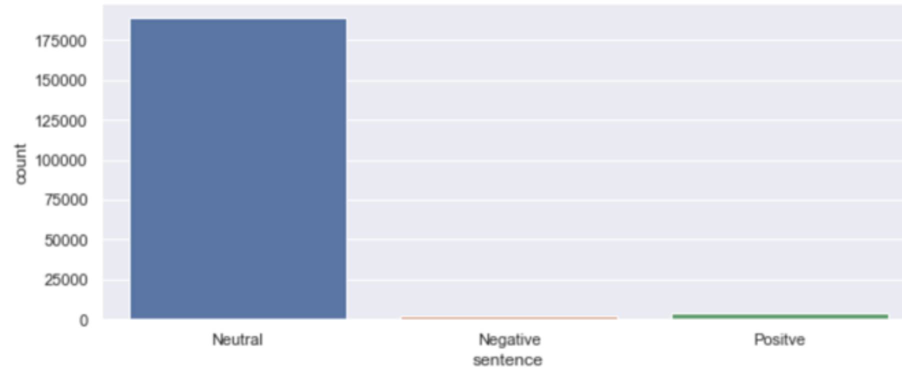


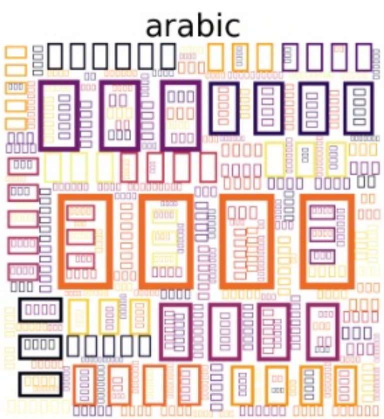


Dataset Analysis

MICHIGAN STATE UNIVERSITY

- Checked the sentiment of all the sentences, to make sure if we are using neutral sentences.
- Plotted the word cloud for different languages.





Methodology

MICHIGAN STATE UNIVERSITY

- **Text Preprocessing:**
  - Converted to lower case
  - Removed special characters
  - Removed of punctuation, htmls, email addresses.
  - Applied stemming
  - Removed stop words for each language separately.

- **Text vectorization:**
  - Bag-of-Words
  - TF-IDF vectorizer
  - transformer, Tokenizer
  - Word2Vec
  - N-gram analysis
  - DistilBERT-base-uncased
  - mBERT

Methodology

MICHIGAN STATE UNIVERSITY

- **Classification modeling:**
  - Naïve Bayes
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - Ensemble
  - SVM
  - LSTM
  - BiLSTM
  - DISTILBERT

- **Training and Testing:**
  - Training – 80%
  - Testing – 20%

Results

MICHIGAN STATE UNIVERSITY

MODEL	ACCURACY
BoW + Naïve Bayes	87.1
BoW + Decision Trees	78.5
TF-IDF + Naïve Bayes	85
TF-IDF + Decision Trees	76
mBERT + Custom NN	72
LSTM + 1 Dense	87
LSTM + 2 Dense	86.7
Bi-LSTM + 2 Dense	85.6

Table 1: Dataset-1 Results

Results

MICHIGAN STATE UNIVERSITY

Model	Accuracy
Unigram(Word) + NB	82
Unigram(Word) + LR	88
Unigram(Word) + SVM	88
Unigram(Word) + RF	87
Bigram (Word) + NB	33
Bigram (Word) + LR	35
Bigram (Word) + SVM	35
Bigram (Word) + RF	35
Trigram (Word) + NB	13
Trigram (Word) + LR	15
Trigram (Word) + SVM	15
Trigram (Word) + RF	15
3 Char gram + NB	84
3 Char gram + LR	90.3
3 Char gram + SVM	91.5
3 Char gram + RF	89
4 Char gram + NB	74
4 Char gram + LR	82
4 Char gram + SVM	83
4 Char gram + RF	80
Word2Vec + LR	68.3
Char+Word+Pos+ NB	91.7
Char+Word+Pos + LR	92.3
Char+Word+Pos + RF	91
Char+Word+Pos + Ensemble	92.6
Fine-tuned DistilBERT	88.7
Fine-tuned mBERT	**93**

Table 2: Dataset-2 Results

Conclusion

MICHIGAN STATE UNIVERSITY

- We have experimented with various embedding techniques, machine learning and deep learning models.
- Observed that considering position along with character and word analysis plays a crucial role during embedding phase.
- Among all our implementations, the model with mBERT has given the best accuracy of **93%.**

Future Work

MICHIGAN STATE UNIVERSITY

- Improve the performance of low-resource language identifications.

- Develop a user interface to accept text or document as input and identify and parse the text to desired language.

- Explore domain specific language detection as language usage vary significantly depending on domain or industry.

References

1. N. Sarma, S. R. Singh and D. Goswami, "Word Level Language Identification in Assamese-Bengali-Hindi-English Code-Mixed Social Media Text," 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesia, 2018, pp. 261-266, doi: 10.1109/IALP.2018.8629104.
2. W. B. Canvar and J. M. Trenkle. N-gram based Text Categorization. Proceedings of Symposium on Document Analysis and Information Retrieval, University of Nevada, Las Vegas, pp. 161-176, 1994.
3. B. Ahmed, S.-H. Cha, and C. Tappert. Language Identification from Text Using N-gram Based Cumulative Frequency Addition. Proceedings of Student/Faculty Research Day, CSIS, Pace University, 2004.
4. Pujeri*, B. P., & Sai D, J. (2020). An anatomization of language detection and translation using NLP techniques. *International Journal of Innovative Technology and Exploring Engineering*, *10*(2), 69–77. https://doi.org/10.35940/ijitee.b8265.1210220.
5. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4.
6. Christian Bartz, Tom Herold, Haojin Yang, Christoph Meinel(2017). Language Identification Using Deep Convolution Recurrent Neural Networks. https://arxiv.org/abs/1708.04811.
7. Priyanka Mathur, Arkajyoti Misra, Emrah Budur(2015). Language Identification from Text Document. https://cs229.stanford.edu/proj2015/324_report.pdf
8. Bhat, Irshad Ahmad, et al. "Universal Dependency parsing for Hindi-English code-switching." *arXiv preprint arXiv:1804.05868* (2018).
9. D. Patel and R. Parikh, "Language Identification and Translation of English and Gujarati code-mixed data," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-4, doi: 10.1109/ic-ETITE47903.2020.410.
10. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text Classification Algorithms: A Survey. *Information* **2019**, *10*, 150. https://doi.org/10.3390/info10040150
11. Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

THANK YOU!