As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 7(12): 169-172, 2021

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0712031
Available online at: http://www.ijmtst.com/vol7issue12.html

Automatic Language Detection using Natural Language Processing

N Srilatha | J J C Prasad Yadav | Padagala Apurva

Independent Researcher

To Cite this Article

N Srilatha, J J C Prasad Yadav and Padagala Apurva. Automatic Language Detection using Natural Language Processing. *International Journal for Modern Trends in Science and Technology* 2021, *7* pp. 169-172. https://doi.org/10.46501/IJMTST0712031

Article Info

Received: 01 October 2021; Accepted: 07 December 2021; Published: 10 December 2021

ABSTRACT

Language identification or detection (LD) is the task of automatically detecting the language present in a document based on the content of the document. In this work, we demonstrate the effectiveness of our method over real-world documents set (each document is mostly monolingual) collected from the web. The proposed system consists of several stages: (1) Pre-processing, (2) Text representation using BoWmodel, and (3) Classification.The experimental findings show that the achieved accuracy is 95.72.

Key words:language detection, Multilingual, Text classification and Natural language processing.

INTRODUCTION

Language detection or identification (LD) [1] is the task of determining the natural language that a document or part thereof is written in. Recognizing text in a specific language comes naturally to a human reader familiar with the language. Table 1 presents excerpts from Wikipedia articles in different languages on the topic of Natural Language Processing ("NLP"), labelled according to the language they are written in. Without referring to the labels, readers of this article will certainly have recognized at least one language in Table 1, and many are likely to be able to identify all the languages therein.

English	Natural language processing is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages.
Italian	L'Elaborazione del linguaggio naturale è il processo di trattamento automatico mediante un calcolatore elettronico delle informazioni scritte o parlate nel linguaggio umano o naturale.
Chinese	自然語言處理是人工智慧和語言學領域的分支學科。
Japanese	自然言語処理は、人間が日常的に使っている自然言語をコンピュータに処理させる一連の技術であり、人工知能と言語学の一分野である。

Table 1: Excerpts from Wikipedia articles on NLP in different languages.

Research into Language detection aims to mimic this human ability to recognize specific languages. Over the years, several computational approaches have been developed that, usingspecially designedalgorithms, and indexing structures, are able to infer the language being used without the need for human intervention. The capability of such systems could be described as super-human: an average person may be able to identify a handful of languages, and a trained linguist or translator may be familiar with many dozens, but most

of us will have, at some point, encountered written texts in languages they cannot place. However, LD research aims to develop systems that can identify any human language, a set which numbers in the thousands [2].

In a broad sense, LD applies to any modality of language, including speech, sign language, and handwritten text, and is relevant for all means of information storage that involve language, digital or otherwise. However, in this survey we limit the scope of our discussion to LD of written text stored in a digitally encoded form.

Research to date on LD has traditionally focused on monolingual documents [3]. In monolingual LD, the task is to assign each document a unique language label. Some work has reported near perfect accuracy for LD of large documents in a small number of languages, prompting some researchers to label it a "solved task" [4]. However, to attain such accuracy, simplifying assumptions must be made, such as the monolinguality of each document, as well as assumptions about the type and quantity of data, and the number of languages considered. The ability to accurately detect the language that a document is written in is an enabling technology that increases accessibility of data and has a wide variety of applications. For example, presenting information in a user's native language has been found to be a critical factor in attracting website visitors [5]. Text processing techniques developed in natural language processing and Information Retrieval ("IR") generally presuppose that the language of the input text is known, and many techniques assume that all documents are in the same language. To apply text processing techniques to real-world data, automatic LD is used to ensure that only documents in relevant languages are subjected to further processing. In information storage and retrieval, it is common to index documents in a multilingual collection by the language that they are written in, and LD is necessary for document collections where the languages of documents are not known a-priori, such as for data crawled from the World Wide Web. Another application of LD that predates computational methods is the detection of the language of a document for routing to a suitable translator. This application has become even more prominent due to the advent of Machine Translation ("MT") methods: for MT to be applied to translate a document to a target language, it is generally necessary to determine the source language of the document, and this is the task of LD. LD also plays a part in providing support for the documentation and use of low-resource languages. One area where LD is frequently used in this regard is in linguistic corpus creation, where LD is used to process targeted web crawls to collect text resources for low-resource languages.

A large part of the motivation for this article is the observation that LD lacks a "home discipline", and as such, the literature is fragmented across several fields, including NLP, IR, machine learning, data mining, social medial analysis, computer science education, and systems science. This has hampered the field, in that there have been many instances of research being carried out with only partial knowledge of other work on the topic, and the myriad of published systems and datasets.

Finally, it should be noted that this survey does not make a distinction between languages, language varieties, and dialects. Whatever demarcation is made between languages, varieties and dialects, a LD system is trained to identify the associated document classes. Of course, the more similar two classes are, the more challenging it is for a LD system to discriminate between them. Training a system to discriminate between similar languages such as Croatian and Serbian [6], language varieties like Brazilian and European Portuguese [7], or a set of Arabic dialects [8] is more challenging than training systems to discriminate between, for example, Japanese and Finnish. Even so, as evidenced in this article, from a computational perspective, the algorithms and features used to discriminate between languages, language varieties, and dialects are identical.

RELATED WORK

LI as a task predates computational methods – the earliest interest in the area was motivatedby the needs of translators, and simple manual methods were developed to quickly identifydocuments in specific languages.LI is in some ways a special case of text categorization, and previous research has examined applying standard text categorization methods to LI [9].TheDeep Learning-Based models for LI [10] for code-mixed socialmedia corpora illustrated and it is concluded that for large corpus these models are

suitable. The recordable fact may be both text and audio to the spoken language identification. The work presented in [11] focused on audio-based language identification. The complexity of language identification in transfer learning while code switching [12, 13, 14] illustrated.LI as Text Categorization LI is in some ways a special case of text categorization, and previous research has examined applying standard text categorization methods to However, LI has characteristics that make it different from typical text categorization tasks.

Text categorization[16, 17, 18] tends to use statistics about the frequency of words to model documents, but for LI purposes there is no universal notion of a word: LI must cater for languages where whitespace is not used to denote word boundaries. Furthermore, the determination of the appropriate word tokenization strategy for a given document.

In LI, classes can be somewhat multi-modal, in that text in the same language can sometimes be written with different orthographies and stored in different encodings but correspond to the same class.

In LI, labels are non-overlapping and mutually exclusive, meaning that a text can only be written in one language. This does not preclude the existence of multilingual documents which contain text in more than one language, but when this is the case, the document can always be uniquely divided into monolingual segments. This contrasts with text categorization involving multi-labelled documents, where it is generally not possible to associate specific segments of the document with specific labels.These distinguishing characteristics present unique challenges and offer opportunities, so much so that research in LI has generally proceeded independently of text categorization research.

METHODOLOGY

We have used the language detection data set which is publicly available here [19]. The data set consists of the text of 17 different languages as described below.The data set resulted 10267 unique feature and no stop word removal or stemming methods specific to any language not used. We have used 80 and 20 percent of the data set for training and testing purpose respectively.

Table 1: Each language documents in the data set

S. No	Language	# Of Documents
1	English	1385
2	Portuguese	739
3	French	1014
4	Greek	365
5	Dutch	546
6	Spanish	819
7	German	470
8	Russian	692
9	Danish	428
10	Italian	698
11	Turkish	474
12	Swedish	676
13	Arabic	536
14	Malayalam	594
15	Hindi	62
16	Tamil	469
17	Kannada	369
18	Hindi	63

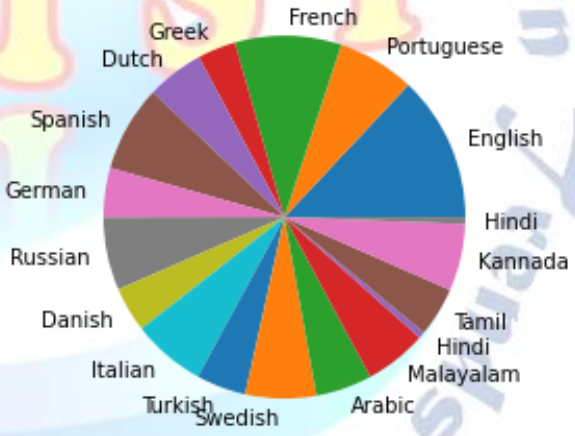

Fig: Pie chart of languages of dataset showing the percentage of them in the dataset

RESULTS AND ANALASIS

We have used multinomial Naïve Bayes classifier for classification. We have achieved 97 percent accuracy. Even though the dataset is very complex with variety diverse of languages the performance of classifier is good. Language specific pre-processing techniques like stop word removal and stemming can reduce feature vector size furthered. We can also use efficient feature selection or extraction model to minimize time and space constraints.

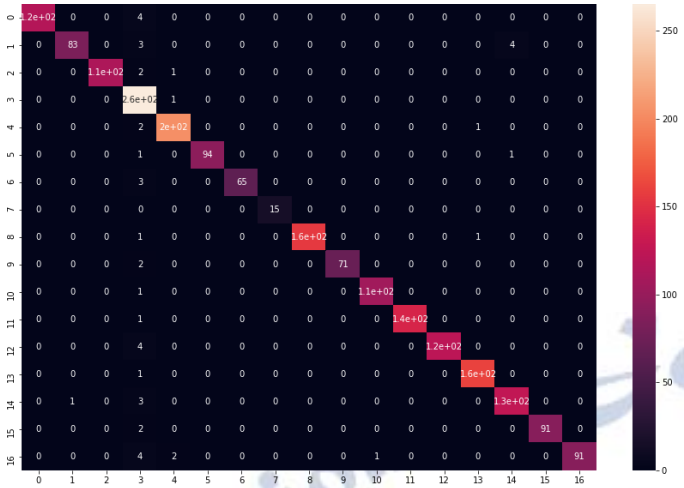

Fig: Confusion matrix resulted from classification

CONCLUSION

We have presented a deep neural network based languageidentification scheme that achieves near perfect accuracy inclassifying dissimilar languages and about 90% accuracy onhighly similar languages.At this point, wethink, further improvement can only be achieved by designingrule based features by talking to language experts or nativespeakers.Language specific feature engineering processes need to apply for efficient language identification.

REFERENCES

[1] TommiJauhiainen et al.," Automatic Language Identification in Texts: A Survey", 10/2018, Journal of Artificial Intelligence Research.

[2] Gary F. Simons et al., "Ethnologue: Languages of the World", Twentieth Edition. SIL International, Dallas, USA, 2017. http://www.ethnologue.com.

[3] Baden Hughes et al., "Reconsidering Language Identification for Written Language Resources", InProceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006),pages 485–488, Genoa, Italy, 2006.

[4] Paul McNamee, "Language Identification: A Solved Problem Suitable for Undergraduate Instruction", Journal of Computing Sciences in Colleges, 20(3):94–101, 2005.

[5] AnettKralisch and Thomas Mandl, "Barriers to Information Access Across Languages on the Internet: Network and Language Effects", In Proceedings of the 39th Annual Hawaii International Conference on System Sciences, volume 3, page 54b, Kauai, USA, 2006.

[6] Nikola Ljubeˇsi´c and Denis Kranjci´c, "Discriminating between VERY Similar Languages among Twitter Users", InProceedings of the 9th Language Technologies Conference, pages 90–94, Ljubljana, Slovenia, 2014.

[7] Marcos Zampieri and BinyamGebrekidanGebre, " Automatic Identification of Language Varieties: The Case of Portuguese", In Proceedings of The 11th Conference on Natural Language Processing (KONVENS 2012), pages 233–237, Vienna, Austria, 2012.

[8] Marcos Zampieri, Liling Tan, Nikola Ljubeˇsi´c, J¨org Tiedemann, and PreslavNakov, " Overview of the DSL Shared Task 2015", In Proceedings of the Joint Workshop on Lan-guage Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial), pages 1–9, Hissar, Bulgaria, 2015b.

[9] William B. Cavnar and John M. Trenkle, "N-Gram-Based Text Categorization", InPro- ceedings of SDAIR-94, Third Annual Symposium on Document Analysis and Information Retrieval, pages 161–175, Las Vegas, USA, 1994.

[10] AnupamJamatia et al., Deep Learning-Based Language Identification in English-Hindi-Bengali Code-Mixed Social Media Corpora", J. Intell. Syst. 2019; 28(3): 399–408. https://doi.org/10.1515/jisys-2017-0440.

[11] GundeepSingh et al., "Spoken Language Identification Using Deep Learning", Hindawi , Computational Intelligence and Neuroscience Volume 2021, Article ID 5123671. https://doi.org/10.1155/2021/5123671.

[12] Aguilar, Gustavo and Solorio ,Thamar, "From English to Code-Switching: Transfer Learning with Strong Morphological Clues", 2019.

[13] Utsab Barman et a., "Code Mixing: A Challenge for Language Identification in the Language ofSocial Media", Proceedings of The First Workshop on Computational Approaches to Code Switching, pages 13–23, October 25, 2014.

[14] DeepthiMave, "Language Identification and Analysis of Code-SwitchedSocial Media Text", Proceedings of The Third Workshop on Computational Approaches to Code-Switching, pages 51–61, July ,2018.

[15] Giovanni Molina et al., "Overview for the Second Shared Task on Language Identification in Code-Switched Data", Proceedings of the Second Workshop on Computational Approaches to Code Switching, pages 40–49,Nov, 2016.

[16] G.Ravi Kumar, S.Rahamat Basha, Surya Bhupal Rao, "A Summarization on Text Mining Techniques for Information Extracting from Applications and Issues", Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 324-332, DOI:https://doi.org/10.26782/jmcms.spl.5/2020.01.00026

[17] Surya Bhupal Rao, S.Rahamat Basha, G.Ravi Kumar, "A Comparative approach of Text Mining: Classification, Clustering and Extraction Techniques", Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 120-131, DOI:https://doi.org/10.26782/jmcms.spl.5/2020.01.00010

[18] S. Rahamat Basha, J. Keziya Rani, JJC Prasad Yadav, "A Novel Summarization-based Approach for Feature Reduction, Enhancing Text Classification Accuracy" Engineering, Technology & Applied Science Research, ISSN -1792-8036, Vol. 9, No. 6, Dec 2019, PP 5001-5005.

[19] https://www.kaggle.com/basilb2s/language-detection.