International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

# Language Detection using Machine Learning

[2]MANSOOR J, [1]Prof.Md. IRSHAD HUSSAIN B

[2]Dept of Master of Computer Applications, University B.D.T College of Engineering

Davangere, Karnataka, India

[2]Assist Professor, Department of Master of Computer Applications, University B.D.T College of Engineering

Davangere, Karnataka, India

## ABSTRACT

We will build a deep learning model able to detect the languages from short pieces of text with high accuracy using neural networks. Dialect Perception is an important and challenging field of research because of various complexities like different sizes and fonts of text, line orientation, different illumination conditions and complex backgrounds in natural scene images. The focus of this paper is to detect and identify text. Language identification ("LI") is the problem of determining the natural language that a document or part there of is written in. Automatic LI has been extensively researched for over fifty years. Today, LI is a key part of many text processing pipelines, as text processing techniques generally assume that the language of the input text is known. Research in this area has recently been especially active. This article provides a brief history of LI research, and an extensive survey of the features and methods used in the LI literature. We describe the features and methods using a

unified notation, to make the relationships between methods clearer.

## I.INTRODUCTION

Most references define dialect as the accent. However, the accent refers to the speaker's pronunciation, while dialect is the speaker's grammatical, lexical, and phonological variation in pronunciation [1]. For our work, we will consider dialect as the pronunciation pattern or the language vocabulary used by a specific community of native speakers [2], those who are usually based in a certain geographical region. Dialect represents an important characteristic of a speaker's voice signature, as it can provide information about the speaker's origin, gender, age, and health status. Automatic Dialect Identification (ADI) has attracted both academia and industry for its promising positive impact on society. Robust ADI is expected to improve Speech Recognition Systems (SRS), which exist in most of today's electronic devices; ADI is also expected to enhance human computer interaction applications and secure remote access communication. In addition, ADI will help in providing new services for e-health and telemedicine, especially important for older

and/or homebound people. Dialect Identification is assumed to be challenging due to its sensitivity to language changes, and regional limitations [3]. The approaches to dialect identification are similar to those used in language identification. These approaches can be classified into two modeling classes: acoustic and phonotactic. Acoustic modeling - as the word implies - deals with spectral feature modeling, while the phonotactic approach deals with speech via phone recognition, language models, and their subsequent scoring. The following subsections highlight both of these approaches.

The rest of this paper is organized as follows: Section II focuses on dialect identification modeling schemes, while Section III covers most of the existing databases in the area of speech recognition. Section IV discusses work done on dialect identification from a linguistic and methodology point of view, and finally Section V gives a brief summary and potential topics to consider for future work.

## II.LITERATURE SURVEY

In 1974, Doddington and Leopard [4], Leopard [5] have explored frequency of occurrences of certain reference sound units in different languages. The average accuracy of 64% and 80% have been achieved using five and seven languages respectively

In 1977,House and Neuberg [6] conducted LID studies on manually phonetic transcribed data. The language related information has been extracted from a broad phonetic transcription instead of using acoustic features extracted from speech signal. Speech signal has been considered as a sequence of symbols chosen from a set. The elements of the set are defined as follows: stop consonant, fricative consonant, vowel and silence. Language identification experiment has been carried out on eight languages. In this work, Hidden Markov Model (HMM) has been trained using broad phonetic labelled data derived from phonetic transcription.

In 1980, Li and Edwards [7] developed automatic LIDsystem based on automatic acoustic phonetic segmentation of speech and automatic LID system has been developed using five languages. Hidden Markov Models(HMM) have been used for developing language models. Recognition accuracy of 80% has been achieved with this approach

In 1994, Muthusamy et al. [8] have proposed a benchmark for language identification task. Perceptual studies with listeners from different language backgrounds have been conducted. The experiments have been conducted on ten languages from OGI-MLTS database. The result obtained from the subjects reported as the benchmark for evaluating the LID performances obtained from automatic LID system.

In 1994, Berkling et al. [9] have analyzed phoneme based features for language recognition. They have performed the LID study on three languages: English, Japanese and German.

The phonemes which can provide the best discrimination between language pairs have used to build the superset. The experimental analysis drawn the conclusion that, to develop a LID system with large number of languages, it may be useful to reduce the number of features despite a small loss in LID accuracy.

In 1994, Tucker et al. [10] have conducted LID experiments with the languages belong to same language family. Sub-word models for English, Dutch and Norwegian languages have been developed for carrying out the LID study. Two types of language models: language independent and language-specific models have been developed in this study. Three techniques namely (i) the acoustic differences between the phonemes of each language, (ii) the relative frequencies of phonemes of each language and (iii) the combination of previous two sources have been explored for classifying the languages. The third technique provides average LID accuracy of 90% for three languages.

In 1994, Zissman and Singer [11] have carried out a comparative study using four approaches: (i) Gaussian mixture model based classification, (ii) phoneme recognition followed by language modeling (PRLM), (iii) parallel PRLM (PRLM-P) and (iv) language-dependent parallel phoneme recognition (PPR). The OGI-MLTS corpus has been used to evaluate the performances of the four LID approaches. The LID study showed that, best performance is obtained with PRLM-P system, which does not require labelled speech corpus for developing language models.

## COMPARISION AND DISCUSSION

| Author name | Features | Remarks |
|---|---|---|
| A.S.House , E.P.Neuberg [6] | Broad phonetic transcription (i.e. ,stop consonant, fricative consonant, vowel and silence) | Phonotactic Information is Language specific |
| K.P.Li , T.J.Edwards [7] | Acoustic-Phonetic-information | Recognition Accuracy of 80% has been achieved |
| K.M.Berkling, T.Arai, e.Bernard [9] | PLP coeffi cients with 56 dimensions | To develop LID system with large number of languages. |
| R.C.F.Tucker, M.Carey, E.Parris [10] | Acoustic differences between the phonemes, relative frequency of phonemes and combination of previous two sources of information | 90% LID accuracy is Achieved |
| M.A.Zissman, E.Singer [11] | MFCC | PRLM-P provides best accuracy of 79.2% |
| S.Kadambe, J.Hieronymus [12] | Phoneme inventory, phonemotactics, syllable structure, lexical and prosodic differences | 88% accuracy is achieved. Language-specific information can be captured using higher order linguis tic knowledge |
| T.J.Hazen, V.W.Zue [13] | Phonotactic, acoustic phonetic and prosodic information | Phonotactic information is most useful information for |

| | | |
|---|---|---|
| | | LID task |
| S.M.Siniscalchi, J.Reed, T.Svendsen, C.H.Lee [14] | Manner and place of articulations of sound units | Universal set of language-specific fundamental units is proposed |
| G.R.Botha. E.Barnard [15] | *n*-gram statistics as features used for text based LID | The SVM classifier outperforms other classifiers and 99.4% accuracy is achieved |
| V.Ramasubramanian, A.K.V.S.Jayaram, T.V.Sreenivas [16] | MFCC | The performance of E-HMM based system is Superiorcompared to GMM |
| R.Tong, B.Ma, H.Li,E. S.Chng [17] | Lexical constraints and phonotactic patterns | TOPTs derived from UPRs outperform those from language. |

Table 1:Features and Remarks of Literature Survey

## III.DATASETS

The dataset that is used for this task consisted of .json files in which the Tweet objects are stored. Each Tweet object contains different types of relevant information about its nature, such as the unique identifier of the tweet itself, the text that it contains, the information about its author, time and location at the point of creation, etc. However, only parts of this information are considered relevant for the classification task. The files that are used contain tweets collected using the Twitter API in April 2012, where in total around 22,000 tweets in 16 different languages are randomly collected at different time points during two days. Languages appearing in less than 3 tweets are discarded and due to insufficient domain knowledge, Indonesian and Malay are grouped together to one language. The language distribution across the dataset is shown in Fig. 1. When collecting the data, we complied with the Twitter's Terms of Use. As expected, almost half of the total number of tweets are written in English. The languages following English by the number of tweets are Malay, Japanese, Korean, Spanish, Portuguese, and Dutch. Even though the distribution of the number of tweets per language in the dataset is highly skewed, it corresponds quite well to the distribution of Twitter languages given in [18]

To make the language labeling of the tweets easier and reduce the manual work, the open-source language detection library Chromium Compact Language Detector 2 (CLD2) is used. Finally, the column indicating the tweet language is added to the existing .csv file. However, since only around 8,000 tweets are obtained using this rather time consuming approach, additional tweets are collected from the users with the user ID already existing in the dataset, assuming the majority of users would tweet only in one or two languages.

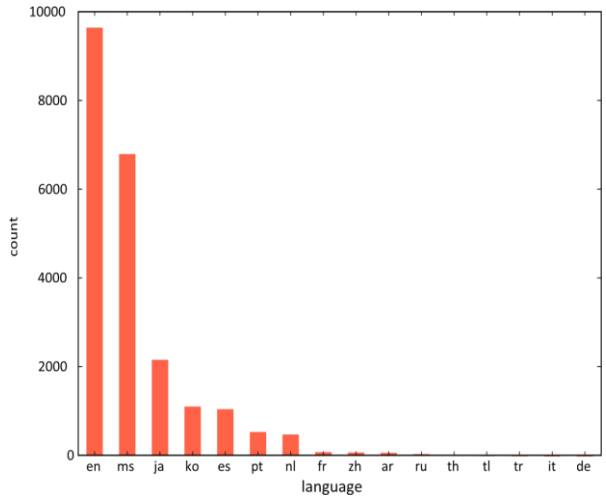


Figure 1: Distrubution of languages in dataset [18]

## IV.METHODOLOGY

The general task of language detection is to predict for a given piece of text the language in which the text is written. This approach is to solving the language detection task would be to show the text to a certain language expert, who would then decide on a language the text is written.

In the machine learning approach to solving language detection problem ,we are given a certain amount of data(i.e, a set of texts in different languages) and the labels (languages to which those texts belong)

**Preprocessing**

The preprocessing task is usually the first part in a machine learning document processing pipeline, preceding the extraction of features from the data. In this paper, preprocessing included editing the tweet texts and assigning them the corresponding language labels. As the first preprocessing step, all the links and expressions of addressing a particular user (the *@user name* form) are removed from every tweet text using simple regular expressions, as they are considered irrelevant for the differentiation between languages. In addition, all the emoticons are removed too, since they maintain the same form across the languages. The text is then converted to lowercase, all multiple white spaces are trimmed, and all the punctuation marks are removed. This procedure transformed all the texts into an equal format, to improve the accuracy when performing their mutual comparisons.

**Classification**

**SVM.**Support vector machines(SVMs) are supervised learning models ,used mostly for classification and regression problems

SVM classification is described for the case of only two classes for simplicity, since multi-class classification is just an extension of that model [19]. The multi-class support in this paper is handled according to the one-vs-one scheme. SVM classification is focused on trying to maximize the *margin*, i.e. the distance of the data points of both classes from the decision boundary based on structural risk minimization [20]. One way to achieve this is by solving the dual optimization problem [21]. The dual optimization problem is defined as follows:

Max $\alpha$subject to 0

where $\mathbf{x}_i$ are the feature vectors and $y_i$ are the corresponding class labels. The function $k(\mathbf{x}_i,\mathbf{x}_j)$ is the socalled *kernel* function, which describes the similarity between two documents and allows the extension of SVMs to

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\leq \alpha_i \leq C,\ i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

nonlinear problems. The parameter $C$ is a regularization constant, which allows for some points in the training set to be misclassified, in order to avoid overfitting. All data points with $\alpha_i > 0$ are the so-called *support vectors*.

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{n} y_i \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b\right)$$

Where b is the bias term . One of the most important points to consider when choosing an SVM as a classification method.

**Logistic Regression**

*Logistic regression*, despite having the word "regression" as part of its name, is a linear model for *classification* rather than regression [22]. The logistic regression classification paradigm is described here for the two class case only. It is a type of probabilistic statistical model, where the probabilities describing the possible assignments to different classes are modeled using the logistic function , which is defined as:

$$P[y_i = +1|x_i, w] = \frac{e^{w^\top x_i}}{1 + e^{w^\top x_i}}$$

where $y \in \{-1, +1\}$ is the assigned class label, $x_i$ is the data point, $w$ is the regression coefficient, and $P[y = +1|x_i, w]$ is the probability of $x_i$ being drawn from the positive class. A new data point $x_i$ gets assigned to a class with the highest probability. As an optimization problem, two-class L2-penalized logistic regression minimizes the following cost function: [22]

$$\min_w \frac{1}{2} w^\top w + C \sum_{i=1}^{n} log(e^{-y_i X_i^\top w} + 1)$$

where $\frac{1}{2} w^\top w$ is the L2-regularization and $C$ is the inverse regularization constant

V.EXPERIMENTS AND RESULTS

Using the obtained dataset and network architecture introduced in section 3 ,we conducted several experiments to assess the performances of our proposed network architecture on several kinds of input data .

While performing our experiments we had a range of different questions in mind those are

- Can we increase the classification accuracy with the help of a network, that combines a CNN with a LSTM, compared to a CNN-only approach?
- Is the network able to reliably discriminate between languages?
- Is the network robust against different forms of noise in the input data? – Can the network easily be extended to handle other languages as well?

In the above several range of questions that are came out when the performing on the experiments. We are having the solution or the type of results on those question or solving methods

**Accuracy Measures**

A good classifier is defined as one for which the number of true positives (TP) and true negatives (TN) is high, while at the same time keeping the number of false positives (FP) and false negatives (FN) low. To sum up those outcomes, two different accuracy measures were used in assessing the classifiers performance: micro- and macro-averaged F1-score. In order to understand the F1-score, precision and recall need to be explained first. *Precision* is defined as:

$$\text{precision} = \frac{TP}{TP + FP}$$

and it measures the ability of the classifier not to assign a sample to the class to which it does not belong. *Recall* is defined as:

$$TP$$

$$\text{recall} = \frac{}{TP + FN}$$

and it measures the ability of the classifier to find all the samples that belong to that class (both assigned to it and the ones not assigned). *F1-score* is defined as:

$$\text{F1-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

and it is the weighted average (harmonic mean) of precision and recall. The *micro-averaged* F1-score calculates the metrics globally by counting the total number of TPs, FPs, and FNs, while the *macroaveraged* F1-score calculates the metrics for each label and finds their unweighted mean, not taking label imbalance into account [23] Because of the large difference in sample sizes between the languages in the dataset used in this paper, the difference between the micro- and macro-averaged F1-scores is expected to be large as well.

## VI.CONCLUSION

In this paper, different algorithmic approaches to language detection for short texts in social media are investigated. So many experiments are done and also facing few problems on the performing the operation. In the result part we gave the solution for those problems adding the accuracy measure for language detection. The best approach is includes the use of the well-known classifiers such as SVM and logistic regression and the combination of both.

## VII.FUTURE WORK

Future work is required in dealing with language detection in multilingual documents. Previous work shows the indentifying the small amount piece of data. The main intend to investigate if these results are also applicable to language indentification in multilingual documents and other open question is the extension of the geretive mixture models to unknown language indentification.

## REFERENCES

[1] J. K. Chambers and P. Trudgill, ―Dialectology‖, chapter one, pp. 4-9, 2nd edition, Cambridge University press, 1998.

[2] L. Gang, H. John, and L. Hansen, ―A Systematic Strategy for Robust Automatic Dialect Identification‖, 19th European Signal Processing Conference (EUSIPCO 2011), pp. 2138-2141, 2011.

[3] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, ―Dialect identification using Gaussian mixture models", Proceedings of the Speaker and Language Recognition Workshop (Odyssey '04), pp. 41-44, Toledo, Spain, 2004.

[4] R. Leonard, G. Doddington, Automatic language identification. Technical Report RADC-TR74-200 (Air Force Rome Air Development Center, Technical Report) August 1974

[5]. R. Leonard, Language Recognition Test and Evaluation. Technical Report RADCTR-80-83 (Air Force Rome Air Development Center, Technical Report). March 1980

[6]. A.S. House, E.P. Neuberg, Toward automatic identification of the languages of an utterance. J.Acoust. Soc. Am. **62**(3), 708–713 (1977)

[7]. K.P. Li, T.J. Edwards, Statistical models for automatic language identification, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 884–887, April 1980

[8]. Y. Muthusamy, R. Cole, M. Gopalakrishnan, A segment-based approach to automatic language identification, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 353–356, April 1991

[9]. K.M. Berkling, T. Arai, E. Bernard, Analysis of phoneme based features for language identification, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. I/289–I/292, April 1994

[10]. R.C.F. Tucker, M. Carey, E. Parris, Automatic language identification using sub-word models,in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. I/301– I/30, April 1994

[11]. M.A. Zissman, E. Singer, Automatic language identification of telephone speech messagesusing phoneme recognition and N-gram modeling, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1 pp. I/305–I/308, (1994)

[12]. S. Kadambe, J. Hieronymus, Language identification with phonological and lexical models, in*International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 3507– 351, May 1995

[13]. T.J. Hazen, V.W. Zue, Segment-based automatic language identification. J. Acoust. Soc. Am.**101**, 2323–2331 (1997)

[14].S.M.Siniscalchi,J.Reed,T.Svendsen,C.-H.Lee,Universalattributecharacterizationofspokenlanguages for automatic spoken language recognition. Comput. Speech Lang. **27**(1), 209–227 (2013)

[15]. G.R. Botha, E. Barnard, Factors that affect the accuracy of text-based language identification.Comput. Speech Lang. **26**(5), 307–320 (2012)

[16].V. Ramasubramanian, A.K.V.S. Jayram, T.V. Sreenivas, Language identification using parallel sub-word recognition - an ergodic HMM equivalence, *European Speech Communication Association (EUROSPEECH)* (Geneva, Switzerland), September 2003

[17]. R. Tong, B. Ma, H. Li, E.S. Chng, A target-oriented phonotactic front-end for spoken languagerecognition. IEEE Trans. Audio Speech Lang. Process. **17**(7), 1335–1347 (2009)

[18]. Delia Mocanu, Andrea Baronchelli, Bruno Gon¸calves, Nicola Perra, and Alessandro Vespignani. The twitter of babel: Mapping world languages through microblogging platforms. *CoRR*, abs/1212.5238, 2012.

[19]. Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292, March 2002.

[20]. Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[21]. K.-R. Mu¨ller, S. Mika, G. R¨atsch, S. Tsuda, and B Sch¨olkopf. An introduction to kernelbased learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–202, 2001.

[22]. Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[23]. F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.