JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Language Identification & Analysis of Code-Switching Language

Nidhi Deepak, Sara Mouzem, Ramya Reddy

Information Technology

Stanley College of Engineering & Technology for Women, Hyderabad, India

Abstract: In a multilingual or socio-lingual configuration Intra-sentential Code Switching (ICS) or Code Mixing (CM) is frequently observed nowadays. In the world most of the people know more than one language. The CM usage is especially apparent in social media platforms. Moreover, ICS is particularly significant in the context of technology, health and law where conveying the upcoming developments are difficult in one's native language. In applications like dialog systems, machine translation, semantic parsing, shallow parsing, etc. CM and Code Switching pose serious challenges. To do any further advancement in code-mixed data, the necessary step is Language Identification. So, Here we present a study of Language Identification in English-Telugu Code Mixed Data. Considering the paucity of resources in code mixed languages, we proposed Logistic Regression where we compare the increase in Accuracy and F1- score before and after Feature Extraction. We have observed that first we got an Accuracy of 73.16% and after feature extraction we got an Accuracy of 89.77% and F1-Score of 0.8957.

Keywords **– Code-switching , code mixing , language identification, NLP, word-level language identification**

I. INTRODUCTION

Language is defined as a group of signal for communication. Thus, Language learning is the source to real-life and social survival. Knowing a language encompasses the ability to speak and communicate in this language. This also signifies the knowledge of linguistic. Language identification remains a difficult problem, especially in social media text where people combines different language in informal style, closely related to the language pair and the concept code switching occurs.

Code – switching occurs when a person switches between two or more languages during communication. In social medial such as Twitter or Facebook has found that it is strategy to achieve certain personal goals in everyday communicative needs, but also many new challenges, in particular since this type of text is characterized and have a high percentage of getting spelling errors and contain creative spelling such as ("goodn8" for 'goodnight'), Meta tags (URL), Hashtags ("#swachbharat"), abbreviations("LOL" for 'Laugh Out Loud') and so on. Non-English speakers don't have use uni codes to express their language, they use phonetic typing, inserting English elements frequently, and often mix multiple language to express their thoughts. All this makes Language Identification a very challenging task. Usually this phenomena is refereed to as 'code-mixing' which means referring to a language that changes occur inside a sentence whereas, Code-switching is more general and in particular use for inter-sequential phenomena. Code-switching is much more prominent in social media as in the following table below example 1 between English- Telugu.

Example 1
"John/NE nuvvu/TE exams/EN baaga/TE prepare/EN aithene/TE ,/UNIV first/EN classlo/TE pass/EN avuthav/TE ./UNIV
John, you will pass in the first class, even if the exams are well prepared)

Table 1 – Example for English – Telugu Code-mixing

Word Level language identification can be difficult because it has mixing at clause level, word , and sub-word level. For example, Identification in English-Telugu due to the relatedness factor in them. In spite Telugu has a native script most of the Telugu social media text is translated. Due to this condition we type Telugu language in English words for (e.g. ikkada, akkada, illu). Some Telugu words can take the same form as an English word. The words 'peru' pani', ' are some of the examples. The first step in handling

computational modeling of code-switching is to deal with Language Identification first. Language identification is very important for a wide variety of end user applications such as information extraction systems, as well as tools for language assessment for bilingual children. In addition to this, it also enables socio-linguistically and pragmatic studies.

II. PURPOSE OF THE STUDY

Shifting to a place where English is the only common language among Telugu students studying in India tend to shift to English or mix English with Telugu when communicating with others the same group of students however may switch to hindi when engaging in conversation with either Indians especially when bringing up Indian cultural-related topics where some words or names lack the clear equivalent in English parallel to this students intentionally or unintentionally shift to Spanish when among Indian colleagues or friends of other nationalities the purpose of this study is to broadly examine the reasons governing students inclination to switch or mix in an attempt to provide a full account of the context of code-switching

III. LITERATURE SURVEY

A wave of research has been embarked upon on the same area. What is more, many linguists have decided on defining the two processes as being similar or the same, whereas others have discriminated between them. It is worth noting that code-switching and code- mixing are never synonymous, and one of them will not, definitely, mean the other. While Hymes (1974) defines code-switching as the substitution of languages when speakers use a diversity of language registers, Halmari (2004, p. 115) views code-mixing as "the process of mixing of two or more languages within the same conversational episode". Myers-Scotton (1993) denotes the fact that within the same speech act, there are substitutions of linguistic varieties, and this process gives rise to code-switching. Muskellunge & Chimbarange (2012) refer to code as a variety of the same language and the whole system of a language. According to Ester (2021), code-switching is a substitution If two or more languages, dialects or varieties in the course of a single conversation. Code-mixing, however, indicates the hybridisation of two languages whereby speakers may tend to use English roots with another language morphology. According to Waris (2012), code-switching occurrence is not limited to social community as it also takes place in classrooms. For teachers to convey meanings effectively, they, sometimes, intentionally code-switch between the target language and the mother tongue in an attempt to maximise learning absorbance. M. R. Hasan et al.(2019)Presented Twitter data to study public opinions on a product. Firstly, to filter tweets, we have developed an NLP based per-processed data system. Secondly, to evaluate sentiment, we integrate the model definition of Bag of Words and Term Frequency-Inverse Text Frequency (TF-IDF). It is an effort to utilize BoW & TFIDF together to distinguish positive & negative tweets accurately. We have originated that the accuracy of SA can be significantly improved using the TF-IDF vectorizer, and simulation results indicate the efficacy of our proposed method. Using NLP technology, they achieved 85.25 percent precision in sentiment analysis.

IV. DATASET DESCRIPTION

We use the English Telugu code mixed dataset for language identification from the Twelfth International Conference on Natural Language Processing (ICON-2015). The dataset is comprised of Facebook posts, comments, replies to comments, tweets and WhatsApp chat conversations. This dataset contains 1987 code-mixed sentences. These sentences are tokenized into words. And the tokenized words of each sentence are separated by new line. The dataset contains 29503 tokens.

Language Label	Label Frequency	Percentage of Label
Telugu	8828	29.92
English	8886	30.11
Universal	11033	37.39
Named Entity	756	2.56

Table 2 – Analysis of corpus

V. PROPOSED METHODOLOGY

Approaches for Word Level Language Identification (LI) LI is the process of assigning a language identification label to each word in a sentence, based on both its syntax as well as its context. We have implemented baseline models using Logistic Regression. The first step is to read the dataset. After this step the following is shown in the form of a flowchart

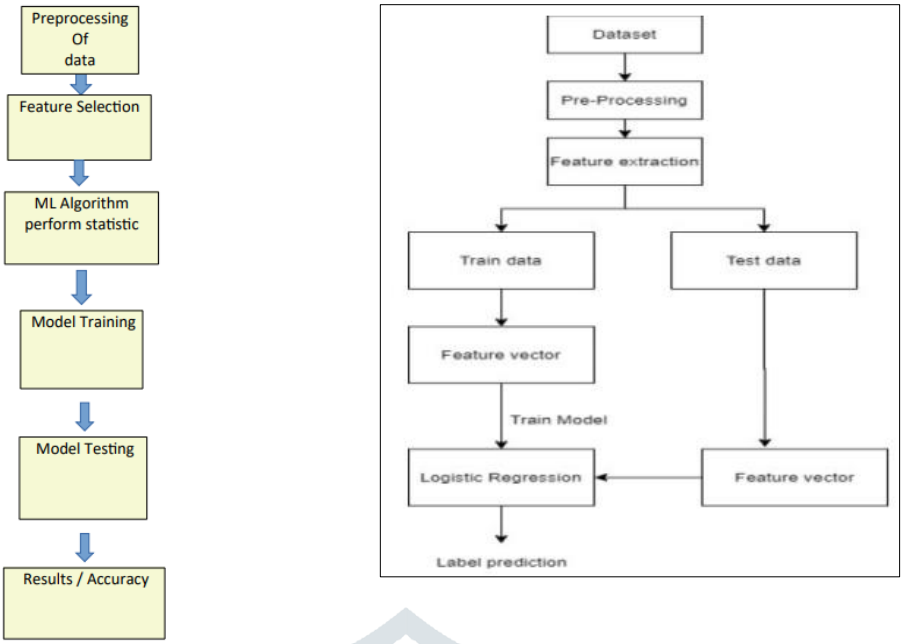

Fig 1- Workflow of the proposed Mode

1) Data Pre- processing

Data preprocessing is an important step to prepare the data to form a machine learning model can understand. There are many important steps in data preprocessing, such as data cleaning, data transformation, and feature selection.

- Lower case: Since we would expect to treat "Food" and "food" as the same word, without creating various predicting powers, I've down-cased each word.
- Contractions: I've replaced contractions with their longer forms such as "isn't": "is not", "can't": "cannot". To do so, I've imported contractions list from here.
- Remove special characters:I've cleaned the data from any special character such as double quotes, punctuation, and possessive pronouns.
- Stop words: I've removed stopwords since they add noise without bringing any information value in modeling. I've downloaded a list of English stopwords from the nltk package and deleted them from the text corpus.
- Tokenization: to process text, we need to split it into smaller chunks. Here, I've split sentences into words using WordPunctTokenizer from the nltk library

2) Feature extraction

In the feature extraction phase, text-based data is transformed into numerical data in the form of a feature vector. This vector represents the unique characteristics of the text and can be generated by any sequence of mathematical transformations . Since machine learning models do not accept the raw text as input data, we need to convert into vectors of numbers. There are different ways of transforming text into numeric vectors. In this analysis, I've applied first Bag of Words, followed by Bag-of-n-Grams, and later I've moved to Tf-Idf which is a more complex representation.

- **Bag of Words (BoW)**

It is a simple but still very effective way of representing text. It has great success in language modeling and text classification. It is based on the word count statistics.

- **Count Vectorizer**

Count Vectorization involves counting the number of occurrences each words appears in a document (i.e distinct text such as an article, book, even a paragraph!). Python's Sci-kit learn library has a tool called CountVectorizer to accomplish this.

- **Bag of n-Grams**

It is an extension of Bag-of-Words and represents n-grams as a sequence of n tokens. In other words, a word is 1-gram (unigram), two words are 2-grams (bigram), etc. It is applied in the NLP pipeline because it retains the original sequence of the text more than the Bag of Words representation. However, it has a very high computational cost, because theoretically k unique words can mean k^2 unique bigrams.

- **Tf-Idf Tf-Idf**

stands for term frequency-inverse document frequency, and instead of calculating the counts of each word in each document of the dataset

*Tf-idf(w, d)= Bow(w, d) * log(Total Number of Documents /(Number of documents in which word w appears))* .....Eq 3.1

If a word appears often in a particular document, but not in so many other documents, it is most likely that the word represents a particular meaning for that document and receives a larger count than before thanks to high Idf. On the other side, if a word appears in many documents, then its Idf is close to 1 and the logarithm turns 1 into 0 and decreases its effect.

- **N- Gram**:

The N-Gram could be comprised of large blocks of words, or smaller sets of syllables. N-Grams are used as the basis for functioning N-Gram models, which are instrumental in natural language processing as a way of predicting upcoming text or speech.

3) Code-Mixed Language Identification

Language Identification is the process of dividing words into one of the mentioned classes. We mainly consider character sequences of the word as features for the system. The challenging part is when a word might belong to two different classes.
Example: Input: plz watch it nd share chusaka nenu cheppanavasarm le meere share chestharuuu
Output: plz-en watch-en it-en nd-en share-en chusaka-en nenu-en cheppanavasarm-en le-te mere_en share-en chestharuuu .
Language Identification is the process of dividing words into one of the mentioned classes. Initially we are taking this problem same as the pos tagging problem.

- lexical feature: – word
- sub-lexical features: – Prefix, suffix character strings – n-grams of word – prefix, Suffix character strings of neighboring words
- other features: – length of the word – neighboring words – pos tag of word – pos tag of previous words

4) Machine Leaning Algorithms

Now after removing all the noise in the data we apply statistical algorithm to perform Language Identification. So here we use Logistic Regression Algorithm as the best way to determine the future sequential order.

- **Logistic Regression**

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for a given set of features(or inputs), X. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Input: Training data

1. For i←1 to k
2. For each training data instance d_i:
3. Set the target value for the regression to
$$z_i \leftarrow \frac{y_j - P(1 \mid d_j)}{[P(1 \mid d_j).(1 - P(1 \mid d_j))]}$$
4. nitialize the weight of instance d_j to $P(1 \mid d_j).(1 - P(1 \mid d_j)$
5. finalize a f(j) to the data with class value (z_j) & weights (wj)
Classification Label Decision
6. Assign (class label:1) if $P(1|d_j) > 0.5$, otherwise (class label: 2)

Fig 2– Pseudo Code for Logistic Regression

5) Training the model

Training the model for 80% of the data with the applied ML Algorithm in order to testing the data

6) Test the model for the test data

Testing 20% of the data on the trained data in order for prediction of the results and Accuracy.

VI. RESULTS

The language identification was performed done by Logistic Regression. The main reason for predicting the wrong language tag is the variation in tag used in the train data of English Telugu words. Our best performance system for tagging the language tag for a word is Logistic Regression with f1-score: 0.89 and accuracy: 89.1788% .
In this work some interesting problems are encountered like Romanization of Telugu words, different types of syntax in social media text...etc.

Label	Precision	Recall	F1-Score
Telugu	0.84	0.81	0.87
English	0.88	0.87	0.88
NE	0.93	0.93	0.95
Universal	0.48	0.39	0.54
Average	0.89	0.90	0.91

Table 3- Experimental results of each tag

Since there is no standard way to transliterate the code mixed data and Romanization contributes a lot to the spelling errors in foreign words. For example, a single Telugu word can have the more than one spelling (Eg. “avaru”, “evaru”, “aivaru”, “yevaru”. Translation into English: “who”). This posed a significant challenge for language identification. Similarly, In social media, chat conversation using SMS language “you” can be written as “U”, “Hai” – “Hi”, “Good” – “goooooood”...etc. Such non standard usage is an issue for language identification

RESULTS BEFORE FEATURE EXTRACTION
1) N-GRAM TF_ID

	Precision	Recall	F1-Score
en	0.75	0.83	0.79
ne	0.13	0.67	0.22
te	0.67	0.84	0.75
univ	0.81	0.62	0.70

Accuracy – 0.7316949152542372 F1_Score – 0.7341706331946041

Next step is to add FEATURE EXTRACTION The technique of extracting the features is useful when you have a large data set and need to reduce the number of resources without losing any important or relevant information.

AFTER FEATURE EXTRACTION
N-GRAM TF-IDF

	precision	Recall	F1-Score
en	0.91	0.88	0.90
ne	0.28	0.63	0.38
te	0.90	0.88	0.89
univ	0.67	0.72	0.69

Acuuracy -0.8433571185864764 F1_Score – 0.8474631517301036

BEFORE & AFTER FEATURE EXTRACTION PLOT

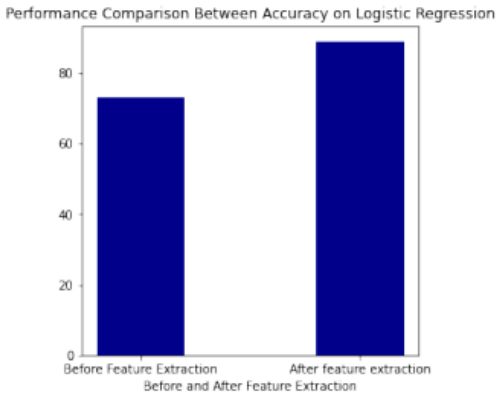

Fig 4- Performance Comparison Between Accuracy on Logistic Regression

Feature extraction helps to reduce the amount of redundant data from the data set. In the end, the reduction of the data helps to build the model with less machine effort and also increases the speed of learning and generalization steps in the machine learning process.

VII. CONCLUSION

The complexity of language identification of codeswitched data depends on the data source, codeswitching behavior, and the typology and relation between the languages involved. We find that the code-switching metrics complement each other in explaining the codeswitching patterns across language pairs. The language identification was performed done by Logistic Regression and gave the best results for our problem. The main reason for predicting the wrong language tag is the variation in tag used in the train data of English Telugu words. Our best performance system for tagging the language tag for a word is conditional random field with 89.77% and F1-Score of 0.8957.

VII. REFERENCES

[1] Deepthi Mave, Suraj Maharjan and Thamar Solorio Department of Computer Science University of Houston , “Language Identification and Analysis of Code-Switched Social Media Text”.

[2] Sarkar, K. and Gayen, V., 2012, November. A practical partof-speech tagger for Bengali. In Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on (pp. 36-40). IEEE.

[3] Anupam Jamatia and Amitava Das TASK REPORT: TOOL CONTEST ON POS TAGGING FOR CODEMIXED INDIAN SOCIAL MEDIA (FACEBOOK, TWITTER, AND WHATSAPP) TEXT @ ICON 2016 In: Proceedings of ICON 2016. 2016

[4]] Neunerdt, M., Trevisan, B., Reyer, M. and Mathar, R., 2013. Part-of-speech tagging for social media texts. In Language Processing and Knowledge in the Web (pp. 139-150). Springer Berlin Heidelberg.

[5] Anupam Jamatia, Bj¨orn Gamb¨ack, and Amitava Das. Part-of-Speech Tagging for CodeMixed EnglishHindi Twitter and Facebook Chat Messages In: Proceedings of Recent Advances in Natural Language Processing. 2015, pp. 239248

[6] Arnav Sharma and Raveesh Motlani POS Tagging For Code-Mixed Indian Social Media Text : Systems from IIIT-H for ICON NLP Tools Contest 12th Inter- national Conference on Natural Language Processing

[7] Burr Settles Biomedical named entity recognition using conditional random fields and rich feature sets. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, pp. 104-107. Association for Computational Linguistics, 2004.