Identifying Kannada and English Code Switch Text

Ramesh Chundi
School of CSA
REVA University
Bangalore, India
rameshreddyc@revainstitution.org

Vishwanath R. Hulipalled
School of C&IT
REVA University
Bangalore, India
vishwanath.rh@reva.edu.in

J.B Simha[1,2]
[1]*Abiba Systems,*
[2]*REVA Academy for Corporate Excellence (RACE)*
REVA University
Bangalore, India
jbsimha@gmail.com

***Abstract* – The users are writing comments or reviews about different events due to the popular usage of social media and smart devices. These comments can be written in both monolingual and bilingual or Code-Switch (CS) text. Nowadays bilingual or code-switch text is common in social media along with monolingual text. Identifying these code-switch text is very important in emotion detection and sentiment analysis. In this paper we focused on the problem of identifying Kannada and English code-switch text by applying different supervised classification techniques. We applied "Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Neutral Network (NN), and Naïve Bayes (NB)" approaches. The experimental results shown that Naïve Bayes and Logistic Regression supervised classification techniques are more accurate than other classification techniques and SVM is least accurate for our data set.**

Keywords – Bilingual, Code-Switch text, Social Media, Supervised Classification Techniques.

I. INTRODUCTION

Popular social media websites such as Youtube, Twitter, and Facebook are used by the visitors to express their views or opinions. Many users are comfortable to use their native language along with English. Using more than one language in a single comment or writing one language in another language script is called as code-switch text. In social media, the data is available in the form of monolingual and bilingual or code-switch text. It is essential to identify the bilingual or code-switch text for emotion detection problems. The emotion can be expressed either in monolingual or bilingual text or emotes. Emotion detection in text is very important in order to obtain useful information for the study on social media [1], [2], [3].

Here we are showing some of the examples [E1] to [E4], which are collected from one of the social media website (youtube.com). These comments have Kannada and English language script. Kannada language is one of the oldest and official languages in southern part of India.

[E1] Sir ಆದಷ್ಟು ಈ ಸಾಂಗ್ ನ್ ನಾನ್ share ಮಾಡ್ತೀನಿ.. Nice song

English Translation: sir I will try my best to share the song.. Nice song

[E2] Dekshakkagi tyaga madiddare jeevana

English Translation: Sacrifice the life for country.

[E3] ಗಿರೀಶ ಕಾರ್ನಾಡ್ ಅವರು ಆತ್ಮಕ್ಕೆ ಶಾಂತಿ ಸಿಗಲಿ ಎಂದು ಹಾರೈಸುತ್ತೇನೆ

English Translation: Let Girish Karnad's soul rest in peace.

[E4] I am proud of you I'm fan of you from this minute.

[E1] is a code-switch text because two languages are used i.e, Kannada and English in a single comment. [E2] is also a code-switch text because Kannada language is written in English script. [E3] is a pure Kannada text comment and [E4] is a pure English text comment. [E1] to [E4], emotions are expressed in all the comments. In [E1] happy emotion is expressed in English text. In [E2] happy emotion is expressed in Kannada language but it is written in English script. In [E3] sad emotion is expressed in pure Kannada language and in [E4] happy emotion is expressed in pure English language.

The rest of the paper is arranged as follows. In section II, we discuss about related work. In section III, data integration and labelling process is discussed. The proposed approach is discussed in section IV. In section V, we discuss the results from our experiments. In section VI, provides conclusions and the scope for future work.

II. RELATED WORK

Identifying code-switch text is one of the basic tasks of Natural Language Processing (NLP). In this section, we will discuss the important related work that has been done on code-switch and bilingual text so far. Many of the researchers are focused on monolingual, but nowadays code-switch text also using on social media to express their views or opinions.

A. Code-Switching and Bilingual Text

One of the problems in natural language processing is code-switching and it has got more attention in the field of research in recent years. There are many approaches have been proposed to identify the code-switch in the text and it is one of the ongoing problems in NLP.

Research on code-switch has been following from 1970s. A few hypotheses are proposed to account for the motivations behind code-switch. For example, diglossia [4], communication accommodation theory [5], the markedness model [6], and conversational analysis model [7]. Code-Switch archives have also got significant considerations in the NLP community [8], [9]. A few investigations have centred on identification in code-switched archives [10], predicting code-switched points in Spanish – English [11], identifying code-switched tokens [12], adding code-switched support to language models [13], and developing POS tagging for code-switching text [14].

Multilingual natural languages processing has started to draw increasingly more consideration in the computational linguistic community because of its wide real world applications. Significant investigations have been accounted in various natural language processing tasks. For example parsing [15], information retrieval [16], text classification [17], and so on.

Bail et al., performed analysis in English-Hindi posts from facebook and they found that at least 4.2% of the data is code-switched [18]. Performed experiments on language identification, transliteration, normalization and POS tagging. The POS tagger accuracy fell by 14% to 65% without using gold language labels and normalization [19]. Addressed the problem of shadow parsing of Hindi-English code-mixed social media text [20], and also addressed the problem of language identification on Bengali-Hindi-English facebook comments [21]. In addition that question classification system for Hindi-Englishcode-mixed languages [22].

Language identification was done on code-mixed English-Kannada social media text by using supervised classification methods [23]. They are embedding a dictionary model to handle word level code-mixing.

Code-Switch can be in different forms, for example, [E1] has both English and Kannada languages which is written in its own script. In [E2] entire Kannada language is written in English script. These two ([E1] and [E2]) are the examples for cod-switch comments. Very less work is carried out in English-Kannada to identify whether the given text is code-switch or not. In our work we want to address the problem of identifying Kannada and English code-switch text.

III. DATA COLLECTION AND ANNOTATION

In this section, we will discuss about data collection and data preparation to apply different supervised classification methods to identify the code-switch text (CS) or Plain Text (PT).

A. Data Collection

We collected text comments from youtube.com, one of the famous social media websites. There are 2266 text comments in the data set, in that some are code-switch text comments and some are plain text comments. Code-switch text comments are two types, one is using more than one language script in a single comment. And another one is writing one language in another language script. Plain text also have two types of comments, one is pure English text and another one is pure Kannada text. [E1] and [E2] are the examples for code-switch comments and [E3] is an example for pure Kannada text comment and [E4] is an example for pure English text comment.

B. Annotation

We had done manual labeling for code-switch text as CS and plain text as PT after collecting the data. [E1] and [E2] are denoted as CS, [E3] and [E4] are denoted as PT. We have 1263 code-switch (CS) text comments and 1003 plain text (PT) comments in our data set after annotation.

IV. PROPOSED APPROACH

We are using python 3 tool to implement our proposed method along with some of the python libraries such as pandas, numpy, preprocessing, sklearn.feature_extraction, sklearn.model_selection, and sklearn.metrices. In Fig.1, we are showing our proposed approach.

A. Input Data Set

The data set has 2266 text comments, which are collected from youtube.com. This data set has Kannada script text, English script text, Numbers and some Symbols. To identify the code-switch text, Kannada and English script is helpful. Numbers and symbols will not make any impact to identify code-switch text.

B. Pre-Processing

The collected raw data has some punctuations, numbers, and symbols, which will not play any important role in identifying code-switch text. We need to remove all these punctuations, numbers, and symbols. We had done this task by using NLTK toolkit and also we split each comment into tokens.

C. Feature Extraction

We can't apply any supervised classification technique directly on the text data. So we should transform text data into numerical features before applying any machine learning algorithm. This process is called feature extraction. We had extracted two features called Bag-of-Words (BOW), and Term Frequency and Inverse Document Frequency (TF-IDF) by using sklearn.feature_extraction model.

Collection of all the words in a document irrespective of their order and meaning is called Bag-of-Words. The Document Term Matrix is formed with the frequency of each word in the document. We applied TF-IDF to improve the word frequency feature. Reducing the weight for words that are common in the document can be done by using IDF to improve the performance.

Finally we extracted 4129 features or columns from the data set.

D. Supervised Classification Techniques

After feature extraction, we had applied supervised classification techniques such as "Logistic Regression, Support Vector Machine, Random Forest, Neural Network, and Naïve Bayes".

- Logistic Regression – It is generally used where the dependent variable (target variable) is binary or dichotomy. The dependent variable takes only two values i.e, "yes" or "no", "0 or 1". Independent variables are categorical or numerical [24].
- Support Vector Machine – SVM is used for both classification and regression. SVM will search for a hyper plane that will separate the two classes in the dataset.
- Random Forest – It is applied to machine learning that develops large number of decision trees to analyse set of variables. It can help to drill down into data and provide more sophisticated analysis. Decision trees are the building blocks of random forest model.
- Neural Network – It is a set of algorithms that are designed to recognize patterns. The patterns they recognize are numerical, vectors, real-world data, image, sound, text, time series etc. It helps in clustering and classification problems.
- Naïve Bayes – It is a classification technique based on Bayes' theorem with an assumption of independence among the predictors. In simple terms, assume that the presence one feature in a class is unrelated to the presence of any other feature. This

is useful for large data sets and it is easy to calculate the posterior probability.

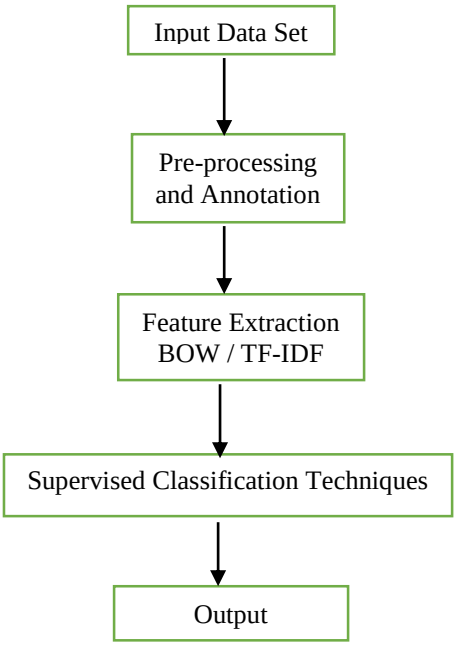

FIG.1: PROPOSED APPROACH

E. Output

The final output will have code-switch text and plain text, which are identified by supervised classification methods. In our data set, we have 1263 code-switch text comments and 1003 plain text comments.

V. RESULTS DISCUSSION

After pre-processing and feature extraction, the data is ready for applying supervised classification techniques. Data set is split into training data and testing data. We have used 10-fold cross validation method for splitting the data set into 70% as train data and 30% as test data. Cross validation is a technique which is used for the assessment of how the results of statistical analysis generalize to an independent data set. We applied different supervised classification techniques to obtain results.

- Confusion Matrix – is a pivot table of the TP, TN, FP and FN of the scored data set. This is the second metric to validate the model performance after accuracy.
- Precision – is the proportion of the TP with the total predicted as positive (TP+FP). It is desirable to tune the models to get precision near to 1.

$$precision = \frac{TP}{TP + FP} \quad (1)$$

- Recall – is the metric used to evaluate the coverage of the TP. It is given by:

$$recall = \frac{TP}{TP + FN} \quad (2)$$

Like precision, it is desirable to have the recall value near to 1.0.

- F1-Score – This is a harmonic mean of precision and recall metric. Closer the numerator and denominator, better will be the model.

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \quad (3)$$

- Support – It is the number of actual occurrences of the class in the specified data set. Support doesn't change between models but instead diagnoses the evaluation process.

TABLE I, shows the results of precision, recall, F1 – score and support of different supervised classification techniques such as "Logistic Regression, SVM, Random Forest, Neural Network, and Naïve Bayes". We had obtained precision, recall, f1 – score, and support for code – switch text and plain text.

Support is same for logistic regression, SVM, Neural Network, and Naïve Bayes supervised classification techniques, i.e., code – switch support is 381 and plain text support is 299. For Random Forest the support is different, i.e., code-switch support is 427 and plain text support is 253.

TABLE I. RESULTS OBTAINED FOR CODE-SWITCH (CS) AND PLAIN TEXT (PT)

CLASSIFICATION TECHNIQUE	CS / PT	PRECISION	RECALL	F1-SCORE	SUPPORT
LOGISTIC REGRESSION	CS	0.82	0.90	0.86	381
	PT	0.85	0.74	0.79	299
S V M	CS	0.82	0.74	0.78	381
	PT	0.71	0.79	0.75	299
RANDOM FOREST	CS	0.88	0.79	0.83	427
	PT	0.70	0.82	0.75	253
NEURAL NETWORK	CS	0.80	0.88	0.84	381
	PT	0.82	0.72	0.77	299
NAÏVE BAYES	CS	0.83	0.88	0.85	381
	PT	0.85	0.74	0.79	299

The confusion matrix is obtained for different supervised classification techniques. Both Logistic Regression and Naïve Bayes correctly classifying 565 out of 680 samples. SVM correctly classifying 519 out of 680 samples. Random Forest correctly classifying 544 out of 680 samples. Neural Network correctly classifying 549 out of 680 samples.

TABLE II. CLASSIFICATION ACCURACY

CLASSIFICATION TECHNIQUE	CLASSIFICATION ACCURACY
LOGISTIC REGRESSION	83
NAÏVE BAYES	83
NEURAL NETWORK	81
RANDOM FOREST	80
SVM	76

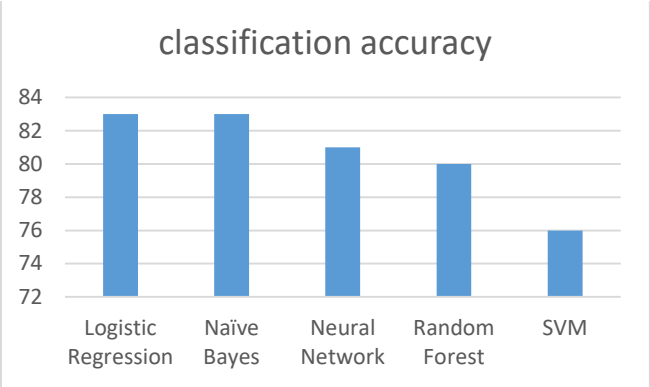

FIG. 2. CLASSIFICATION ACCURACY

TABLE II, shows the classification accuracy of different supervised classification techniques. Logistic Regression and Naïve Bayes both are showing 83% classification accuracy. Neural Network (81%), Random Forest (80%) are close to Logistic Regression and Naïve Bayes.

FIG. 2, shows that SVM (76%) classification technique is showing least classification accuracy when compare with other classification techniques.

VI. CONCLUSIONS

Popular usage of social media, the code-switch is very common for communication in social media in order to express views or opinions. People are very comfortable to use code-switch text in text comments. These code-switch text may have emotion words and these emotion words are important for sentiment analysis.

In this paper, we tried to identify code-switch text and plain text by using supervised classification techniques such as "Logistic Regression, SVM, Random Forest, Neural Network, and Naïve Bayes". Logistic Regression and Naïve Bayes are more accurate than other supervised classification techniques which are followed by Neural Network and Random Forest. SVM is least accurate compare with other supervised classification techniques. In future work, this work can be extended for detecting emotions in English-Kannada code-switch text, and it can be extended further to sentiment analysis.

REFERENCES

[1] S. Lee, S. Li and C. Huang, " annotating events in an emotion corpus," Processing of the 9th international conference Lang.Res.Eval.,2014.

[2] H. Liu, S. Li, G. Zhou, C. Huang and P. Li, "joint Modeling of News Reader's and comment Writer's Emotions," in Processing of the 51st Annual Meeting of the Association for Computational Linguistics, Pages 511 – 515.

[3] B. Pang, L. Lee and S. Vaithyanathan, " Thumbs Up? Sentiment Classification using Machanic Learing Techniques," in Processing of Empirical Methods for Natural Language Proces, 2002, PP.79-86.

[4] J. Blom and J. Gumperz, "Socila Meaning in Linguistic Structures: Code Switching in Nothern Norway," in directions in Socialinguistics. New York, NY, USA: Winston, 1772.

[5] H. Giles and R. Clair, " Language and Social Psychology," London, UK.:1979

[6] C. Myers-Scotton, "Duelling Language: Grammatical Structure in Code-Switching," Oxford, UK.: Clarendon, 1997.

[7] P. Auer, " Code- Switching in Conversation, " Evanston, IL, USA: Routledge, 1999.

[8] H. Adel, N. Vu, K. Kirchhoff, D. Telaar, and T. Scgultz, "Syntactic Features for code- switching Factored language models," IEEE/ACM Transactions and Audio, Speech, and Language process, vol23, no.3, pp.431 -440, March 2015.

[9] D. Garrette, H. Alpert-Abrams, T. Berg-kirkpatrik, and D. Klein, " Unsupervised Code-Switching for Multilingual Historical Document Transaction," in processing of Annual Conference of the North America Chapter association for Computational Linguistic, 2015,pp.1036 -1041.

[10] W. Ling, G. Xiang, C. Dyer, A. Black, and I.Trancoso, "Microblogs as Parallel Corpora," in processing of 51st Annula Meeting of the Association for computational Linguistic,2013.

[11] T. Solorio and Y. Liu, "Learning to Predict code-Switching Points" in proceedings of Emprirical Methods fr Natural Language Process, 2008, PP.973-981.

[12] C. Lignos and M. Marcus, " Toward Web- Scale Analysis of code-Switching, " in Proceedings of Annual Meeting of the Linguistic Society America, 2013.

[13] Y. Li and P. Fung, " Code-Swich Language Model with Inversion constraints for Mixed Language Speech Recognition," in Proceedings of the International Conference of the Computational Linguistic,2012, PP.1671-1680.

[14] A. Jamatia, B. Gamback, and A. Das, " Part-of-Speech Tagging for Code-Mixed English Hindi Twitter and Facebook Chat Messages," in Proceedings of Recent Advanced Natural Language Process, 2015, PP.239-248.

[15] D. Burkett and D. Klein, "Two Languages are better than one (for syntactic parsing)," in Proceedings of the confrence Empirical Methods for Natural Language Process, 2008, PP.877-886.

[16] W. Gao, J. Blizer, M. Zhou, and K. Wong, " Exploiting Bilingual Information to improve Web search," in proceedings of 47th Annual Meeting of the Association for Computational Linguistic / 4th International Joint Conference of Natral Language Process, AFNAP,2009, PP.1075-1083.

[17] M. Amini, C. Goutte, and N. Usunier, " Combining Coregularization and Consensusbased Seld – Training for Multilingual Text Categorization," in proceedings of 33rd International ACM SIGIR Conference Res. Develop. Inf.Retrieval, 2010, PP.475-482.

[18] K. Bali, J. Sharma, M. Choudhury, and Y. Vyas, " I Am Borrowing Ya Mixing? An Analysis of English – Hindi Code Mixing in Facebook," in Proceedings of the First Workshop on Computational Approaches to Code – Switching, 2014, pages 116-126.

[19] Y. Vyas, S. Gella, J. Sharma, K. Bali, and M. Choudhury, "POS Tagging of English – Hindi Code – Mixed Social Media Content," in proceedings of the 2014 conference on Empirical Methods in Natural Language Processing(EMNLP), Pages 974-979.

[20] A Sharma, S. Gupta, R. Motlani, P. Bansal, M. Srivastava, R. Mamidi, and D. M. Sharma, " Shallow Parsinf Pipeline for Hindi-Englis Code – Mixed Social Media Text," 2016 arXiv preprint arXiv : 1604.03136.

[21] U. Barman, A. Das, J. Wagner, and J. Foster, " Code Mxing a: A Challenge for Language Identification in the language of Social Media," in proceedings of the First Workshop on Computational Approaches to Code Switching, 2014, pages 13-23.

[22] K. C. Raghavi, M. k. Chininakotla, and M. Shrivastava, " Answer Ka Type Kya He?: Learning to classify Questions in Code- Mixed Language," in proceedings of the 24th International Conference on World Wide Web, 2015, pages 853-858, ACM.

[23] S. Lakshmi B S and Shambhavi B R, "An Automatic Language Identification System for Code-Mixed Englisg-Kannada Socila Media Text," 2nd IEEE International Conference on Computational System and Information Technology for Sustainable Solutions, 2017.

[24] Ratnakar Pandey. (2017, November 4). Logistic-Regression-Using-Scikit-Python. Retrieved July 30, 2019, from http://datafai.com