A Language Identification System Using CRF-based Approach for North-East Indian Regional Social Media Text

[1]Priyadarshini Lamabam, [2]Kunal Chakma
[1]Post Graduate Computer Science Teacher, [2]Assistant Professor
[1]Computer Science Department,
[1]Kendriya Vidhyalaya No.1 Imphal, Imphal, India

Abstract : ***Identification of the languages at the document level has been considered an almost solved problem in some application areas, but language detectors fail to perform well in the social media context due to phenomena such as utterance internal code-switching, lexical borrowings and phonetic typing. In such an environment, automatic language identification for the code-mixed Social Media Texts has captured attention from the Natural Language Processing Research community. We describe our Conditional Random Field(CRF)–based system for automatic language identification of social media content of code-mixed English and Manipuri texts. A dataset of Twitter and Facebook posts that exhibit code-mixing between English and Manipuri was selected. Experimentation on CRF models was done using various features and the performances have been observed.***

IndexTerms – Natural Language Processing,code-mixed,CRF,trigrams,bigrams.

I.INTRODUCTION

Natural Language Processing(NLP) domain considers the text processing as the core research problem. The NLP research community have already developed many state-of-the-art technologies for processing formal texts written in most popular languages. With the proliferation of Internet and Social Media sites such as Facebook, Twitter etc., a new category of content has been generated by the users which is often referred as Social Media Content (SMC). The type of texts generated on Social Media is informal in nature. They are often bi-lingual and in some cases tri-lingual in nature where mixing of two or more than two languages are prevalent.Mixing of languages is known as Code Mixing or Code Switching .Therefore, the performance of the state-of-the-art text processing techniques/technologies prove to be a failure on such informal texts. As a result such informal texts have introduced new research challenges in the NLP domain.

Many studies have been carried on the topic of code-mixing and code-switching in conversations for several years. Bilingualism is commonly practiced in many countries, but it has not been linguistically studied in detail in computer-mediated communication, especially in the field of social media. For any kind of automatic text processing system, it is very essential to be able to identify the language from a specific segment of text. After the various investigations on language identification and computational analysis of code switching for several years, only few works on automatic language identification for multilingual code-mixed texts has been reported. This is because the available language detectors fail in the context of social media texts due to the phenomena of utterance internal code-switching, lexical borrowings(borrowings of words from other language mainly English) and phonetic typing (typing of other languages using Roman alphabet). Therefore, automatic language identification for the code-mixed social Media texts has become one of the most important and challenging task in the field of Natural Language Processing.

Many investigations were done to find out why code-mixing occurs in social media. Some studies showed that linguistic motivations influence the people for code-mixing in highly bilingual societies[1]. However, in the area of social media, code-mixing often takes place at the message beginnings or through simple insertions among the texts, and also to mark in-group membership in short text messages[2], chat messages[3], Twitter and Facebook posts[4], comments and emails.

Speakers whose first language uses non-Roman alphabet, hardly use their native script while expressing their ideas and expressions in social media. They used transliterated text where their script is converted into Roman script for convenience. This increases the code-mixing likelihood. Such cases are found In South-east Asia and India. People in India, especially the urban population frequently insert English words or phrases through Anglicism or code-mixing to express their thoughts when they speak. When posting texts on social media, Indians often use their native language mixed with English.So, all these reasons contribute to the code-mixing phenomenon in social media and the dominance of English is receding though it is still the most popular language. The following is an example of code-mixed post from twitter.

Tweet: Friends with benefits. *Sina* **best** *tare* **.** *Aduna loi tinnarase* **lol .** *Fagi twbni ko* .
Translation: Friends with benefits. This is the best. So let us all make friends lol. Just kidding.

This post is written in two languages: English (**bold**) and Manipuri (italic). Different types of language mixing phenomena have been discussed and defined. We described the datasets to investigate the code-mixing between English and Manipuri. The corpus collected for this task is selected from Twitter and Facebook posts that exhibit code mixing between English and Manipuri.Manipuri belongs to Tibeto-Burman Language and is one of the scheduled languages in the Indian Constitution. The resources of this language is very less which leads to few linguistically ecplored works on the formal documents as reported in [5].

As far as code-mixing is concerned, no significant system has been designed which can automatically identify the code-mixed English-Manipuri social media text.Therefore, research in this language processing will help the Non-Manipuri speaking people to understand the

language and bring it up to the global platform. Some recent works on code-mixing in English, Hindi and Bengali languages have been reported in [6][7] so following their footsteps our objective is to develop a language identification system for code-mixed English-Manipuri Social Media text.

The paper is organized as follows: section II introduces the description of code-mixing and the few previous works on language identification of some popular languages. The code-mixed corpus acquisition process is described in section III. Section IV andV discusses about the experimentation and the analysis of the results. Finally, the paper is concluded in section 6 with the future highlightment.

II.RELATED WORK

Since 1980s, code-mixing or code-switching has been recognized as a byproduct of two or more languages. The various investigations has been done on language identification for half a century [8] and that of computational analysis of code switching for several decades [9]. Still it is found that only few works on automatic language identification for multilingual code-mixed texts have been reported. Several researchers tried to find out the reasons of code-mixing.They have reported that Linguistic motivations for sociological and conversational importance influence the people in highly bilingual societies for code-mixing[10].Descriptions about inter-sentential, intrasentential and intra-word code mixing were done according to the researchers. Researchers in [11], showed that facebookers tend to mainly use inter-sentential switching (59%) over intra-sentential (33%) and tag switching (8%), and 45% for real lexical needs, 40% in talking about a particular topic, and 5% by content clarification.

For automatic language identification, few previous works on Hindi, Bengali, English, etc., have been done but none of them worked on code-mixed Manipuri with English.An initial study on automatic language identification with the dataset of Bengali-Hindi-English Facebook comments was presented in [6]. They have used systems such as dictionaries, Support Vector Machine (SVM) and Conditional Random Field (CRF). Lack of transliterated dictionary for Bengali and Hindi and their phonetically-typed nature, make use of the training set words as dictionaries. Dictionaries like British National Corpus (BNC), LexNormList and SemEvalTwitter along with training set words are used for English. Experimentations on SVM and CRF were done with various features such as char-n grams,presence in dictionaries, length of words and capitalization.

In [7] the researchers have presented some different techniques for the identification of English-Hindi and English-Bengali language mixed in Facebook posts using character n-grams, dictionaries and Support Vector Machine classifiers. N-gram modeling experimentation was performed on the training data for n = (1, 2, 3, 4, 5, 6, 7).For English, a lexical normalization dictionary prepared for Twitter was used. The Samsad English-Bengali dictionary [12] and the Bengali lexicon transliterated into Romanized text using the Modified Joint-SourceChannel mode [13] was used for training Bengali language. Incorporation of features like N-gram with weights, dictionary-based,Minimum Edit Distance(MED)-based weight and word context information were done for SVM.

Different datasets of Nepali-english and Spanish-English were taken and language classification experiments were done using dictionary-based method,linear kernel Support Vector Machines(SVMs) and a k-nearest neighbor approach as reported in [14].The British National Corpus, lexical normalization dictionary and the training set words are used as dictionaries.Their SVM system, uses char n-grams, dictionary-based labels, length of words, capitalization and contextual clues as features.In [15], a CRF-based system for language identification of four language pairs namely, English-Spanish, English-Nepali, English-Mandarin and Standard Arabic-Arabic Dialects. It uses lexical, contextual, character n-gram and special character features, and therefore, replication can be easily done across the languages.

In [16], the analysis on english-hindi code mixing from facebook has been reported by the researchers. They created the corpus from the facebook pages of some popular public figures and also from BBC news corpus.They annotated the data using matrix,normalization,word origin,Named entities and POS tagging. Their analysis has shown a significant amount of Code Mixing of English in Hindi matrix and Hindi in English matrix.Hindi words embed in English using formulaic patterns of Nouns and Particles while English language get mixed with Hindi at various forms ranging from single words to multi-word phrases. They have also indicated that code-mixing in social media needs a deeper analysis of structural and discourse linguistics.

The researchers in [17] have tested the limits of the existing word level language identification systems such as linguini[18], polyglot[19], langid.py[20] and Compact Language Detector2 (CLD2)[21]. They have prepared a synthetic code-mixed dataset of 28 languages. Due to lesser accuracy of the previous systems,they extended the existing algorithms to Random, MaxWeighted, CoverSet and Optimal. Random algorithm assigns a randomly chosen label to a word from a possible set of labels which are obtained by setting a threshold value on the confidence scores of the classifiers. Maxweighted assigns the label of the classifier with the highest confidence. CoverSet assumes that code-mixing happens only with a few language though there is no restriction in the number of languages. optimal algorithm compute the set of possible labels based on a threshold value.If the actual(gold standard) label of a word belongs to the set,then it is assigned as the tag for the word. The extended algorithms outperformed than the existing algorithms significantly.

The researchers in [22] have developed two twitter language identification systems for tweets consisting of nine languages and written in three non-latin scripts.The two systems are logistic regression classifier (LogR) and prediction by partial matching (PPM). LogR uses the character features and meta features. For PPM, with the given training data,it tries to minimise crossentropy and choose the language that would compactly encode the text to classify.The language identification systems discussed above were not available for testing on our data so no comparisions could be reported.

Manipuri, is a regional language and there is lack of linguistically studied resources so only few NLP tools have been developed earlier for the formal documents. Its resource is very less in comparison to the languages such as English, Hindi, Chinese, Korean, etc.. the tools are Part of Speech tagger [23] developed using hand written linguistic rules and affix stripping method. Another part of speech tagger is developed using CRF in [24]. Name entity recognition systems have been developed by [25] from CRF using features from manual assumption. Another name entity recognizer is also developed from SVM in [26].As of now, no significant research on Manipuri social media texts has been reported.

III.CORPUS ACQUISITION

For Natural Language Processing (NLP), Corpus acquisition is the most important requirement. Without the data collection, no NLP work can be started.The higher the number of data, the better is the NLP task.

We started our data collection from Twitter through twitter4j[27] API. We started searching tweets with different query terms consisting of popular Manipuri terms, stop words and phrases in Roman transliterated form. After several attempts, we observed that English-Manipuri code-mixed tweets were not retrieved as the tweets were obtained from different languages but not in Manipuri.There could be several possible reasons for which the API could not retrieve the expected tweets. A possible reason could be due to the Roman transliterated form of the query terms,the same word could possibly belong to some other language besides Manipuri. Another possible reason is that the people who speak the Manipuri language mostly tweet in either monolingual English or Manipuri only. The third reason is the smaller size of the Manipuri population as compared to the people who speak other Indian languages. Moreover, twitter API does not allow to fetch data older than 7 days.

Therefore, the tweets are collected manually from popular Manipuri twitter pages such as *Kangla, Leima Photography* where we find code-mixing is frequent. Our initial collection after removing retweets stands at 2000 out of which only 700 tweets are code-mixed. As a further approach, two Facebook confession groups (*HRD Confessions Ghari* and *Herbert School, Imphal-Confessions*) were selected to obtain the publicly available posts of 778 and 522 respectively giving a total of 1300 code-mixed facebook posts. Unlike Twitter, the Facebook posts are quite long as they have no limitations in the number of characters.The data was collected using Facebook graph API explorer. Our final corpus consists of 2000 code-mixed posts where 700 are from Twitter
and 1300 from Facebook.

A. Tokenization

The whole code-mixed corpus of 2000 has been tokenized by CMU Tokenizer[28]. Some instances are found where the CMU Tokenizer failed to tokenize the words such as

- Words followed/preceded by symbols. Eg. *Nungsibiradi],Luhongba/,*etc.
- Symbols between the words. Eg. *Manipur/kangleipak, cheiraoba-Bangkok,*etc.

Such cases exist even after tokenization due to the noisiness of the social media text,but no pre-processing has been adopted. The posts are kept in the original state as they are tokenized.

Table 1: Annotation Tagset

Tags	**Description**
en	English word
mni	Manipuri word
univ	Universal word
acro	Acronym
acro_mni	Acronym + Manipuri suffix
ne	Named entity
ne_mni	Named entity + Manipuri suffix
mixd	English+Manipuri suffix or Manipuri + English suffix
undef	Undefined or other language

B. Annotation and its agreement

The tokenized Twitter and Facebook posts are randomly shuffled to mix the facebook and twitter posts. The first 1400 is selected as the training data and the remaining 600 as the test data. The training data is manually annotated with the tags as given in Table 1. No previous work has been done on such texts, so no standarad tagset was available for use. Therefore, we finalized some guidelines and accordingly the tagging HAS been done.

- English and Manipuri tokens are tagged with *en* and *mni* respectively.
- The named entities like name of the persons ,locations,organisatios,language,religion,community are tagged with *ne* tag
- The named entities if followed by Manipuri are respectively tagged with *ne_mni* tag.
- Acronyms like HRD,RIMS are tagged with *acro* tag and *acro_mni* is for those acronyms with Manipuri suffix.
- English tokens followed by Manipuri suffix and vice versa are tagged with *mixd* tag.
- A univ tag is attached to a word or a token if
  -All characters of the token are either symbols or numbers.
  -It contains certain repetitions characterised by regular expressions.
  (eg., *hahaha,hehe,lol*,etc.)
  -It is a an URL or a hashtag or mention-tags (eg., *@ManipuriSMS*)
- A word or a token is considered as undef if
  -It is of other language besides English and Manipuri
  -It is followed/preceded by symbols due to inability of the tokenizer (eg., *Nungsibiradi],clz12/E,* etc.)
  -It is formed by symbols between two unrelated words. (eg., *Manipur/kangleipak,,Cheiraoba-Bangkok,*etc.)

-It is unrelated combination of two tokens (eg., *khara-A*,etc.)

Two annotators were involved and an Inter-annotator agreement of 95.52% was measured on randomly selected 100 posts with a giving a kappa[29] value of 0.9552.

C. Data characteristics

The characteristics of our data set is revealed by the word-level statistics which is given in Table 2. The total number of tokens of our dataset is 37648. We have 54% of the total tokens from Manipuri. English tokens contribute to 18% and we have 4% as named entities. The remaining is contributed from the other categories.

The different nature of the social media text is proven by the examples given below

- creative spellings (eg., *2sn* for tuition, *clz* for college)
- word play (eg., *albuuuumz* for album, *killllll* for kill)
- abbreviations (eg., *OMG, SMS*)

Another interesting characteristic is that some words (eg., d,c,to,a,up,) share the same surface but they have different meanings in English and Manipuri.This is due to the phonetic similarity of English and Manipuri. We find that the numbers are attached to the end of the words if they are to be written more than one time, indicating the number of words (eg.,*pareng2,Fajabne2,little2,bye2*,etc.)

Table 2: Word-level statistics

Tags	Count
en	6832
mni	20461
univ	7475
mixd	94
undef	621
acro	216
acro_mni	2
ne	1878
ne_mni	69

D. **Code-Mixing Types**

The various types of code-mixing obtained from our corpus are given below. Bold-face indicates English segments and italics Manipuri. The corresponding translation in English is given after each post.

- ***Inter-sentential code-mixing***:If switching of language occurs outside the sentence.
  E.g.: #840 **I lov u** donna maibam... *eina nggi nafamda tinnari kanda fongdok hwdrae*...
  #840 I lov u donna maibam... I have not proposed you when we were friends...
- ***Intra-sentential code-mixing***:If switching of language occurs inside the sentence or clause boundaries.
  E.g.: #1225 **confession** se 2006 tagi hourammadi fadoune... waiii chaini **confession** na... 2008 **pass out**
  #1225 it would have been better if the confession were started by 2006... there would have been lot of confession... 2008 pass out
- ***Word-level code-mixing***:If switching of language occurs inside a word.
  E.g.: #1180 2006-08 ki akhoina 11 karakpg **section G** gi **group** fotodo leiradi **upload** amta twbirko
  #1180 if the group photo of 11 section G 2006-08 is available please upload

Here, "fotodo" is a word-level code-mixing where "foto" stands for "photo" and "do" is a Manipuri suffix.

IV.EXPERIMENTAL WORK

Considering the dataset of 2000,shuffling has been done for mixing the 700 twitter and the 1300 facebook posts and comments.Conditional Random Fields(CRFs) belong to statistical modeling methods. They are often used in pattern recognition and machine learning arears for structured prediction. An ordinary classifier can predict a label for a single sample without considering the neighboring samples but CRF can consider context into account for better predictions.

CRF models are implemented using Miralium[30] , a machine learning toolkit. Both the training and the testing files need to be in a particular format for working with CRF. They consist of multiple tokens where each token contains a fixed number of columns separated by tab. A template file is used where the features are defined.

A bootstrapping process is adopted for CRF and the performance of the CRF models are evaluated in each iteration. The model building process continues until the model stabilizes.

The various crf models are built iteratively using the following feature sets.

1. The first three and the last three characters

The experimentation starts with the initial 1400 posts for building the classification model.It is tested with the next 200 untagged posts. The models are built iteratively and the performance of the models are evaluated by computing the weighted precision, recall and F1-score. The results of the various iterations of the models are given in Table 3 and the individual measures of the various tags in Table 4.

Table 3: Iterations of the CRF-based models with the first feature

Iterations	Weighted F1-score
1	0.910
2	0.913
3	0.912

Table 4: Individual Tag measures of the first feature

Tags	Precision	Recall	F_1-score
en	0.905	0.870	0.887
mni	0.943	0.952	0.947
univ	0.989	0.961	0.975
ne	0.496	0.640	0.559
ne_mni	0.000	0.000	0.000
acro	0.952	0.667	0.784
acro_mni	0.000	0.000	0.000
mixd	0.500	0.444	0.471
undef	0.792	0.585	0.673

2. The previous and the next token

The second feature we took is the previous and the next token for every current token. If there is next token but no previous token then the current token itself acts as the previous token.If there is previous token but no next token,the current token acts as the next token.Using this logic we prepare the crf models and the iterations are presented in Table 5
and the individual measures in Table 6.

. Table 5: Iterations of the CRF-based models with the second feature

Iterations	Weighted F1-score
1	0.835
2	0.84
3	0.837

Table 6: Individual Tag measures of the second feature

Tags	Precision	Recall	F_1-score
en	0.834	0.788	0.810
mni	0.893	0.878	0.885
univ	0.916	0.863	0.888
ne	0.349	0.621	0.447
ne_mni	0.000	0.000	0.000
acro	0.944	0.567	0.708
acro_mni	0.000	0.000	0.000
mixd	0.500	0.333	0.400
undef	0.696	0.492	0.577

3. Character bigrams

The third feature taken is the sequence of every two characters. Eg.,for the word ”school”,the character bigrams are sc,ch,ho,oo and ol. The various iterations are presented in Table 7 and the individual measures in Table 8.

Table 7: Iterations of the CRF-based models with the third feature

Iterations	Weighted F1-score
1	0.916
2	0.916
3	0.912

Table 8: Individual Tag measures of the third feature

Tags	Precision	Recall	F_1-score
en	0.869	0.893	0.881
mni	0.929	0.972	0.950
univ	0.979	0.970	0.974
ne	0.772	0.517	0.619
ne_mni	0.000	0.000	0.000
acro	0.917	0.733	0.815
acro_mni	0.000	0.000	0.000
mixd	0.600	0.333	0.429
undef	0.829	0.523	0.642

4. Character Trigrams

The third feature taken is the sequence of every three characters. Eg., for the word ”school” the character trigrams are sch,cho,hoo and ool. The various iterations are presented in Table 9 and the individual measures in Table 10.

Table 9: Iterations of the CRF-based models with the fourth feature

Iterations	Weighted F1-score
1	0.906
2	0.912
3	0.906

Table 10: Individual Tag measures of the fourth feature

Tags	Precision	Recall	F_1-score
en	0.897	0.864	0.880
mni	0.948	0.951	0.949
univ	0.984	0.960	0.972
ne	0.491	0.665	0.565
ne_mni	0.000	0.000	0.000
acro	0.952	0.667	0.784
acro_mni	0.000	0.000	0.000
mixd	0.444	0.444	0.444
undef	0.795	0.538	0.642

5. Combination of the features

In this case,all the features are combined and the various iterations are presented in Table 11 and the individual measures in Table 12.

Table 11: Iterations of the CRF-based models with the combined feature

Iterations	Weighted F1-score
1	0.916
2	0.919
3	0.914

Table 12: Individual Tag measures of the combined feature

Tags	Precision	Recall	F_1-score
en	0.839	0.944	0.888
mni	0.946	0.954	0.950

univ	0.984	0.963	0.974
ne	0.861	0.581	0.694
ne_mni	0.000	0.000	0.000
acro	0.950	0.633	0.760
acro_mni	0.000	0.000	0.000
mixd	1.000	0.444	0.615
undef	0.521	0.569	0.544

Table 13: Comparisions of the CRF-based Models

Features	**Weighted F1-score**
A1	0.840
A2	0.913
A3	0.914
A4	0.912
A5	**0.919**

Where,
A1=previous and next token
A2=first three and last three characters
A3=character bigrams
A4=character trigrams
A5=combination of all the features

From the various experimentations we find that the CRF model built using bigrams has given the highest F1-score of 0.916 and combination of all the features has given a higher F1-score of 0.919.

V.RESULT AND ERROR ANALYSIS

We have built the various CRF models using different features. With the previous and the last token as the feature,the F1 scores of English and Manipuri are 0.810 and 0.885 which is not upto the mark. With the inclusion of other features, there is a significant improvement in the weighted F1-scores. We find that the CRF model built using the character bigrams has given the highest F1-score of 0.916. The English and the Manipuri tokens have given an increased F1-scores of 0.881 and 0.950. The remaining tokens like acronyms have also shown an improvement. And now, with the combination of all the features, the highest weighted F1-score is achieved i.e.,0.919. Therefore,we can establish the fact that the CRF model built using the combination of all the features is the system that can be used for the language identification of the code-mixed social media posts.

For calculating the performance of the various models, weighted F1-scores have been used.This is because the distribution of the different types of tokens is uneven. The percentage of the Manipuri tokens is high as compared to English and universal tokens. The remaining tags contribute to lesser percentage. So, the individual precision, recall and F1-scores of the developed models can be biased towards some tags as compared to others. Therefore, to have a better evaluation of the models we have computed the weighted F1- scores. This is calculated by weighting the measure (precision, recall and F1-score) of an individual class by the total number of instances of that class in the training data. In one of the iterations of CRF-based models, we find that the F1-measure of Manipuri tokens (0.950) is higher than English tokens (0.881). Therefore, computing the weighted F1-measure gives a more realistic measure of the system which is 0.916

A. Statistical Significance

Suppose that we have generated two classification models, M_1 and M_2 from our data.We have performed 10-fold cross-validation to obtain a mean error rate for each. To find out the better classifier, a statistical significance test is done to prove that the two classifier models are statistically significant. Null hypothesis states that the two models are the same. So if we can reject the null hypothesis,we can prove that the two models are statistically significant and choose the one with lower error rate. Using any number of cross-validation we can obtain

$$\overline{err}(M_1) \quad (1)$$

and

$$\overline{err}(M_2) \quad (2)$$

These mean error rates are just estimates of error on the true population of future data cases. Assuming the samples follow a t-distribution with k-1 degrees of freedom,to reject null hypothesis, a test is done that computes t-statistic with k-1 degrees of freedom. The formula for t-test is

$$t = \frac{\overline{err}(M_1) - \overline{err}(M_2)}{\sqrt{var(M_1 - M_2)/k}} \quad (3)$$

Where

$$\text{var}(M_1\text{-}M_2) = \sum_{i=1}^{k} [err(M_1)_i - err(M_2)_i - (\overline{err}(M_1) - \overline{err}(M_2))]^2 \quad (4)$$

A significance level is choosen eg., sig = 0.05 or 5% means M_1 and M_2 are significantly different for 95% of population. Confidence limits for our error estimates are obtained as z=sig/2. Ift <z or t >-z, then t value lies in rejection region and we can reject null hypothesis that mean error rates of M_1 and M_2 are same and conclude that statistically significant difference between M_1 and M_2. Otherwise, conclude that any difference is chance. Now among the four CRF models, the bigram-based and the first and last three character based models are selected to test for statistical significance since they give higher F1-scores than the remaining models. We have 3 iterations and the error rate of each iterations are observed and average them to calculate the mean error rates of the two models. The value of k is 3 and the t-test is calculated giving the value as -37.75 which is less than z i.e, -0.025 that shows that the two models are statistically significant. They accept with each other for 95 % of the population

VI.CONCLUSION AND FUTURE WORK

Social media revolution has created a challenge to the language processing system designed for formal texts It becomes more difficult when the corpus is confined to the regional languages like Manipuri. We have presented an initial study on automatic language identification of code-mixed English and Manipuri from Twitter and Facebook posts. After the experimentation on CRF-based models,it is found that character bigrams feature of CRF has given an F1-score of 0.916 and combination of the four features has given the highest F1-score of 0.919

Obtaining code-mixed texts for a regional language like Manipuri was quite difficult so we have worked only on a small datset containing combination of Twitter and Facebook posts. Due to the smaller size of the dataset, there is else diversity in the nature of the data. Therefore,we could easily obtain the high F1-measure in CRF without the employment of variety of features. We plan to extend our dataset in the future and explore more number of features. These techniques can also be applied to other datasets and observe the performance.

REFERENCES

[1] D. C. S. Li., Cantonese-english code-switching research in hong kong: a y2k review

[2] S. Sotillo., Ehhhh utede hacen plane sin mi???:@ im feeling left out:(form, function and type of code switching in sms texting., ICAME 33 Corpora at the centre and crossroads of English linguistics (2012) 309–310.

[3] Z. Bock., Cyber socialising: Emerging genres and registers of intimacy among young south african students., Language Matters: Studies in the Languages of Africa 44(2) (2013) 68–91.

[4] L. A. Shafie, S. Nayan., Languages, code-switching practice and primary functions of face book among university students., Study in English Language Teaching.

[5] S. S. R. K, S. Gunasekaran, A. K. M, K. P. Soman, A Short Review about Manipuri Language Processing 3 (3) (2014) 99–103.

[6] U. Barman, A. Das, J. Wagner, J. Foster, Code Mixing : A Challenge for Language Identification in the Language of Social Media, Proceedings of The First Workshop on Computational Approaches to Code Switching (2014) (2014) 13 –23.

[7] A. Das, Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text.

[8] E. M. Gold., Language identification in the limit.

[9] A. K. Joshi., Processing of sentences with intra-sentential code- switching, proceedings of the 9th International conference on Computational linguistics,ACL 145–150.

[10] P. Auer., Code-switching in conversation:language, interaction and identity.

[11] T. Hidayat, AN ANALYSIS OF CODE SWITCHING USED BY FACEBOOKER (a Case Study in a Social Network Site).

[12] S. Bisvas., Samsad bengali-english dictionary sahitya samsad 3.

[13] A. Das, T. Saikh, T. Mondal, A. Ekbal, S. Bandyopadhyay, English to Indian Languages Machine Transliteration System at NEWS 2010 (July) (2010) 71–75.

[14] U. Barman, J. Wagner, G. Chrupaa, J. Foster, DCU-UVT : Word-Level Language Classification with Code-Mixed Data.

[15] G. Chittaranjan, Word-level Language Identification using CRF : Code-switching Shared Task Report of MSR India System (2014) 73–79.

[16] J. kalika, Sharma, I am borrowing ya mixing ? An Analysis of English-Hindi Code Mixing in Facebook (2014) 116–126.

[17] S. Gella, K. Bali, M. Choudhury, Testing the Limits of Word level Language Identification.

[18] J. M. Prager., Linguini: Language identification for multilingual documents, proceedings of the 32nd Annual Hawaii International Conference IEEE. (1999) 11pp.

[**19**] M. Lui, T. Baldwin., Cross-domain feature selection for language identification proceedings of 5th International Joint Conference on Natural Language Processing.

[**20**] langid.py., https://github.com/saffsd/langid.py.

[**21**] Compact language detector2, https://code.google.com/p/cld2/.

[22] S. Bergsma, P. Mcnamee, M. Bagdouri, C. Fink, T. Wilson, Language Identification for Creating Language-Specific Twitter Collections (Lsm 2012) (2013) 65–74.

[23] B. S. P. Kh Raju Singha, K. D. Singha., Part of speech tagging in manipuri: A rule-based approach, proceedings of IJCA 51(14).

[24] A. E. Thoudam Doren Singh, S. Bandyopadhyay., Manipuri pos tagging using crf and svm: A language independent approach.

[25] T. S. B. S. C. Kishorjit Nongmeikakpam, Leisram Newton Singh, S. Bandyopadhyay., Crf based name entity recognition in manipuri: A highly agglutinative indian language, proceedings of 8th International Conference on Natural Language, IIT Kharagpur.

[26] T. D. Singh, K. Nongmeikapam, A. Ekbal, S. Bandyopadhyay, Named Entity Recognition for Manipuri Using Support Vector Machine (2009) 811–818.

[27] Twitter api., http://twitter4j.org/en/.

[28] O. Owoputi, B. O. Connor, C. Dyer, K. Gimpel, N. Schneider, N. A. Smith, Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters.

[29] The kappa statistic, https://en.wikipedia.org/wiki/Cohens+kappa.

[30] Miralium, https://code.google.com/p/miralium.