Sentiment Orientation from Code-mixed Social Media Data

Kavita Asnani[1] and Floyd Avina Fernandes[2]

[1]State Higher Education Council, Directorate of Higher Education, Porvorim, Goa.
[1]Adjunct Faculty, Department of Computer Science, Goa College of Arts, Science and Commerce,Sanquelim, Goa.
[2]Conaug Technology, Goa.

Collecting and evaluating data is becoming an effectively admissible challenge in the highly connected world. In the 21st century, with the advent of social networks getting popular, the social media information is getting archived at alarming rates. The use of local language in informal fashion is very common on social media platform. In natural language processing, Sentiment Analysis (SA) is a specialized process of determining user orientation from opinion data floating on social web. Code-mixed social media data in specific is challenging to process, due to mixing of varied languages used to portray the linguistic efficiency. In this paper, we propose a model called Code-mixed Sentiment Analyzer (cmSentiAnalyzer) to derive sentiment orientation from code-mixed sentences. Our proposed model has used language features across code-mixed languages to map the words occurring in different languages to a common space. Our experiments reveal that cmSentiAnalyzer outperforms baseline approaches in sentiment analysis for code-mixed text by 2% in accuracy and 89% of average precision.

Keywords: Code-Mixed social media text, Sentiment Analysis, Pointwise Mutual Information (PMI), Sentiment Classification

1. INTRODUCTION

Code-mixing (CM) refers to mixing of two or more languages together Thara and Poornachandran [2018]. In multilingual countries, multiple languages exist and are used for daily communication. Social media sites generate large volumes of data which contribute to deriving useful patterns. The influence of local languages is seen to be dominating in such data. With the use of multiple languages in code-mixed social media data, there is high probability of words occurring in multiple scripts. Natural Language Processing(NLP) researchers have been addressing various aspects in the domain of code-mixed data like language identication, Parts of Speech (POS) tagging and Named Entity Recognition(NER)Chittaranjan et al. [2014] Vyas et al. [2014] Sequiera et al. [2015] Solorio et al. [2014] Rao and Devi [2016]. Such NLP tasks have an intermediate requirement of transliteration of data written in roman script and native script. Globally, people communicate in different languages; such as parts of Europe, Africa and South East Asia Jamatia et al. [2019]. Hindi and English are designated as the official languages by Government of India and therefore people have prior knowledge of Hindi/English language in addition to their mother tongue Raghavi et al. [2015]. It is seen that spoken languages where code-mixing has been common in psycho- and socio-linguists exists for half a century; but the research work on computational processing of code-mixed text has been done in the early 1980s Joshi [1982]. Based on the switch of language, code-mixing has inter-sentential and intra sentential code switching Thara and Poornachandran [2018]. In inter-sentential switching, language switch happens at sentence boundaries. For example “Kya hua.. anybody called? ”. The hindi context (‘kya ’ ‘hua ’) is switched to English (‘anybody ’and ‘called ’). In intra-sentential code switching, the shift is done within a sentence, with no pauses to illustrate a shift. For example “Woh mera friend hai ”i.e. hindi is blended with english Thara and Poornachandran [2018].

In this paper, we have made an attempt to perform sentiment analysis on code-mixed data for hindi-english code-mixed text. Our proposed model has been referred to as Code-mixed Sentiment Analyzer (cmSentiAnalyzer). It works at token level where each token is mapped to the relevant language dictionary and further sentiment classification is done to derive sentiment orientation.

In Section 2 the related work is presented. In Section 3, the cmSentiAnalyzer model is presented and Section 4 reports experiments and results. Section 5 presents conclusion and future work.

2. RELATED WORK

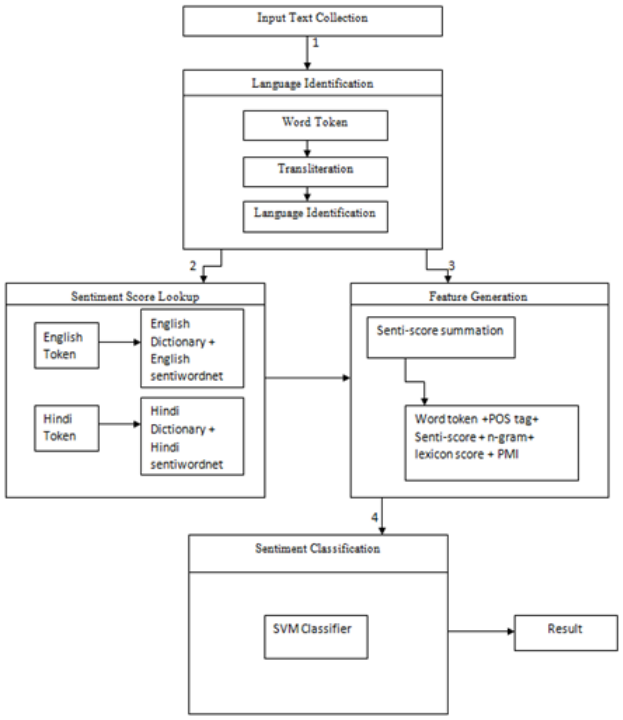
One of the core aspects of research in code-mixing has been sentiment analysis. Pravalika et al. [2017] devised an approach for sentiment analysis of domain specific code-mixed social network data with two approaches based on the lexicon and machine learning methods. Support Vector Machine (SVM) classier has been used to perform classication on the training data which obtained accuracy of 86% for lexicon based approach, while 72% for the machine learning approach Pravalika et al. [2017]. Analysis on word level language identification model with input of English and Hindi code-mixed sentences with different features like n-gram, distance and POS tags were applied to the respective languages Malgaonkar et al. [2017]. A scoring algorithm is used in Malgaonkar et al. [2017] to assign the score to the words and subsequently recorded the accuracy of 92.68% and 91.72% for positive and negative cases respectively. . A survey on multiple languages with different dialects has been done in Bhargava et al. [2016]. The machine learning techniques for language specic SentiWordNet with n-grams features were used with various classiers Bhargava et al. [2016]. Word level language identication in social media text have been addressed using dictionary mapping approach. SVM classifier have been used to perform classication on features like n-gram pruning and dictionary modules. Sentiment analysis on Hindi-Marathi code-mixed languages has been performed in Ansari and Govilkar [2018]. Language identification has been done by extricating slang present in the text. To get the original text transliteration has been performed followed by POS tagging of the tokens generated and thus obtained tokens have been mapped to the respective dictionaries for computing senti-score which resulted in the features which subsequently have been classified using Naive bayes and SVM Ansari and Govilkar [2018]. Analysis of code-mixed data for Hindi and English have been done in two phases namely, language identification and judging the sentiment Sharma et al. [2015a]. In language identification, each word is identified and tagged as /H or //E for respective language and transliteration of Hindi words have been done to get the original text Sharma et al. [2015a]. Sentiments have been judged by mapping words from WordNet and applying POS tagger Sharma et al. [2015a]. The transliteration based precision was 85% while precision based on sentiment obtained was 80%.

Subjective graph-based lexicon for Hindi language which was dependent on WordNet have been proposed in Arora [2013]. A list of seedwords has been built and expanded using WordNet obtained synonym and antonym relations Arora [2013]. Each word in the seed list have been considered as node and have been connected to their synonym and antonym. The accuracy reported for classication of reviews showed 74% and with human annotators it was found to be 69%. POS tagging on social media content for English-Hindi bilinguals have been proposed in Vyas et al. [2014].

The common language related challenges in the domain of code-mixing are identified as transliteration, slang and lack of annotated data Vyas et al. [2014]. The results demonstrate the computation of co-ocurrence score of two words based on semantic orientation which have been used to compute text orientation as positive, negative and neutral.

3. SYSTEM MODEL

This research paper proposes cmSentiAnalyzer model which aims to determine the sentiment orientation of code-mixed text. The proposed architecture of cmSentiAnalyzer is shown in Figure 1. It comprises of four main modules: (i) Data preprocessing module. (ii) Sentiment score lookup module. (iii) Feature Generation module (iv) Sentiment classification.



. Figure 1. Architecture of cmSentiAnalyzer

FIRE 2013 [1], an annotated code-mixed social media text dataset is used; where each token is associated with its respective language. Annotated FIRE 2013 dataset provides social media data with each token providing the associated language tag label. FIRE 2013 labelled dataset is extracted. Named entities are taken into consideration. Words labeled as /H or /E are looked-up and separated for compilation.

Example 1 shows the sample annotated text and Example 2 shows an example of the positive sentence.

The main modules of our proposed cmSentiAnalyzer model which contribute to computation of semantic orientation are presented in Section 3.1, Section 3.2 and Section 3.3.

3.1 Sentiment Score Lookup

cmSentiAnalyzer follows look-up approach for language assignment to each token. WordNet is used to analyze the existence of a word and if the word exists then it displays the synonyms with

[1]http://cse.iitkgp.ac.in/resgrp/cnerg/qa/fire13translit/FIRE-Data/HindiEnglish-FIRE2013-AnnotatedDev.txt

Christian\E dharma\H =धर्म ke\H = के vivah\H= विवाह vidhi\H =विधि
Gayatri\H = गायत्री mantra\H= मंत्र english\E translation\E
Tere\H = तेरे bin\H=बिन nai\H =नई lagta\H = लगता dil\H= दिल mera\H= मेरा dolna\H= डोलना music\E release \E date\E
Telangana\E par\H= पर congress\E vivaad\H=विवाद
Hawa\H= हवा mahal\H = महल tourist\E guide\E
Ek\H= एक ahnabi\H= अजनबी sa\H= सा ehsaas\H = एहसास dil\H= दिल ko\H= को sataye\H=सताए movie\E name\E
Barkha\H= बरखा mein\H= में bhi\H= भी dil\H=दिल pyasa\H= प्यासा hain\H= हैं song\E release\E

. Example 1. Sample from Annotated Dataset

synset ids. Similarly Hindi token is viewed in the Hindi WordNet. Hindi WordNet of IITB [2] have been used for Hindi word look-up. For each English word, after reviewing through the English WordNet, the SentiWordNet is used. The English SentiWordNet provides the senti scores of English words. Similarly, for Hindi words, Hindi SentiWordNet is used. It consists of words in Devanagari script having positive and negative decimal scoring. Once the positive and negative score is obtained final score is calculated. The final score assigns the final sentiment to the entire document in terms of positive or negative.

3.2 Feature Generation

Features are extracted based on senti-score summation. POS tagging, n-grams, lexicon score and Pointwise Mutual Information (PMI) are used for senti-score summation.

3.2.1 *POS Tagging:.* Part of Speech tagging in language processing is used to assign a peculiar tag to a word such as verb, adjective, adverb etc. cmSentiAnalyzer have considered four POS tags namely noun, adjective, adverb and verb. POS tagging matrix have been created to store the values.

3.2.2 *N-grams:.* The n-grams are obtained from the input text and the feature vector is constructed from them. The n-gram representations are used to train the data file for n-gram generation.

3.3 Lexicon Score

The lexicon score is computed using feature list for English and Hindi word representation which includes values of positive score, negative score, total positive count, total negative count, max positive and max negative. Lexicon score matrix was created to store the values.

3.4 PMI based Scoring

PMI has been used to compute the weight of each word in the training set. Each word contains a value for each possible class namely positive and negative. The look- up approach has been used as the dictionary maintains a count of every word for positive and negative labelled sentence. It

[2]http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php

counts, the number of times the word has occurred in respective labelled category. It computes the count of all positive words in the positive sentiment while the negative words in negative sentiment. Thus obtained feature comprise of feature vectors for the given code-mixed text.

3.5 Sentiment Classification

The Support Vector Machine (SVM) classifier has been applied for performing sentiment classification on the feature vectors.

The experimental results are presented in Section 4.

4. EXPERIMENTAL RESULTS

The dataset used for the evaluation of the proposed cmSentiAnalyzer model comprises of 500 documents. For each token of the document text, English and Hindi transliterated form is obtained and accordingly it is augmented with the language identification tag. As referred to in the Figure 1 and Section 3, our proposed cmSentiAnalyzer is implemented. We used look-up approach as referred to in Section 3.1, to map each token to the respective WordNet and SentiWordNet language digital dictionaries. Thus obtained sentiscore is used for positive or negative sentiment classification of the document. We implemented morphological feature assignment for each language token for POS tagging as referred to in Section 3.2.1. Stanford POS tagger has been used for tagging of English tokens and shallow parser has been used for tagging Hindi words. All the values obtained have been stored in POS matrix for further evaluation. As referred to in Section 3.2.2, the n-gram representation is implemented using unigrams and bigrams. The updated feature vector is appended with lexicon score value which is computed, using lexicon matrix attributes namely positive score, negative score, total positive count, total negative count, max positives and max negatives. The order of token occurrence as referred to in Section 3.2.4, is used for computing the PMI score and accordingly the feature vector is updated. The proposed cmSentiAnalyzer model is designed to work in the supervised settings using the SVM classifier. The updated feature vectors comprise the vector base which is spilt into training and testing.

Table 1 shows the working of the cmSentiAnalyzer model.

Reference	Dataset	Approach	Code-mixed	Task	Performance
Baseline model [1]	FIRE 2013 FIRE 2014 Facebook Youtube	Lexicon approach	English-Hindi language	Sentiment Analysis	Precision: 80%
cmsentianalyzer model	FIRE 2013	Machine Learning (Classification approach)	English-Hindi language	Sentiment Analysis	Precision: 89% Accuracy:82 %

. Table 1. cmSentiAnalyzer Performance Comparison

The evaluation carried out on code-mixed data used 450 sentences for training and 50 sentences for testing. A two class confusion matrix has been constructed for accuracy and average precision as shown in Table 2.

```
[[ 0  9]
 [ 0 41]]
```

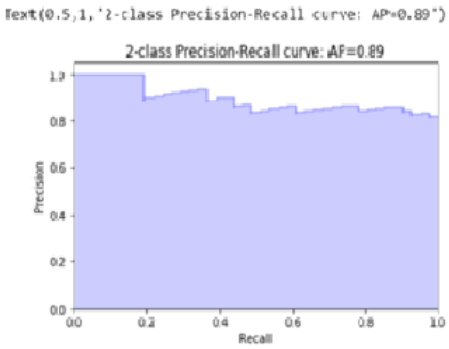
. Table 2. cmSentiAnalyzer: Confusion Matrix

The accuracy and average precision was calculated using Equation 1 and Equation 2 respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$AP = a_0 + \sum_n (R_n - R_{n-1}) \tag{2}$$

We evaluated the results against baseline model Sharma et al. [2015b] and the comparison is shown in Table 1. The cmsentianalyzer model outperforms the baseline as it recorded the accuracy of 82% which contributed to 2% rise and an average precision of 89% in judging the sentiment of code mixed text. Figure.2 shows the Precision-Recall Curve.



. Figure 2. Precision-Recall Curve

5. CONCLUSION AND FUTURE WORK

This research paper presents a novel framework based on model performance baselining to effectively determine the sentiment orientation from code-mixed text. ThecmSentiAnalyzer model presented in this paper, addresses the problem by using linguistic features which represents random occurrences of words in multiple languages into a common space.In order to derive sentiments in a precise manner, feature generation is done using POS tags, n-grams and PMI scoring. These features are further directed to the SVM classifier for classification. It concludes that cmSentiAnalyzer performs better than the baseline model as the accuracy obtained is 82% and average precision is 89% and precision-recall curve is obtained. We observe a curve ranging from 80% to 99% of test values to predictions probabilities. This system will enable a better way in understanding and analysing code-mixed social media data in user/customer conversations.

As future work, the main focus will be to extend the dataset by incorporating various information sources so that the supervised machine learning algorithm i.e. TF-IDF vectorization can be incorporated as it weights down the common words thereby giving higher weightage for unique

words. It would be legitimate to experiment with other languages and other types of social media text, such as tweets, comments on Youtube and Instagram. Building upon this study, one can delve into new techniques in various tasks

REFERENCES

Ansari, M. A. and Govilkar, S. 2018. Sentiment analysis of mixed code for the transliterated hindi and marathi texts. *International Journal on Natural Language Computing (IJNLC) Vol 7*.

Arora, P. 2013. Sentiment analysis for hindi language. *MS by Research in Computer Science*.

Bhargava, R., Sharma, Y., and Sharma, S. 2016. Sentiment analysis for mixed script indic sentences. In *2016 International conference on advances in computing, communications and informatics (ICACCI)*. IEEE, 524–529.

Chittaranjan, G., Vyas, Y., Bali, K., and Choudhury, M. 2014. Word-level language identification using crf: Code-switching shared task report of msr india system. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*. 73–79.

Jamatia, A., Das, A., and Gambäck, B. 2019. Deep learning-based language identification in english-hindi-bengali code-mixed social media corpora. *Journal of Intelligent Systems 28,* 3, 399–408.

Joshi, A. 1982. Processing of sentences with intra-sentential code-switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*.

Malgaonkar, S., Khan, A., and Vichare, A. 2017. Mixed bilingual social media analytics: case study: Live twitter data. In *2017 international conference on advances in computing, communications and informatics (ICACCI)*. IEEE, 1407–1412.

Pravalika, A., Oza, V., Meghana, N., and Kamath, S. S. 2017. Domain-specific sentiment analysis approaches for code-mixed social network data. In *2017 8th international conference on computing, communication and networking technologies (ICCCNT)*. IEEE, 1–6.

Raghavi, K. C., Chinnakotla, M. K., and Shrivastava, M. 2015. " answer ka type kya he?" learning to classify questions in code-mixed language. In *Proceedings of the 24th International Conference on World Wide Web*. 853–858.

Rao, P. R. and Devi, S. L. 2016. Cmee-il: Code mix entity extraction in indian languages from social media text@ fire 2016-an overview. In *FIRE (Working Notes)*. 289–295.

Sequiera, R., Choudhury, M., Gupta, P., Rosso, P., Kumar, S., Banerjee, S., Naskar, S. K., Bandyopadhyay, S., Chittaranjan, G., Das, A., et al. 2015. Overview of fire-2015 shared task on mixed script information retrieval. In *FIRE workshops*. Vol. 1587. 19–25.

Sharma, S., Srinivas, P., and Balabantaray, R. C. 2015a. Sentiment analysis of code-mix script. In *2015 international conference on computing and network communications (CoCoNet)*. IEEE, 530–534.

Sharma, S., Srinivas, P., and Balabantaray, R. C. 2015b. Text normalization of code mix and sentiment analysis. In *2015 international conference on advances in computing, communications and informatics (ICACCI)*. IEEE, 1468–1473.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*. 62–72.

Thara, S. and Poornachandran, P. 2018. Code-mixing: A brief survey. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2382–2388.

Vyas, Y., Gella, S., Sharma, J., Bali, K., and Choudhury, M. 2014. Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 974–979.

Dr. Kavita Asnani is working as an Associate Professor in Teaching, Learning and Educational Technology at State Higher Education Council Directorate of Higher Education, Goa. She is also an Adjunct Faculty at Department of Computer Science, Goa College of Arts, Science and Commerce, Sanquelim Goa She has completed her graduation in B.E.(Computer Engineering) from Shivaji University Masters in Engineering(IT) from Goa University and was awarded PhD in Computer Science and Technology in December 2018 from Goa University. She has 19 years of teaching experience in engineering college at UG and PG level. She has guided students at M.E. and B.E. projects. Her research areas are Data Mining and Natural Language Processing(NLP). She was a member of the development project entitled "Rural Health Education Project", CIDA/ACCC Project # 703Z, a five year Canadian College Partnership Program funded by Canadian International Development Agency (CIDA). She is Principal Investigator for some Research Projects. He is a member of the ACM, IEEE and life member of the CSI.

Ms Floyd Avina Fernandes is an UI /UX Designer in Conaug Technology Pvt Ltd. She has completed her M.E. in Computer Science and Engineering from Goa College of Engineering. She has completed her Masters under the guidance of Dr. Kavita Asnani. She has two years experience in the industry. She has worked on research projects like Parts of Speech Tagging for Indic Languages, Sentiment Analysis during her M.E. Program. Her research interests are Machine Learning, Natural Language Processing and Sentiment Analysis on Indic Languages.