Emotion Detection using Online Machine Learning Method and TLBO on Mixed Script

Shashank Sharma, PYKL Srinivas, Rakesh Chandra Balabantaray
IIIT Bhubaneswar
Odisha
E-mail: sohamshashank@gmail.com, a114011@iiit-bh.ac.in, rakesh@iiit-bh.ac.in

Abstract

Due to rapid modernization of our societies, most people, if not all, have access to online social media and mobile communication devices. These people hail from diverse cultures and ethnicity, and interact with each other more often on these social media sites. Moreover, due to their distinct backgrounds, they all have an influence on the common language in which they communicate. Also, many users employ a myriad of shorthand, emoticons and abbreviations in their statements to reduce their effort. This calls for a means to assist in better communications through social media.

In our work, we have researched on understanding the underlying emotions and sentiments of these interactions and communications. Our focus was on analyzing the conversations by Indians in the code-mix of English and Hindi languages and identifying the usage patterns of various words and parts of speech. We have categorized statements into 6 groups based on emotions and improved the model using TLBO technique and online learning algorithms. These features were integrated in our application to assist the mobile device users in quickly sort and prioritize their messages based on the emotions attached with the statements and provide much more immersive communications with their friends and family.

Keywords: code mix, mixed script, emotions, TLBO, online machine learning

1. Introduction

Today social media has become a one stop solution for all information needs, whether it's about chatting with friends or growing your professional network or delivery of news content. We are all connected on social media websites. As our connectivity expands, our network of friends and acquaintances is growing beyond boundaries. We interact with a mixture of people with different roots who are bilingual or even multilingual. In these kinds of scenarios, people tend to mix two or more languages while interacting with others. This mixing of two languages happens when both people who are communicating are not experts in a common language, thus they tend to mix a few words from one language to another to interpret or complete the conversation.

This kind of scenario is mostly seen when the mother tongue of both the persons interacting is different and neither one of them is fluent in their national language. For example, if we have two people from India, one's mother tongue is Hindi and the other's is Marathi, during a conversation in Hindi, the Marathi person may mix a few common Marathi terms in his speech. The person mixes terms from another language when he is trying to use a very rarely used or complicated word and may doubt that the other person would be unable to understand it in the base language, and thus he replaces it with a more commonly used word from another language.

We find many instances of such usage where there is mixture of two languages being written in Roman script on social media sites. This is referred as Code mixing or Code Switching or Mixed Script. Few authors have differentiated code mixing or mixed script and switching but for ease we have used these terms interchangeably. Linguists have explored a lot of different reasons and the frequency of mixing two languages. In this paper we have worked specifically on analyzing and understanding the emotions within the mixture of two languages i.e. English and Hindi, written in Roman script.

Commonly, we found that the sentences in Indian mixed script usually had few terms from the associate official language i.e. English, but the grammar rules that are followed were from the base language (Hindi, Marathi, Bengali etc.)

Example: *"ye awesome nahi hai !!"*

Here, the word "awesome" is in English language, but this statement has been written by a Hindi speaker which follows the grammatical rules from Hindi language. So it is better to convert this sentence in pure Hindi language for natural language processing.

The next most notable characteristic which was found on Indian social media was the mother tongue influence (MTI). If a Bihari, Marathi, Bengali, or Malayali person speaks in Hindi language, the pronunciation of the words has a MTI (Pal, 2013) from their mother tongue language. This effect was also seen on social media content, e.g. a word 'bhi' in Hindi was found written as 'bi', 'vi', 'bee', 've' formats, and these different spellings variations can be referred as creative spellings.

We even find news articles being presented in code mixed format, to show the importance of the word or drag focus of readers to that point. When a mono-language parser or interpreter tries to understand the sentiment of such text, the unidentified words are left out as those are not parts of the base language. This leads to reduced analysis of the text as a whole.

Interpreting a language is essential since we need to communicate, understand, translate, answer questions and even to retrieve information from web if a person doesn't know the word in international auxiliary language.

Nowadays every person is connected using various social networking sites, and receives a lot of messages every minute. A lot of text is flowing into our phones. We need to process the natural language to prioritizing our messages based on the content and the sender. So we have designed an app where users can prioritize their messages which are being received on the phone. Also, our app can read messages from different social media sites and notify the user by blinking the screen with the set of emoticon (Happy :), Surprise :O, Sad :(, Angry :@, Fear :'(and Neutral :|) along with corresponding colors.

Our work is focused on language identification (LI) and POS tagging of mixed script. We have also tried to improve the language development process and detection of emotions in mixed script by combining machine learning and human knowledge. Though this mixed script phenomenon has been recognized by linguistics over 40 years ago, we don't find a strong (large) linguistic resource for Hindi language in Roman or in Devanagari script. Another aspect of standard dictionaries is that they can't be used for analysis on social media sites as it doesn't include internet slang words. So, we have taken reviews from the user about his/her creative spellings using our mobile app. The other interesting notification features we have included is that the app reads the messages received from various social media sites, prioritizes them and notifies the user by displaying an emoticon/emoji on the screen based on the emotion of the message received from user. These implementation procedures have been discussed below.

In the next section we have discussed the related work from this field and section 3 describes the way we have handled various issues and implemented the algorithm.

2. Related Work

The topology of code switching has been analyzed by many of the linguistics around 50 years back, where they have studied how functional and linguistic factors affect code switching behavior. Also, according to a survey, inarticulate bilingual speakers were able to code switch without grammatical errors (Poplack, 1980). Most of the people think that code switching is a random event but according to Lance (1975), it is rule governed. Code switching may be used to achieve interaction effects during communication (Gumperz, 1971, 1976; Valdes falls, 1978). We do agree that code switching is an indicator of degree of bilingual proficiency. Also, code switching was identified as one of the modes of communication by Pedro Pedraza(1978). Code mixing and code switching has been analyzed from structural, psycholinguistic and sociolinguistic dimension (Muysken, 2001; Senaratne, 2009).

For language identification on mixed script, most of the researchers have used Conditional Random Fields (CRF) based model. Chittaranjan et al.(2014) experimented CRF on 4 language pairs. CRF (Lafferty et al., 2001) is a probabilistic framework for labeling and segmenting structured data, such as sequences, trees and used for assigning labels to a set of observation sequence.

Language identification task involves language modeling and classification. Dunning (1994) was the first to try character n gram models for language identification. Different machine learning approaches can be used for classification techniques like support vector machines (Kruengkrai et al., 2005), normalized dot product (Damashek, 1995), k-nearest neighbor and relative entropy (Sibun and Reynar, 1996) have also been used for language identification.

In Indian social media, the mixed script is a mixture of English-mother tongue language, but these are written in Roman script. So the words from one language are written using a scripting language of other, this phenomenon is known as Phonetic typing. We need to transliterate these words into one language. Most of the transliteration systems were designed to make foreign language e.g. English or national language (Hindi) readable to all. So these literatures were transliterated to various Indian languages so that people who cannot communicate or read English could understand the literature in their mother tongue language like Marathi, Bengali, Tamil etc. We came across a lot of relevant work in transliteration. English to Devanagari script transliteration was performed (Aggarwal, 2009) by using Statistical Machine Translation Tool known as Moses. In our work, we have transliterated English language written in Roman script to Devanagari script.

We have used an approach to find the base language of the speaker and have translated mixed from another language to base language. A rule based system known as AnglaHindi was been designed (Sinha and Jain, 2003) to translate from English to other Indian languages.

Parts of speech are very useful for judging the sentiment of the sentence. One of the recent works was done using Maximum Entropy Markov Model to tag POS for Hindi language (Dalal et al., 2006), where multiple features are used to predict the tag for a word. Gradable adjectives (Hatzivassiloglou and Wiebe, 2000) such as 'extremely' which are a part of Adjectives play an important role in subjective languages.

A huge amount of work has been done on developing lexicon dictionary for sentiment analysis on English language. SentiStrength (SO-CAL) is a set of lexicons where each word is given a score ranging from -5 to +5, where -5 responds to most negative emotion word and +5 responds to most positive word. This list was created by human coders. A semi-automated technique has been used to construct a lexicon list (Whitelaw et al., 2005) where every lexicon has 5 attributes describing each word.

Hindi SentiWordNet (Joshi et al, 2010) has been developed by translating English SentiWordNet. Words not found in the HSWN are searched with closest meaning words from synset to judge the polarity of the sentence (Pooja and Sharvari, 2015) but instead of entirely depending on the WordNet we have taken help from users to judge unclassified statements.

3. Our Approach

Our primary objective was to detect the underlying emotions within mixed texts written in roman script. This process involves various steps. The mixed script had to be preprocessed for identifying the emotion/s in the sentence. Preprocessing involved identifying the language of each word in the mixed script to discern base language of the speaker, so that we could apply the same parts of speech (POS) tagger to understand the grammar of the language better. The better we know the structure of the language; the better would be its interpretation and response. Correct grammar could also make the translation from one language to another precise. So, firstly we had to identify the language of every word in the mixed script. We have used CRF (Conditional Random Field) which is a probabilistic model for labeling sequential data. In CRF, each feature is a function that takes in as input: a sentence (s), the position (i) of a word in the sentence, the label (l_i) of the current word, the label (l_{i-1}) of the previous word and outputs a real-valued number (the numbers are often either 0 or 1). CRF model (referred as M_1) is trained with a huge dictionary of English words, and for Hindi words we used a dictionary by "IIT Kgp" which had 30,823 transliterated Hindi words (Roman script) followed by the same word in Devanagari and also contains Roman spelling variations for the same Hindi word. We have used the same Hindi word list (Gupta et al., 2012) as a dictionary to identify language and also for getting the right transliteration pair. Though our dictionary consisted of 2 lakh words and 31,000 words approximately from English and Hindi language respectively, we were unable to identify the language of all the words in the mixed script sentence, as dictionary based approach is not exhaustive and secondly, a lot of chat acronyms & text shorthand have spawned a new language on internet. Also, these set of words cannot be found in any standard language dictionaries. These text shorthand notations are written in roman script and are in English language. These notations have been referred to as characteristics of mixed script found on Indian social media (Sharma et al., 2015) and have been categorized into phonetic typing, short forms, word play, and slang words. Again, we had to adopt a dictionary/rule based approach – to detect and correct these creative spellings. We trained our CRF model with a huge list of 5000 creative spellings and slang words of English language to detect and correct the words for POS tagging.

Even after applying CRF model with dictionary of words, few words were again left unrecognized due to limitations with dictionary based approach. To identify the language of these words we firstly tried to identify the base language of the speaker. Base language can be considered as the mother tongue of the speaker or the first language in which the user likes to initiate the conversation.

Ex 1.*"ye bahut important hai bro :@ !!"*

Here, the base language of the speaker is Hindi, where two words are in English (i.e. 'important' and 'bro').

The base language is guessed by various factors such as the number of words from the base language used in the sentence, language of the starting word, language of the prepositions or stop words etc. The first advantage of identifying base language is to approximate the correct language's part of speech tagger to be applied. Secondly, there were few words which belonged to both the languages and were wrongly tagged as English, in the first phase of LI. Ex 2. *"mujhe ye item banana hai"*. In this sentence, the word 'banana' is a Hindi as well as English word. But this would be tagged as English in the first phase of LI if we presume this word is not present in our Hindi dictionary. By identifying the base language we recheck ambiguous words and tag 'banana' as Hindi word. In this way we improve our LI model. For identifying ambiguous words, we created a list of common words from English and Hindi language and then guessed the language of these words with respect to base language. This way of identifying language gave much better accuracy than window based approach (Sharma et al., 2015).

Now, as we have identified the language of each mixed script we need to translate the entire sentence to the base language. By knowing the base language of the statement, we get to know which language's grammatical rules the statement follows. Considering example statement 1, as the base language is Hindi, we need to transliterate Hindi words from Roman script to Devanagari script i.e. converting WX (a transliteration scheme for representing Indian languages in ASCII) to UTF (the universal character code standard to represent characters) notation, and pure English words need to be translated from English to Hindi using Shabdanjali dictionary. The example statement 1 after following the above rules, become: "ये बहुत महत्वपूर्ण है भाई :@ !!"

It has been argued that we could skip LI and directly translate the text into English language. But if we blindly translate every mixed script to pure English language, we may miss the context of the sentence, as has been validated by most of the people who use Google translate, which has an accuracy of 57% in translating text (Patil & Davis, 2014).

The next step towards emotion detection was to tag the statements with accurate parts of speech tagger based on the base language. We identified that adjectives and adverbs express positive or negative orientations and verbs and nouns are used to express opinions. For example: 'dislike' and 'love' are verbs and 'hero' and 'villain' are nouns. We need to understand the lexical category or word class or POS of the language to recognize the emotions attached with the sentence better.

According to a study, researchers got very low accuracy in tagging POS of a code mix script. Instead of using a probabilistic based approach in judging the language of mixed script which depends on the preceding language of the word or chunking words belonging to the same language for POS tagging, we have used a standardized POS tagger based on the identified base language. Stanford POS tagger is used for English and Sivareddy's POS tagger for Hindi language is used.

We have used multi class SVM and multinomial logistic regression (M_2) based approach to detect the emotions of the sentence. As there is no pre annotated dataset available in code mix format having the corresponding emotions assigned, we have used dataset released by "FIRE 2014 Shared Task on Transliterated Search" and few posts were manually collected from various websites like Facebook and Youtube. To make our app capable in judging the emotions of statement written in pure English language, we have also considered a dataset with 4000 statements categorized into 6 different emotions. The mixed script statements collected did not have corresponding emotion attached with it. So, we have manually tagged these statements into 6 categories. These 6 categories of emotions are: Happy, Surprise, Sad, Angry, Fear and Neutral. We have also segregated smileys into 6 categories and incorporated in our model to improve the emotion detection in the statements.

We have considered 300 mixed script statements which were manually tagged and the model was built. As these statements were not so broad and filled with emotions, many statements were categorized as Neutral. So, we have used a bootstrapping based approach where different lexical and semantic relations between the Hindi words and English words are considered from Hindi WordNet and English WordNet respectively to correlate similar words and push into the above emotions category which has reduced neutral statements and also helped us to expand our static Hindi dictionary to some extent.

As our model was totally based on the lexicon dictionary which is even small in size, we integrated our application to take reviews from the user for unclassified sentences by using TLBO technique and learnt the model online using logistic regression. Teaching learning based optimization (TLBO) technique has been used to achieve a global optimum solution from different users' reviews. TLBO is a population-based iterative learning algorithm for large scale non-linear optimization problems for finding the global solutions. The TLBO (Rao et al., 2011) process is divided into teaching phase and learning phase, where teacher influence the output of learners in the class. The teacher is considered the most intelligent person who shared his or her knowledge with learners and capability of the teacher affects the outcome of the learners. Teacher tries to distribute knowledge among learners which in turn increases intelligence level of whole class.

In our problem, statements those emotions were judged or statements which cannot be judged by our lexicon dictionary are considered as learners and users are considered as teachers to train the model, by tagging unidentified emotions of statements in our scenario. Correct emotion category which will be assigned by the teacher (i.e. user) is the outcome of this technique. The model efficiency is improved by two methods: firstly by learning among learners, which is similar to supervised learning based approach (M_2) and secondly by the teacher (user). User tries to improve the model by assigning emotions to unclassified statements, which in turn increases model capabilities. This approach uses mean value of the population to update the solution, where the opinions of the users are considered to get the global optimum. TLBO technique does not require any parameters for tuning and it is easy to implement.

For every unclassified statement user is prompted to select the right category to which a statement belongs. The categories are Happy, Surprise, Sad, Angry, Fear and Neutral. These statements which are judged by the user by using TLBO technique act as train set for updating our cloud based model using online/ incremental logistic regression based learning technique.

Online machine learning based algorithm is suitable in this situation as the model needs to be updated each time it gets a review from the user. This review acts as a new training instance for the model. By using this approach our model is always updated by considering recent history and we are able to create a repository of pre-annotated statements with their corresponding polarity.

4. Results

To test our model, we have divided our dataset of 300 mixed script statement into train and test set, where 200 statements were randomly selected as train set and 100 as test set. Multinomial logistic regression and multi class SVM algorithm was modeled on this dataset and the results were compared from human annotated emotions of statements. We achieved a precision of 0.74 by multinomial logistic regression and 0.70 by SVM multi class classifier. Our model may be improved incrementally by training more statements in each category of emotions. To implement the same, we have provided option to the user of the app to tag uncategorized statements to emotions which will improve our model and create a broader dataset of mixed script statements with their corresponding polarity.

5. Conclusion

In this paper, we have described the capabilities of our app which can read messages received from various social chats from different senders and prioritize them and notify the emotion attached to the message by displaying that kind of smiley and colors on the screen. The users can then understand the significance of the message received and if interested, they can go ahead and read it or ignore. We have used online machine learning based approach to handle instant update of the model to give results dynamically. During this process, we were able to identify the language of ambiguous words which were common in Hindi and English and tag lexical category or parts of speech in mixed script by identifying the base language of the speaker. We can create a language resource of mixed script statements with their corresponding polarity by using TLBO technique.

6. Acknowledgements

We would like to show our gratitude to our parents for sharing their pearls of wisdom with us during the course of this research.

We thank Arjun Roy Choudhury for his comments and valuable feedback that greatly improved the manuscript.

7. Bibliographical References

Aggarwal, A., (2009). “Transliteration involving English and Hindi languages using syllabification approach”, *In Doctoral dissertation, Indian Institute of Technology, Bombay Mumbai.*

Canasai Kruengkrai, Prapass Srichaivattana, Virach Sornlertlamvanich, and Hitoshi Isahara. (2005). “Language identification based on string kernels”, *In Proceedings of the 5th International Symposium on Communications and Information Technologies (ISCIT- 2005),* pp. 896--899, Beijing, China

Chamindi Dilkushi Senaratne, (2009), “Sinhala-English code-mixing in Sri Lanka: A sociolinguistic study”, chapter Code-mixing as a research topic. *LOT Publications.*

Chittaranjan, G., Vyas, Y., & Choudhury, K. B. M. (2014, October). “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*, pp. 73--79.

Dalal, A., Nagaraj, K., Sawant, U. and Shelke, S., (2006). “Hindi part-of-speech tagging and chunking: A maximum entropy approach”, *In Proceeding of the NLPAI Machine Learning Competition*.

Gumperz, J. J. (1971). “Bilingualism, bidialectalism and classroom interaction”, *In Language in Social Groups*. Stanford: Stanford. University Press.

Gumperz, J. J. (1976). “The sociolinguistic significance of conversational code-switching”, *Working Papers of the Language Behavior Research Laboratory*. 46. Baerkeley: University of California.

Joshi, Aditya, A. R. Balamurali, and Pushpak Bhattacharyya, (2010) "A fall-back strategy for sentiment analysis in hindi: a case study." *In Proceedings of the 8th ICON.*

Joachims, T. (1998). “Text categorization with support vector machines: Learning with many relevant features”, *In Springer Berlin Heidelberg*. pp. 137—142

Kanika Gupta and Monojit Choudhury and Kalika Bali.(2012).“Mining Hindi-English Transliteration Pairs from Online Hindi Lyrics”, *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC '12)*, Istanbul, Turkey, pp. 2459--2465.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lance, D. (1975). “Spanish-English code-switching. In: El lenguaje de los Chicanos”, Edited by E. Hern Sndez-Chavez, A. Cohen, and A. Beltramo. Arlington, Va. Center for Applied Linguistics

Marc Damashek. (1995). “Gauging similarity with n-grams: Language-independent categorization of text”, *In Science*, 267(5199), pp. 843--849.

Pal, S., (2013). "Mother Tongue Influence on Spoken English", *In Conference proceedings ICT for language learning*, libreriauniversitaria. it Edizioni. Vancouver, pp. 454.

Pandey, Pooja, and Sharvari Govilkar, (2015)."A Framework for Sentiment Analysis in Hindi using HSWN."*In International Journal of Computer Applications 119.19.*

Patil, S. and Davies, P.(2014).“Use of Google Translate in medical communication: evaluation of accuracy”, *BMJ*, *349*, pp. 7392.

Pedraze, P. (1979). “Ethnographic observations of language use in El B'arrio”, *Ms. New York: Center for Puerto Rican Studies.*

Penelope Sibun and Jeffrey C. Reynar. (1996). “Language identification: Examining the issues”*, In Proceedings of SDAIR’96*, pages 125--135.

Pieter Muysken. (2001). “The study of code-mixing. In Bilingual Speech: A typology of Code-Mixing”, Cambridge University Press.

Poplack, S. (1980). “Sometimes I’ll start a sentence in English y termino en espanol: Toward a typology of code-switching”, *Linguistics* 18, pp. 581--616.

Rao, R.V., Savsani, V.J. and Vakharia, D.P., (2011). “Teaching–learning-based optimization: a novel method for constrained mechanical design optimization problems”, *In Computer-Aided Design*, *43*(3), pp.303--315.

Sharma, S., Srinivas, P., and Balabantaray, R. C. (2015).“Text normalization of code mix and sentiment analysis”.*In Advances in Computing, Communications and Informatics (ICACCI),* International Conference on IEEE. pp. 1468--1473

Sinha, R.M.K. and Jain, A., (2003). “AnglaHindi: an English to Hindi machine-aided translation system”, *In MT Summit IX, New Orleans, USA*, pp.494--497.

Sivareddy’s Hindi Parts-Of-Speech Tagger: http://sivareddy.in/downloads

Stanford Log-linear Part-Of-Speech Tagger: http://nlp.stanford.edu/software/tagger.shtml

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M., (2011). “Lexicon-based methods for sentiment analysis”, *In Computational linguistics,37(2),* pp.267--307.

Ted Dunning.(1994). “Statistical identification of language”, *Technical Report* MCCS-94-273, Computing Research Lab, New Mexico State University.

V. Hatzivassiloglou and J. Wiebe, (2000). “Effects of adjective orientation and gradability on sentence subjectivity”, *In COLING 2000.*

Valdes Fallis, G. (1978). “Code-switching as a deliberate verbal strategy: a microanalysis of- direct and indirect requests-among bilingual Chicano speakers”, To appear in R. Dunin (ed.), *Latino Language and Communicative Behavior.* New Jersey: Ablex Publishing Corp.

Whitelaw, C., Garg, N. and Argamon, S., (2005). “Using appraisal groups for sentiment analysis”.*In Proceedings of the 14th ACM international conference on Information and knowledge management,* pp. 625--631