Sentiment Analysis of Code-Mixed Indian Languages: An Overview of SAIL_Code-Mixed Shared Task @ICON-2017

Abstract

Sentiment analysis is essential in many real-world applications such as stance detection, review analysis, recommendation system, and so on. Sentiment analysis becomes more difficult when the data is noisy and collected from social media. India is a multilingual country; people use more than one languages to communicate within themselves. The switching in between the languages is called code-switching or code-mixing, depending upon the type of mixing. This paper presents overview of the shared task on sentiment analysis of code-mixed data pairs of Hindi-English and Bengali-English collected from the different social media platform. The paper describes the task, dataset, evaluation, baseline and participant's systems.

Introduction

The past decade witnessed rapid growth and widespread usage of social media platforms by generating a significant amount of user-generated text. The user-generated texts contain high information content in the form of news, expression, or knowledge. Automatically mining information from user-generated data is unraveling a new field of research in Natural Language Processing (NLP) and has been a difficult task due to unstructured and noisy nature. In spite of the existing challenges, much research has been conducted on user-generated data in the field of information extraction, sentiment analysis, event extraction, user profiling and many more.

According to Census of India, there are 22 scheduled languages and more than 100 non scheduled languages in India. There are 462 million internet users in India and most people know more than one language. They express their feelings or emotions using more than one languages, thus generating a new code-mixed/code-switched language. The problem of code-mixing and code-switching are well studied in the field of NLP BIBREF0 , BIBREF1 . Information extraction from Indian internet user-generated texts become more difficult due to this multilingual nature. Much research has been conducted in this field such as language identification BIBREF2 , BIBREF3 , part-of-speech tagging BIBREF4 . Joshi et al. JoshiPSV16 have performed sentiment analysis in Hindi-English (HI-EN) code-mixed data and almost no work exists on sentiment analysis of Bengali-English (BN-EN) code-mixed texts. The Sentiment Analysis of Indian Language (Code-Mixed) (SAIL _Code-Mixed) is a shared task at ICON-2017. Two most popular code-mixed languages namely Hindi and Bengali mixed with English were considered for the sentiment identification task. A total of 40 participants registered for the shared task and only nine teams have submitted their predicted outputs. Out of nine unique submitted systems for evaluation, eight teams submitted fourteen runs for HI-EN dataset whereas seven teams submitted nine runs for BN-EN dataset. The training and test dataset were provided after annotating the languages and sentiment (positive, negative, and neutral) tags. The language tags were automatically annotated with the help of different dictionaries whereas the sentiment tags were manually annotated. The submitted systems are ranked using the macro average f-score.

The paper is organized as following manner. Section SECREF2 describes the NLP in Indian languages mainly related to code-mixing and sentiment analysis. The detailed statistics of the dataset and evaluation are described in Section SECREF3 . The baseline systems and participant's system description are described in Section SECREF4 . Finally, conclusion and future research are drawn in Section SECREF5 .

Related Work

With the rise of social media and user-generated data, information extraction from user-generated text became an important research area. Social media has become the voice of many people over decades and it has special relations with real time events. The multilingual user have tendency to mix two or more languages while expressing their opinion in social media, this phenomenon leads to generate a new code-mixed language. So far, many studies have been conducted on why the code-mixing phenomena occurs and can be found in Kim kim2006reasons. Several experiments have been performed on social media texts including code-mixed data. The first step toward information gathering from these texts is to identify the languages present. Till date, several language identification experiments or tasks have been performed on several code-mixed language pairs such as Spanish-English BIBREF5 , BIBREF6 , French-English BIBREF7 , Hindi-English BIBREF0 , BIBREF1 , Hindi-English-Bengali BIBREF8 , Bengali-English BIBREF1 . Many shared tasks have also been organized for language identification of code-mixed texts. Language Identification in Code-Switched Data was one of the shared tasks which covered four language pairs such as Spanish-English, Modern Standard Arabic and Arabic dialects, Chinese-English, and Nepalese-English. In the case of Indian languages, Mixed Script Information Retrieval BIBREF9 shared task at FIRE-2015 was organized for eight code-mixed Indian languages such as Bangla, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, and Telugu mixed with English.

The second step is the identification of Part-of-Speech (POS) tags in code-mixed data and only handful of experiments have been performed in it such as Spanish-English BIBREF10 , Hindi-English BIBREF11 . POS Tagging for Code-mixed Indian Social Media shared task was organized for language pairs such as Bengali-English, Hindi-English, and Telugu-English. However, to best of the authors' knowledge no tasks on POS tagging were found on other code-mixed Indian languages. Again, Named Entity Recognition (NER) of code-mixed language shared task was organized for identifying named entities in Hindi-English and Tamil-English code-mixed data BIBREF12 .

Sentiment analysis or opinion mining from code-mixed data is one of the difficult tasks and the reasons are listed below.

Sentiment analysis of Hindi-English code-mixed was performed by Joshi et al. JoshiPSV16 which used sub-word level representations in LSTM architecture to perform it. This is one of the initial tasks in sentiment analysis of HI-EN code-mixed dataset. There are several applications on code-mixed data which depends on sentiment analysis such as stance detection, aspect based sentiment analysis. However, there are several tasks available on sentiment analysis of Indian language tweets BIBREF13 , BIBREF14 . The shared task on sentiment analysis in Indian languages (SAIL) tweets focused on sentiment analysis of three Indian languages: Bengali, Hindi, and Tamil BIBREF13 .

Dataset and Evaluation

This section describes statistics of the dataset and the evaluation procedure. Preparing a gold standard dataset is the first step towards achieving good accuracy. Several tasks in the field of NLP suffer from lack of gold standard dataset. In the case of Indian languages, there is no such code-mixed dataset available for research purpose. Thus, we developed the dataset and the details are provided below.

Dataset

Data collection is a time consuming and tedious task in terms of human resource. Two code-mixed data pairs HI-EN and BN-EN are provided for developing sentiment analysis systems. The Twitter4j API was used to collect both Bengali and Hindi code-mixed data from Twitter. Initially, common Bengali and Hindi words were collected and then searched using the above API. The collected words are mostly sentiment words in Romanized format. Plenty of tweets had noisy words such as words from other languages and words in utf-8 format. After collection of code-mixed tweets, some were rejected. There are three reasons for which a tweet was rejected.

a tweet is incomplete, i.e. there is not much information available in the tweet.

a tweet is spam, advertisement or slang.

a tweet does not have either Bengali or Hindi words.

The hashtags and urls are kept unchanged. Then words are automatically tagged with language information using a dictionary which is developed manually. Finally, tweets are manually annotated with the positive, negative, and neutral polarity. Missed language tags or wrongly annotated language tags are corrected manually during sentiment annotation.

Any of the six language tags is used to annotate the language to each of the words and these are HI (Hindi), EN (English), BN (Bengali), UN(Universal), MIX (Mix of two languages), EMT (emoticons). MIX words are basically the English words with Hindi or Bengali suffix, for example, Delhite (in Delhi). Sometimes, the words are joined together by mistake due to the typing errors, for example, jayegiTension (tension will go away). UN words are basically symbols, hashtags, or name etc. The statistics of training and test tweets for Bengali and Hindi code-mixed datasets are provided in Table TABREF23 . Some examples of HI-EN and BN-EN datasets with sentiment tags are given below.

BI-EN: Irrfan Khan hollywood e abar dekha debe, trailer ta toh awesome ar acting o enjoyable. (positive)

Tagged: Irrfan/EN Khan/EN hollywood/EN e/BN abar/BN dekha/BN debe/BN ,/UN trailer/EN ta/BN toh/BN awesome/EN ar/BN acting/EN o/BN enjoyable/EN ./UN

Translation: Irrfan Khan will be seen in Hollywood again, trailer is awesome and acting is also enjoyable.

BI-EN: Ei movie take bar bar dekheo er matha mundu kichui bojha jaye na. Everything boddo confusing and amar mote not up to the mark. (negative)

Tagged: Ei/BN movie/EN take/BN bar/BN bar/BN dekheo/BN er/BN matha/BN mundu/BN kichui/BN bojha/BN jaye/BN na/BN ./UN Everything/EN boddo/BN confusing/EN and/EN amar/BN mote/BN not/EN up/EN to/EN the/EN mark/EN ./UN

Translation: After watching repeated times I can't understand anything. Everything is so confusing and I think its not up to the mark.

HI-EN: bhai jan duaa hei k appki film sooper dooper hit ho (positive)

Tagged: bhai/HI jan/HI duaa/HI hei/HI k/HI appki/HI film/HI sooper/EN dooper/HI hit/HI ho/HI ./UN

Translation: Brother I pray that your film will be a super duper hit.

HI-EN: yaaaro yeah #railbudget2015 kitne baaje start hooga ? (neutral)

Tagged: yaaaro/HI yeah/EN #railbudget2015/EN kitne/HI baaje/HI start/EN hooga/EN ?/UN

Translation: Friends, when will #railbudget2015 start?

Evaluation

The precision, recall and f-score are calculated using the sklearn package of scikit-learn BIBREF15 . The macro average f-score is used to rank the submitted systems, because it independently calculates the metric for each classes and then takes the average hence treating all classes equally. Two different types of evaluation are considered and these are described below.

Overall: The macro average precision, recall, and f-score are calculated for all submitted runs.

Two way: Then, two way classification approach is used where the system will be evaluated on two classes. For positive sentiment calculation, the predicted negative and neutral tags are converted to other for both gold and predicted output by making the task as binary classification. Then, the macro averaged precision, recall, and f-score are calculated. Similar process is also applied for negative and neural metrics calculation.

Baseline

The baseline systems are developed by randomly assigning any of the sentiment values to each of the test instances. Then, similar evaluation techniques are applied to the baseline system and the results are presented in Table TABREF29 .

System Descriptions

This subsection describes the details of systems submitted for the shared task. Six teams have submitted their system details and those are described below in order of decreasing f-score.

IIIT-NBP team used features like GloVe word embeddings with 300 dimension and TF-IDF scores of word n-grams (one-gram, two-grams and tri-grams) as well as character n-grams (n varying from 2 to 6). Sklearn BIBREF15 package is used to calculate the TF-IDF. Finally, two classifiers: ensemble voting (consisting of three classifiers - linear SVM, logistic regression and random forests) and linear SVM are used for classification.

JU_KS team used n-gram and sentiment lexicon based features. Small sentiment lexicons are manually prepared for both English and Bengali words. However, no sentiment lexicon is used for Hindi language. Bengali sentiment lexicon consists of a collection of 1700 positive and 3750 negative words whereas English sentiment lexicon consists of 2006 positive and 4783 negative words. Finally, Naïve Bayes multinomial is used to classify and system results are presented in Table TABREF29 .

BIT Mesra team submitted systems for only HI-EN dataset. During preprocessing, they removed words having UN language tags, URLs, hashtags and user mentions. An Emoji dictionary was prepared with sentiment tags. Finally, they used SVM and Naïve Bayes classifiers on uni-gram and bi-gram features to classify sentiment of the code-mixed HI-EN dataset only.

NLP_CEN_AMRITA team have used different distributional and distributed representation. They used Document Term Matrix with N-gram varying from 1 to 5 for the representation and Support Vector Machines (SVM) as a classifier to make the final prediction. Their system performed well for n-grams range 5 and minimum document frequency 2 using the linear kernel.

CFIL team uses simple deep learning for sentiment analysis on code-mixed data. The fastText tool is used to create word embeddings on sentiment corpus. Additionally, Convolutional Neural Networks was used to extract sub-word features. Bi-LSTM layer is used on word embedding and sub-word features together with max-pooling at the output which is again sent to a softmax layer for prediction. No additional features are used and hyper-parameters are selected after dividing training corpus to 70% and 30%.

Subway team submitted systems for HI-EN dataset only. Initially, words other than HI and EN tags are removed during the cleaning process. Then, a dictionary with bi-grams and tri-grams are collected from training data and sentiment polarity is annotated manually. TF-IDF scores for each matched n-grams are calculated and weights of 1.3 and 0.7 are assigned to bi-grams and tri-grams, respectively. Finally, Naïve Bayes classifier is used to get the sentiment.

Results and Discussion

The baseline systems achieved better scores compared to CEN@AMRIT and SVNIT teams for HI-EN dataset; and AMRITA_CEN, CEN@Amrita and SVNIT teams for BN-EN dataset. IIIT-NBP team has achieved the maximum macro average f-score of 0.569 across all the sentiment classes for HI-EN dataset. IIIT-NBP also achieved the maximum macro average f-score of 0.526 for BN-EN dataset. Two way classification of HI-EN dataset achieved the maximum macro average f-score of 0.707, 0.666, and 0.663 for positive, negative, and neutral, respectively. Similarly, the two way classification of BN-EN dataset achieved the maximum average f-score of 0.641, 0.677, and 0.621 for positive, negative, and neutral, respectively. Again, the f-measure achieved using HI-EN dataset is better than BN-EN. The obvious reason for such result is that there are more instances in HI-EN than BN-EN dataset.

Most of the teams used the n-gram based features and it resulted in better macro average f-score. Most teams used the sklearn for identifying n-grams. IIITH-NBP team is only team to use character n-grams. Word embeddings is another important feature used by several teams. For word embeddings, Gensim and fastText are used. JU_KS team has used sentiment lexicon based features for BN-EN dataset only. BITMesra team has used emoji dictionary annotated with sentiment. Hashtags are considered to be one of the most important features for sentiment analysis BIBREF16 , however they removed hashtags during sentiment identification.

Apart from the features, most of the teams used machine learning algorithms like SVM, Naïve Bayes. It is observed that the deep learning models are quite successful for many NLP tasks. CFIL team have used the deep learning framework however the deep learning based system did not perform well as compared to machine learning based system. The main reason for the above may be that the training datasets provided are not sufficient to built a deep learning model.

Conclusion and Future Work

This paper presents the details of shared task held during the ICON 2017. The competition presents the sentiment identification task from HI-EN and BN-EN code-mixed datasets. A random baseline system obtained macro average f-score of 0.331 and 0.339 for HI-EN and BN-EN datasets, respectively. The best performing team obtained maximum macro average f-score of 0.569 and 0.526 for HI-EN and BN-EN datasets, respectively. The team used word and character level n-grams as features and SVM for sentiment classification. We plan to enhance the current dataset and include more data pairs in the next version of the shared task. In future, more advanced task like aspect based sentiment analysis and stance detection can be performed on code-mixed dataset.