IJCRT.ORG ISSN : 2320-2882
INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

MULTILINGUAL TEXT CLASSIFICATION USING SENTIMENT ANALYSIS

[1]P Venkata Tanusha, [2]S Kayani, [3]P Sruthi, [4]A VishnuPriya, [5]V Sukanya
[1]B.Tech IV, [2]Professor, [3]B.Tech IV , [4]B.Tech IV, [5]B.Tech IV
[1,2,3,4,5] Information Technology,
[1,2,3,4,5] Vignan's Institute of Engineering for Women, Visakhapatnam,
India

***Abstract**:* Sentiment analysis (SA) using code-mixed data from social media has several applications in opinion mining ranging from customer satisfaction to social campaign analysis in multilingual societies. We use a Hindi-English (Hi-En) and Telugu-English (Tel-En) code-mixed datasets for sentiment analysis and perform empirical analysis comparing the suitability and performance of various state-of-the-art SA methods in social media. To do any further advancement in code-mixed data, the necessary step is data preprocessing, Word Variation, sentiment analysis, sentence classification into positive, negative and neutral.

Index Terms - **Heartbeat Sentiment analysis, Code-mixed data, Word variation, Campaign analysis**.

I. INTRODUCTION

Machine Learning is an artificial intelligence discipline geared toward the technological development of human knowledge. It allows computers to handle new Situations via analysis, self-training, observation and experience. Machine Learning is often confused with data mining and Knowledge Discovery Database (KDD), which share a similar methodology. Machine Learning facilitates the continuous advancement of computing through exposure to new scenarios, testing and adaptation. While employing pattern and trend detection for improved decisions in subsequent(though not identical situations). In multilingual societies like India, users generally combine the prominent language, like English, with their native languages. This process of switching texts between two or more languages is referred to as code-mixing. Millions of internet users in India communicate by mixing their regional languages with English which generates enormous amount of code-mixed social media texts. For example, "tum bahut super ho" , meaning *"you are superb"*, is a Hi-En code-mixed text. The linguistic complexity of code-mixed content is compounded by the presence of spelling variations, transliteration and non-adherence to formal grammar. The Code-mixed data on social media presents inherent challenges like word or phrase contractions ("message" to "msg"), and non-standard spellings (such as "wowww" or "suppeerrrrr"), etc. Along with diverse sentence constructions, words in Hindi can have multiple variations when written in English which leads to a large amount of sparse and rare tokens. For instance, "bahut"(very) can be written as "bahout", "bahot", "bhout", "bauhat", or "bhot", etc. Some models are proposed to understand the meaning from multilingual data as the system cannot understand the code mixed data but the accuracy is very less. This is challenging for sentiment analysis as traditional semantic analysis approaches do not capture meaning of the sentences. Scarcity of annotated data available for sentiment analysis also limit the advances in the field.

In this paper, we propose an ensemble model where we combine the outputs of dense networking and character-trigrams based LSTM to predict the sentiment of Hi-En code-mixed data. While the LSTM model encodes deep sequential patterns in the text and dense networking is used to convert the text to input to the LSTM. These results reveal that our model is able to outperform other traditional machine learning approaches as well as the deep learning models proposed in literature.

II. EXISTING SYSTEM

Information extraction from user-generated code-mixed data is difficult due to its multilingual nature. Language identification tasks have been performed on several code-mixed language. NLP specific tasks such as POS and Madan Gopal Jhanwar†, Arpita Das.An Ensemble Model for Sentiment Analysis of Hindi-English Code-Mixed Data.arXiv:1806.04450v1 [cs.CL] 12 Jun 2018,Microsoft India Development Center,Nurendra Choudhary, Rajat Singh, Ishita Bindlish, Manish Shrivastava, Sentiment Analysis of Code-Mixed Languages leveraging Resource Rich Languages, 19th International Conference on Computational Linguistics and Intelligent Text Processing have also been performed on the codemixed data. Initiatives have been taken by shared task like FIRE-20152 to study retrieval of mixed script of Indian languages. However, these proposed solutions do not align with the problem of sentiment analysis in code-mixed data.

Very less work has been done so far in the area of sentiment analysis of Hi-Encode-mixed data, A shared task for Sentiment Analysis of Indian Language (Code-Mixed) (SAIL Code-Mixed)3 on twitter data was organized at ICON-20174. Patraetal.(2015)summarizes the dataset used, various models submitted by the participants and their results. The best submission for the Hi-En language pair used features like GloVe word embeddings with 300 dimensions and TF-IDF scores of word and character ngrams. They trained an ensemble of linear SVM, Logistic Regression and Random Forests to classify the sentiments.

III. PROPOSED SYSTEM

3.1 Twitter Tweets:

Extracting the tweets from twitter using keys. Twitter API which is a python wrapper is used for performing API requests.

3.2 Preprocessing:

After the data is collected data preprocessing is done to eliminate the incomplete, noisy and inconsistent data. Data must be preprocessed in order to perform any data mining functionality.

3.3 Sentiment Analysis:

Machine Learning is related to prediction-making on some data.There are many machine learning algorithms.In the proposed system we use a parallel ensemble of two models –

- End-to-end deep learning model
- Traditional machine learning model

to classify a sentence into one of the *positive*, *negative* or *neutral* sentiment classes. For the deep learning model, the sentence is fed in the form of character-trigram embedding matrix. To the LSTM layer the embedding matrix is fed which encodes the sequential patterns in the query and outputs a feature representation. This representation then passes through a fully-connected layer, which models the various interactions between these features and outputs the probability of the sentence belonging to each of the three classes. For the traditional machine learning model, we feed the ngram features of the dense networking , which outputs the probability of the sentence belonging to each of the classes. We combine the outputs of both of the models to predict the final sentiment of the sentence.

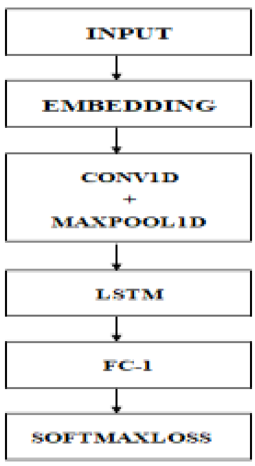

Fig 1:BLOCK DIAGRAM

3.3.1 Algorithm:

We choose LSTMs for the 3-class sentiment classification of the code mixed data, we designed a LSTM based classifier with the following details

Input Features : Each token is represented as a bag-of-character-trigrams vector. Maximum of 100 character trigram features is allowed then truncation is applied and in case of excess and deficit tokens padding is done. To the LSTM unit we fed 128 length embedding matrix for every token . Sparse code-mixed data representation feature is used as it removes the influence of word stem, to solve out-of-vocabulary issues and diverse variations .

Output : To a Fully Connected (FC) layer the output of the end state of the final LSTM layer is connected which models the interactions between these features and the classes . An activation function is used that is softmax to produce correctly normalized probability values.

Loss function : We train the parameters of the classifier with an objective of maximizing their predication accuracy given the target labels in the training set or minimizing the cross entropy error across the set. If o is the output of the network and t is the true label, the cross entropy (CE) loss function is calculated as follows:

$$CE(t,o)=-(t\log(o)+(1-t)\log(1-o)) \qquad (1)$$

The optimal hyper-parameter configuration of the classifier set is shown in Table

Table 1 : Hyperparameters of LSTM classifier

Hyperparameter	**Value**
Batch Size	32
Max length	100
LSTM cells	64
Character Embedding	128
Learning rate	0.01
Optimizer	Adagrad

IV. Dataset and Result

In this section,we will give a brief introduction of the dataset used and its result.

4.1 Dataset:

Joshietal.(2016) released a dataset for sentiment analysis of Hi-En code-mixed data. The dataset contains user comments from Narendra Modi, the Prime Minister of India at the time and public Facebook pages of Salman Khan, a Bollywood actor. The dataset contains 3979 sentences, split into 15% negative, 35% positive and 50% neutral classes. For experimentation, we divided the data into three sets – train set, development set and test set, in the ratio of 70%, 10% and 20% respectively. The train set is used to train the models, development set is used to tune the model parameters.

4.2 Result:

It can be observed that our Ensemble model outperforms other traditional and deep learning based models on a small Hi-En code-mixed data with an accuracy of 84.9.

Fig 2: Training the data

The below Figures depicts the predictions of the data when a HI-ENG sentence is given.

Fig 3: Positive Prediction

Fig 4: Neutral Prediction

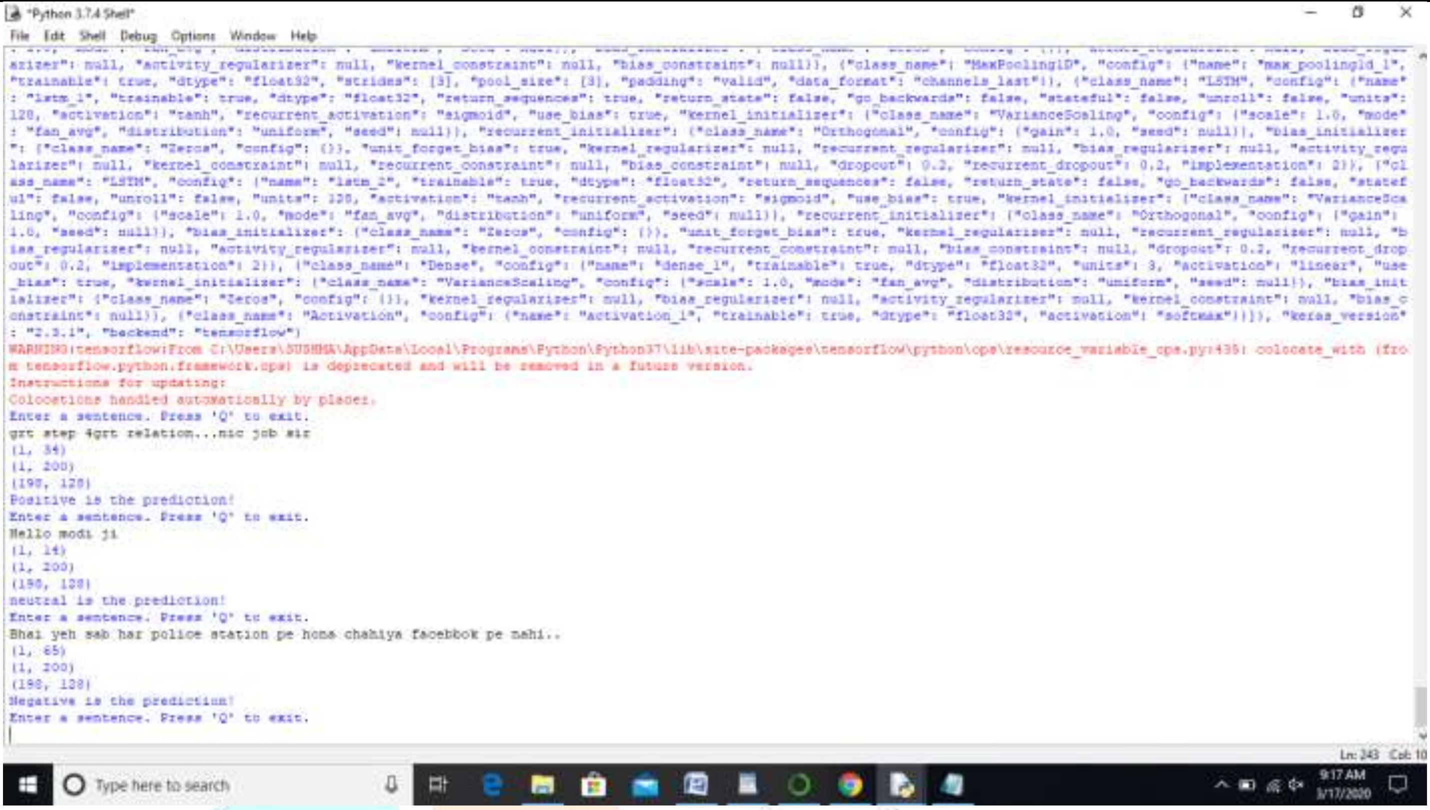

Fig 5: Negative Prediction

V. CONCLUSION

As there is an increase in popularity and impact of social media texts, analyzing the sentiments to maintain the understanding of the society plays a major role. In this paper we deal with sentiment analysis of the sparse and inconsistent Hi-En code-mixed data. Here we mainly deal with the shortcomings of deep learning models on a small multilingual code-mixed data. To identify the sentiment in code-mixed data we further propose a dense networking and char trigram based deep learning model(LSTM).

VI. SCOPE FOR FUTURE WORK

In future, we would like to extend our work to several other language pairs of code-mixed data. It would be interesting to utilize the rich features of individual languages to help identifying sentiments in their code-mixed version.

REFERENCES

[1] Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. Pos tagging of english-hindi code-mixed social media content. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 974–979, 2014

[2] Sida Wang and Christopher D Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, pages 90–94. Association for Computational Linguistics, 2012.

[3] ShiliangZhengandRuiXia. Left-center-rightseparatedneural network for aspect-based sentiment analysis with rotatory attention. arXiv preprint arXiv:1802.00892, 2018.

[4] Madan Gopal Jhanwar†,ArpitaDas.An Ensemble Model for Sentiment Analysis of Hindi-English Code-MixedData.arXiv:1806.04450v1 [cs.CL] 12 Jun 2018,Microsoft India Development Center.

[5] Nurendra Choudhary, Rajat Singh, Ishita Bindlish, Manish Shrivastava, Sentiment Analysis of Code-Mixed Languages leveraging Resource Rich Languages, 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2018), India {nurendra.choudhary,rajat.singh}@research.iiit.ac.in ishita.bindlish@students.iiit.ac.in m.shrivastava@iiit.ac.in

[6] Siaw Ling Lo1 · Erik Cambria2 · Raymond Chiong1 · David Cornforth1, Multilingual sentiment analysis: from formal to informal and scarce resource languages, 20 August 2016 © Springer Science+Business Media Dordrecht 2016

[7] Yu Zhou, Yanxiang Tong, Ruihang Gu, and Harald Gall. Combining text mining and data mining for bug report classification. JournalofSoftware: EvolutionandProcess, 28(3):150–176, 2016.

[8] ShashankSharma,PYKLSrinivas,andRakeshChandraBalabantaray. Text normalization of code mix and sentiment analysis. InAdvancesinComputing, Communicationsand Informatics (ICACCI), 2015 International Conference on, pages 1468–1473. IEEE, 2015.