NUIG-Shubhanker@Dravidian-CodeMix-FIRE2020: Sentiment Analysis of Code-Mixed Dravidian text using XLNet

Shubhanker Banerjee[a], Arun Jayapal[b] and Sajeetha Thavareesan[c]

[a]*National University Of Ireland Galway, Ireland*
[b]*Trinity college Dublin, Ireland*
[c]*Eastern University, Sri Lanka*

Abstract
Social media has penetrated into multi-lingual societies, however most of them use English to be a preferred language for communication. So it looks natural for them to mix their cultural language with English during conversations resulting in abundance of multilingual data – call this *code-mixed data*, available in today's world. Downstream NLP tasks using such data is challenging due to the semantic nature of it being spread across multiple languages. One such NLP task is Sentiment analysis; for this we use an auto-regressive XLNet model to perform sentiment analysis on code-mixed Tamil-English and Malayalam-English datasets.

Keywords
code-mixed, XLNet, auto-regressive, attention

1. Introduction

Social media content results in large data feeds from wide geographies. Since multiple geographies are involved, the data is multilingual in nature resulting in *code mixing* [1] often. Sentiment analysis on code-mixed text allows to gain insights on the trends prevalent in different geographies however is a challenge due to the non-trivial nature involved in inferring the semantics of such data. In this paper, we address these challenges using XLNet[1][2] framework. We have fine tuned the pre-trained XLNet model with the available data without any additional pre-processing mechanisms. The rest of the paper is organized as follows. Section 2 illustrates the related work done in the field. Section 3 explains the dataset and the task. Section 4 demonstrates the approach we used and briefly explains the architecture of XLNet. Results are discussed in Section 5.

[1]Code mixing refers to the linguistic units from different languages being used together by multilingual users
[2]XLNet is an auto-regressive model[2] which is built with the transformer architecture[3] with a two stream attention mechanism[4]. The two stream attention mechanism ensures that the language model obtained through training can predict missing words on the basis of bidirectional context. Bidirectional context in XLNet is achieved by permutation language modelling.

2. Related Work

Multilingual users have the tendency to mix linguistic units in the social media resulting in code-mixed data being easily available. The phenomenon of code-mixing is explained in [5, 6, 7, 8, 9, 10] and provides an analysis on the possible reasons behind code-mixing. This is done by identifying the languages involved in the code-mixed data which looks inevitable. In the past, several approaches were taken and experiments were conducted aimed at the detection of languages in code-mixed data [11][12][13]. A review of many research works on code-mixing is discussed in [14].

2.1. Code-mixed data

Since code-mixed is mostly sourced from social media platforms, the data in it's raw form is highly unstructured and hence corpus creation to organize this unstructured data into datasets for further analysis pose a challenge. For some of the Indian languages, [15] has compiled a Tamil-English code-mixed dataset, the first annotated Tanglish[3] dataset. Similarly, [16] published a dataset for Malayalam-English code-mixed data, where the authors also provided references to the availability of other code-mixed datasets such as Chinese-English and Spanish-English. But significant work hasn't been done in the area of corpus creation for code-mixing for Indian languages. The Indian languages are considered to be under-resourced and so there is less interest in performing NLP tasks on these languages.

2.2. Sentiment-analysis

Sentiment analysis is a well known NLP task that infers the positive, negative and neutral sentiments from a statement in question. However there are very few works on the sentiment analysis over code-mixed data; [17][18] provides an overview of the work done on sentiment analysis of Dravidian code-mixed text. Another work, [19] compares the performance of different transformer architectures on the task of sentiment analysis of code-mixed data. [20] employed an approach based on lexicon to assign sentiment to Hindi-English code-mixed text. [21] illustrates a method to detect hate speech in code-mixed Hinglish dataset. For the purpose of conducting this research they used FIRE 2013 and FIRE 2014 datasets. [22] used a LSTM [23] based approach to improve the state-of-art performance [20] on the hinglish datasets by 18 percent. [24] used shared parameters in a siamese network [25] to project the code-mixed sentences and sentences in standard languages into a common sentiment space. The similarity of projected sentences is an indicative of how similar their sentiments are, similar sentences have similar sentiment. Ensemble based techniques have also been used for sentiment analysis of code-mixed data, [26] proposed an ensemble model of a character-trigrams based LSTM and a word-ngrams naive bayes to detect the sentiment in Hindi-english code-mixed data. [27] have used a multilayer perceptron to perform sentiment analysis on code-mixed data extracted from social media platforms. [28] used an ensemble of a convolutional neural network and a self-attention based LSTM for sentiment analysis of Spanglish and Hinglish text.

---

[3]Tanglish refers to code switching between Tamil and English, a term predominantly used in the Tamil community

Table 1
Dataset size and splits

Dataset	Training	Validation	Testing
Tamil-English	1,335	1,260	3,149
Malayalam-English	4,716	674	1,348

3. Dataset for sentiment analysis

In spoken and written conversations, it is observed that the usage of lexicon, connectives and phrases from English are used in combination with other languages; this can very well be seen in the social media text and in spoken conversations across geographies, especially in India.

Sentiment Analysis in social media has drawn attention in recent years. However, sentiment analysis on Tamil-English (Tanglish) and Malayalam-English code-mixed data are not readily available for research. The authors of [15] and [16] have collected 184,573 sentences for Tamil and 116,711 sentences for Malayalam from YouTube comments which are based on the trailers of the movies released in 2019 for building Tamil-English and Malayalam-English datasets where non-code-mixed sentences were removed from the collection. Further, emoticons were removed and sentence length filters were applied to render the mentioned datasets. In the end two data sets of size 15,744 and 6,738 sentences were reported for Tanglish and Malayalam-English texts.

To get this dataset ready for sentiment analysis [15] and [16] refers to manual annotation activity carried out with three annotators annotating each sentence in the data set. The Krippendorff's alpha (α) is used to measure inter-annotator agreement which is 0.6585 and 0.890 for Tamil and Malayalam code-mixed data sets respectively.

The dataset was provided for this task in three parts training, validation and testing. The number of sentences used for the dataset splits are provided in table 1. Both these datasets are released in DravidianCodeMix FIRE 2020 competition organized by dravidiancodemixed. These comments were grouped into five categories positive, negative, neutral, mixed emotions, or not in the intended languages.

4. Methodology

Language models have been integral to the recent advances made in the field of NLP due to its ability to predict the next token in a sequence. Traditionally this achieved by computing the joint distribution of the tokens in a sequence as a function of conditional probability distribution of each token given other tokens in the sequence.

However, XLNet[1] takes a different approach; when these models are trained on large datasets, it achieves state-of-art performances on downstream NLP tasks. This uses permutation language modelling, which trains an autoregressive model on all possible permutation of words in a sentence – see equation 1. During prediction of a word in a sequence, it takes into account bidirectional context and predicts the masked tokens on the basis of the words/tokens to the right as well as the left of the masked token in the sequence. XLNet is based on the transformer architecture[3], which uses the concept of attention[4] to learn the long range token dependencies. Another important aspect of XLnet is two-stream attention; this refers

Table 2

Precision, Recall and F-score measures on the Test set

Data	Classes	Precision	Recall	F1 Score	Weighted Average-F1	Accuracy
Malayalam-English	Mixed feelings	0.03	0.22	0.05	0.52	0.49
	Negative	0.12	0.40	0.19		
	Positive	0.72	0.50	0.59		
	not-malayalam	0.36	0.58	0.44		
	unknown state	0.42	0.46	0.44		
Tamil-English	Mixed feelings	0.23	0.13	0.17	0.32	0.35
	Negative	0.50	0.16	0.24		
	Positive	0.39	0.73	0.51		
	not-Tamil	0.10	0.40	0.16		
	unknown state	0.03	0.31	0.05		

to attention streams working in parallel, one which encodes the content of the tokens and the other which incorporates the positional information. This property would be useful and so is exploited to perform the sentiment analysis on code-mixed data.

The following equation formally describes the language modelling objective using XLNet. In eqn.1, for a give text sequence x, and set of all permutations of the sequence Z_T and $z \in Z_T$.

$$\max_{\theta} E_{z \sim Z_T} [\sum_{t=1}^{T} \log p_{\theta} \, x_{z_t} | x_{z<t}) \,] \tag{1}$$

For the purpose of experiments, we fine-tuned the XLNet model using the given datasets – refer to table 1 and sentiment analysis was conducted on this labelled dataset. The training and testing were carried out as per the numbers mentioned in Table 1. Two experiments were conducted on the mentioned datasets and for those experiments the XLNet embeddings were fine-tuned for 4 epochs each with a maximum learning rate of 0.005 to perform sentiment analysis of the given datasets. Results of the experiments carried out are illustrated in the next section.

5. Results and Discussion

The experiment results for the experiments outlined in the previous section are provided in table 2. We were able to achieve 0.49 & 0.35 accuracies and 0.52 & 0.32 F-scores on both the datasets respectively. The results are biased towards *Positive* class because of the class-imbalance seen in the training set. Further, it can be seen that the model performs better on the Malayalam-English dataset despite the Tanglish dataset having more samples; this can be attributed to more noise in the Tamil-English data and hence relatively poor performance. Our results do not perform better than the baseline-results described in [15] and [16]. We hypothesize that these results can further be improved by training the model for more epochs with a pre-processing step performed in combination with oversampling and undersampling of the minority and majority classes respectively.

References

[1] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, 2019. arXiv:1906.08237.

[2] K. Gregor, I. Danihelka, A. Mnih, C. Blundell, D. Wierstra, Deep autoregressive networks, in: International Conference on Machine Learning, PMLR, 2014, pp. 1242–1250.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[4] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014. arXiv:1409.0473.

[5] E. Kim, Reasons and motivations for code-mixing and code-switching, Issues in EFL 4 (2006) 43–61.

[6] B. R. Chakravarthi, M. Arcan, J. P. McCrae, WordNet gloss translation for under-resourced languages using multilingual neural machine translation, in: Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation, European Association for Machine Translation, Dublin, Ireland, 2019, pp. 1–7. URL: https://www.aclweb.org/anthology/W19-7101.

[7] B. R. Chakravarthi, R. Priyadharshini, B. Stearns, A. Jayapal, S. S, M. Arcan, M. Zarrouk, J. P. McCrae, Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription, in: Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages, European Association for Machine Translation, Dublin, Ireland, 2019, pp. 56–63. URL: https://www.aclweb.org/anthology/W19-6809.

[8] B. R. Chakravarthi, M. Arcan, J. P. McCrae, Comparison of different orthographies for machine translation of under-resourced Dravidian languages, in: 2nd Conference on Language, Data and Knowledge (LDK 2019), Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[9] B. R. Chakravarthi, P. Rani, M. Arcan, J. P. McCrae, A survey of orthographic information in machine translation, arXiv preprint arXiv:2008.01391 (2020).

[10] B. R. Chakravarthi, Leveraging orthographic information to improve machine translation of under-resourced languages, Ph.D. thesis, NUI Galway, 2020.

[11] S. Mandal, S. Banerjee, S. Naskar, P. Rosso, S. Bandyopadhyay, Adaptive voting in multiple classifier systems for word level language identification, 2015. doi:10.13140/RG.2.1.3976.0246.

[12] K. Bali, J. Sharma, M. Choudhury, Y. Vyas, "i am borrowing ya mixing ?" an analysis of english-hindi code mixing in facebook, 2014, pp. 116–126. doi:10.3115/v1/W14-3914.

[13] T. Solorio, M. Sherman, Y. Liu, L. M. Bedore, E. D. Peña, A. Iglesias, Analyzing language samples of spanish-english bilingual children for the automated prediction of language dominance, Nat. Lang. Eng. 17 (2011) 367–395. URL: https://doi.org/10.1017/S1351324910000252. doi:10.1017/S1351324910000252.

[14] N. Jose, B. R. Chakravarthi, S. Suryawanshi, E. Sherly, J. P. McCrae, A survey of current datasets for code-switching research, in: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), IEEE, 2020, pp. 136–141.

[15] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, J. P. McCrae, Corpus creation for sentiment analysis in code-mixed Tamil-English text, in: Proceedings of the 1st Joint

Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), European Language Resources association, Marseille, France, 2020, pp. 202–210. URL: https://www.aclweb.org/anthology/2020.sltu-1.28.

[16] B. R. Chakravarthi, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, A sentiment analysis dataset for code-mixed Malayalam-English, in: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), European Language Resources association, Marseille, France, 2020, pp. 177–184. URL: https://www.aclweb.org/anthology/2020.sltu-1.25.

[17] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, E. Sherly, J. P. McCrae, Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text, in: Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020). CEUR Workshop Proceedings. In: CEUR-WS. org, Hyderabad, India, 2020.

[18] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, E. Sherly, J. P. McCrae, Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text, in: Proceedings of the 12th Forum for Information Retrieval Evaluation, FIRE '20, 2020.

[19] S. Banerjee, B. R. Chakravarthi, J. P. McCrae, Comparison of pretrained embeddings to identify hate speech in Indian code-mixed text, in: 2nd IEEE International Conference on Advances in Computing, Communication Control and Networking –ICACCCN (ICAC3N-20), 2020.

[20] S. Sharma, P. Srinivas, R. C. Balabantaray, Text normalization of code mix and sentiment analysis, in: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015, pp. 1468–1473.

[21] P. Rani, S. Suryawanshi, K. Goswami, B. R. Chakravarthi, T. Fransen, J. P. McCrae, A comparative study of different state-of-the-art hate speech detection methods in Hindi-English code-mixed data, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 42–48. URL: https://www.aclweb.org/anthology/2020.trac-1.7.

[22] A. Prabhu, A. Joshi, M. Shrivastava, V. Varma, Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text (2016).

[23] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[24] N. Choudhary, R. Singh, I. Bindlish, M. Shrivastava, Sentiment analysis of code-mixed languages leveraging resource rich languages, 2018. arXiv:1804.00806.

[25] S. K. Roy, M. Harandi, R. Nock, R. Hartley, Siamese networks: The tale of two manifolds, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3046–3055.

[26] M. G. Jhanwar, A. Das, An ensemble model for sentiment analysis of hindi-english code-mixed data, arXiv preprint arXiv:1806.04450 (2018).

[27] S. Ghosh, S. Ghosh, D. Das, Sentiment identification in code-mixed social media text, arXiv preprint arXiv:1707.01184 (2017).

[28] A. Kumar, H. Agarwal, K. Bansal, A. Modi, Baksa at semeval-2020 task 9: Bolstering cnn

with self-attention for sentiment analysis of code mixed text, 2020. arXiv:2007.10819.