Enhancing Deep Learning Approach for Tamil English Mixed Text Classification

Neeraj Bhargava(✉) and Anantika Johari

Department of Computer Science Engineering, MDS University, Ajmer, India
profneerajbhargava@gmail.com, ananitkajohari@gmail.com

Abstract. Text Classification with sentiments understanding is an essential task for data processing and predicting user behavior. In case of Multilingual data, the process requires to convert the entire data to machine understandable language or to pre-process the text prior to classification keeping the semantics of the text intact. Deep Learning libraries like Bidirectional Encoder Representations from Transformers (BERT) with word2vector model and Convolutional Neural Network (CNN) for natural language processing (NLP) support both techniques, and the manuscript attempts to enhance pre-processing of the Tamil English Mixed text Classification. The pre-processing of the Tamil English Mixed text addressed the issue of annotated text non-availability.

Keywords: BERT · CNN · Multilingual mask model

1 Introduction

The Data generation with the multilingual viewing support requires the information shown must be correctly classified and acceptable by the viewer. The text processing also requires the semantics/meaning of the text displayed to be precise and acceptable at the viewer end. The application are now created has multilingual code- mixed text like Tamil, Telugu, Kannada with English or Hindi with little support for annotated text provides the expansion and search for better classification techniques. The recent research supports the deep leaning techniques extended to the domain of text sentiment classification. The vectorization of the text/word's representation with low-dimensional data with dense layer approach provide better classification for the given text. The manuscript aims at enhancing the current state of art research for the Tamil-English-Text mixed data available at the UCI Machine learning repository (https://archive.ics.uci.edu/ml/machine-learning-databases/00610/). Bharathi Raja et al. 2020 showcase the corpus building and code-switching as the first attempt for towards processing of the Tamil-English-text. The major contribution was the annotated dataset for sentiment analysis and experimental analysis using various classification algorithms. The manuscript attempts to further enhance the results by deploying NLP based pre-processing to standardize the text for classification. The multilingual masked model (XLM) is also attempted for the comparison. The trade-off between the multilingual and monolingual model efforts (Liu et al. 2019b) towards scaling is also targeted in the study. The result outperforms the

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 829–837, 2023.
https://doi.org/10.2991/978-94-6463-136-4_73

initial analysis done and the classification analysis based on BERT, GBM & GLM based AUTOML and XGBOOST cross lingual results were presented and compared.

2 Related Work

The multilingual text requires code switching; ana alternate between the two languages is benefited with latest NLP statistical measures. However, for the monolingual analysis various corpora exist for English (Hu et al.2004, Wiebe et al. 2005, Jiang et al. 2019) and Indian Language (Agarwal et al. 2018; Rani et. al. 2022) which are of prime interest for various researchers. The Synthetic monolingual text generation is a challenging task for which Deep Learning model namely 1) the General Adversarial Technique (GAN) (Kannan et. al.) and 2) Variational Auto encoders (VAE), (Kingma et. al. 2013) are used for generating diverse and plausible synthetic texts which seems realistic.

One of the oldest classical languages still in use today is Tamil. "The only language of contemporary India that is clearly continuous with a classical past," it was stated. Because of its diversity and high caliber, traditional Tamil literature has been referred to as "one of the great classical traditions and literatures of the world." The language is mostly spoken in the southern region of India, countries like Sri Lanka (Chakravarthi et. al. 2018, 2019, 2020a). The English-Tamil paring was discussed for cross-lingual information retrieval and linguistic comparison (Sanjanasri et. al. 2020). However, the code-mixed data till Chakravarthi et. al. Worked upon is underdeveloped and were not readily available for research. The major classifier namely KNN, Decision Tree, Random Forest, Logistic Regression, Naïve Bayes, SVM, Ed Conv-LSTM, DME, CDME, BERT multilinguals were discussed with various performance metrics measure were showcased and compared for all classifiers.

Multilingual masked language models (MLM) like mBERT (Devlin et al., 2019) and XLM (Lample et. al. 2019) talked on the cross-lingual processing for large transformer models for many languages. The model works well for cross-lingual natural language inference (Bowman et. al., 2015; Williams et al., 2017; Coneau et al., 2018), Question and Answers (Rajpurkar et al., 2016; Lewis et al., 2019) and named entity identification (Pires et al., 2019; Wu et al., 2019). The major contributions share the XLM model outperform the mBERT model. However, most of the research worked on finetuning the model's performance for significant large amount of data. The research gap is seen in terms of no preprocessing of the dataset using existing natural language processing (NLP) techniques to standardize form. To achieve the better classification result we propose the use of BERT model to extract Feature representation.

3 BERT and Auto ML Model

3.1 BERT Model

The most recent language model BERT (Bidirectional Encoder Representations from Transformer) (Devlin et al., 2019) is a pre-trained language representation model that was trained on 16 GB of unlabeled texts, including Wikipedia and Books Corpus, with a total of 3.3 billion words and a vocabulary size of 30,522. Because it uses the masked

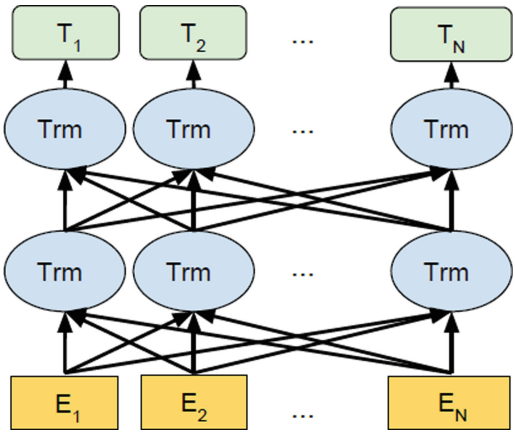

Fig. 1. BERT model for Text Classification from Devlin et al. 2019

language model (MLM) pre-training objective, it has a bidirectional structure shown in Fig. 1. That gives it an edge over other pre-trained language models like ELMo (Mathews et al., 2018) and ULMFiT (Horward et al., 2018). The MLM chooses 15% of the tokens in the input at random and uses context from both sides to predict the word's original vocabulary id. For NLP interpretation and inference, the pre-trained model can be employed straightaway to improve on incoming input.

3.2 GBM – Auto ML Model

XGBoost Gradient Boosting Machines (GBM), H2O Gradient Boosting Machines (GBM), Random Forests (Default and Extremely Randomized Tree variation), Deep Neural Networks and Generalized Linear Models (GLM) are all included in H2OAutoML. We can use this outside method in H2O AutoML since H2O provides a wrapper around the well known XGBoost programme. Additionally, training can now be accelerated on the GPU. The pre-specified models are included to provide each algorithm with rapid, trustworthy defaults. The user-customizable order of the algorithms is set to begin with models (pre-specified XGBoost models) that consistently produce strong results over a wide range of datasets, followed by an adjusted GLM for a quick reference point. From here, we focus on adding a few Random Forests, (H2O) GBM, and Deep Learning models in order to increase the diversity throughout our set of models (for the benefit of the final Stacked Ensembles). Following the training of these prescribed models and their addition to the leader board, we launch a random search using those identical methods. According to our perception or estimated "value" of each task, we explicitly determine the percentage of time spent on each method in the AutoML run, giving some algorithms (like XGBoost GBM and H2O GBM) more time than others (like H2O Deep Learning) (Fig. 2).

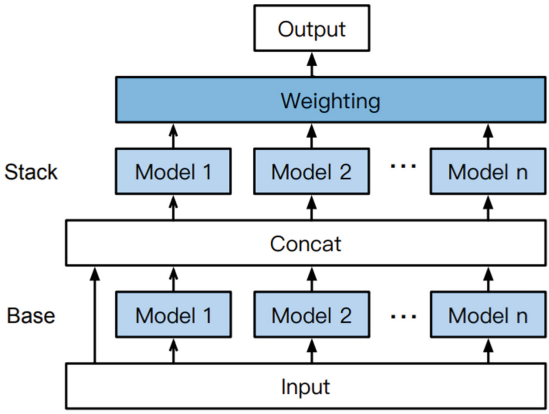

Fig. 2. AutoGluon's multi-layer stacking strategy from Erickson et al. 2020

Table 1. Data Distribution based on the Google Collab analysis

Class	Number of Instances
Positive	10559
Negative	2037
Mixed Feelings	1801
unknown states	850
not-Tamil	497

4 Dataset for Modeling

The Tamil English data set downloaded from the UCI repository consists of 15,744 sentences randomly shuffled and split with almost 11k approx. Sentences used for training and 1K and 3K words used for validation and testing. The exact distribution of the dataset based on the initial process applied in the Google Colab is shown in Table 1.

The dataset was further processed using the standard NLTK packages for removing the string punctuation, convert the entire dataset into lower case to avoid any ambiguity and tokenized the system based on Tamil stemmer Corpus. The final dataset generated can be seen in the below Fig. 3.

The inter annotator agreement is based on the Krippendorff's alpha (α) (Krippendorff's, 1970) stated in the based paper (Chakravarthi et. al. 2018, 2019, 2020b).

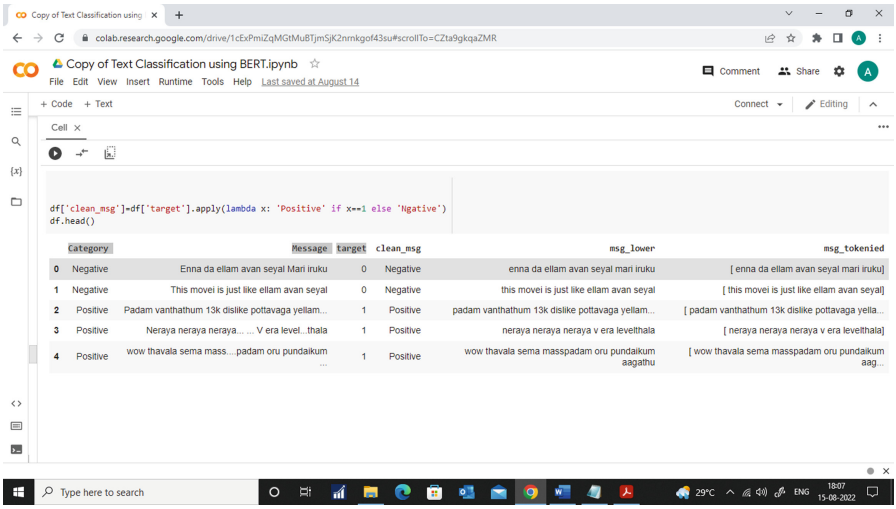

Fig. 3. Pre-processing of the Tamil-English dataset.

Table 2. Comparative Analysis of various model

Model	Precision	Recall	F-Score
mBERT	.6729	1.00	.80
Logistic Regression	.6786	.98	.83
CNN Model	.6623	.92	.81

5 Experiments Results and Discussion

The experimental work was executed on Google Co-LAB with both TPU, and GPU hardware acceleration facility utilized for Tensor and H2O layered network. The results obtained were summarized in the Table 2. For the mBERT, Logistic Regression, CNN Model based on the performance metrics measure namely Precision, Recall, F-Score.

The table clearly shows that almost 67% of the cases belong to the *Positive* class, leading to the inference that the dataset was highly imbalanced. The un-stability check was further scrutinized based on the unsupervised learning approach using the H2O open-source library for supervised and unsupervised learning algorithms. The automated result for various techniques falls in H2O library is depicted in the Table 3, Table 4 and (Fig. 4)

The results show the GBM based Auto-ML has the minimum loss and the class predicted is P1 representing the 'Positive' values for the text case and text imbalance. The results may be further enhanced by predicting the performance metrics for all the other classes. Also, the stemming of the Tamil corpus if enhanced and tried on the different deep learning model like VAE and GAN may give better results.

Table 3. H2O leather board for various Auto ML features

model_id	auc	logloss	aucpr	mean_per_class_error	rmse
GBM_1_AutoML_4_20220808_104409	1	7.57E-17	1	0	9.17E-17
GLM_1_AutoML_4_20220808_104409	1	0.000506411	1	0	0.000587
DRF_1_AutoML_4_20220808_104409	1	0.0960695	1	0	0.123631
XGBoost_2_AutoML_4_20220808_104409	1	0.00689967	1	0	0.00729

Table 4. Predicted Classes for the Vector Text

predict	p0	p1	p2	p3	p4
1	3.20E-05	1.000	1.59E-05	2.89E-05	9.50E-06
2	9.37E-05	0.000	0.999605	8.51E-05	2.87E-05
0	0.999696	0.000	3.97E-05	7.37E-05	2.35E-05
1	3.20E-05	1.000	1.59E-05	2.89E-05	9.50E-06
0	0.999696	0.000	3.97E-05	7.37E-05	2.35E-05
1	3.20E-05	1.000	1.59E-05	2.89E-05	9.50E-06
1	3.20E-05	1.000	1.59E-05	2.89E-05	9.50E-06
2	9.37E-05	0.000	0.999605	8.51E-05	2.87E-05

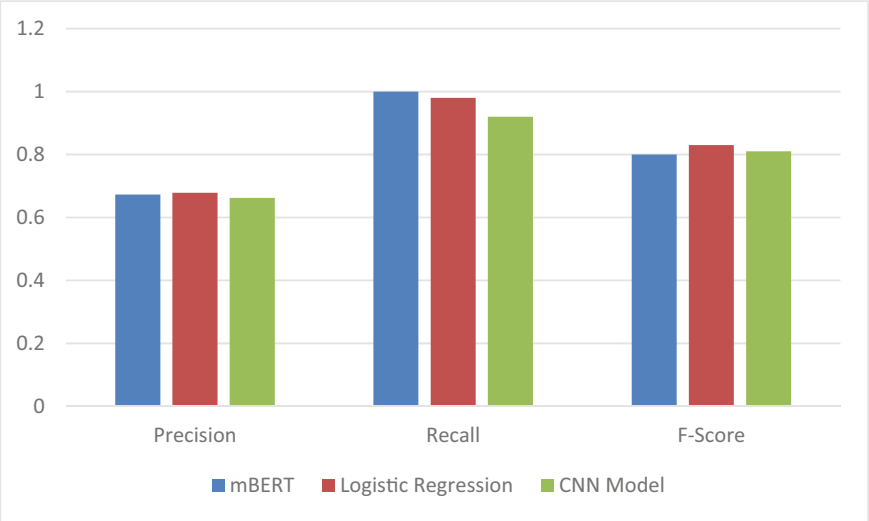

Fig. 4. Comparative Analysis of Various models

6 Conclusion

The manuscript discusses the classification models for the annotated Tamil-English mixed dataset. The dataset was enhanced using the pre-text processing using NLTK library for the mBERT model-based classification while for the H2O library the process is inbuilt function for implementing Gradient Boost model for Random Forest and linear model. The data processed for the Positive class in case mBERT, Logistic Regression and CNN model is calculated while for H2O based solution the prediction is done for all the classes. The performance metric measures-based results were discussed and elaborated.

References

Index of /ml/machine-learning-databases/00610 (uci.edu)

Krippendorff, K., Estimating the reliability, systematic error and random error of interval data. Edu. and Psy. Measurement, 30(1):61–70 (1970).

Hu, M. et al., Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD Int. Conf. on Know. Discovery and Data Mining, KDD '04, page 168–177, New York, NY, USA. ACM (2004).

Wiebe, J., et al., Annotating expressions of opinions and emotions in language. LRE, 39(2):165–210, May (2005).

Kannan, A. et al., Towards building a SentiWordNet for Tamil. In Proceedings of the 13th Int. Conf. on NLP, pages 30–35, Varanasi, India, December. NLP Association of India Author, F.: Article title. Journal 2(5), 99–110 (2016).

Rajpurkar P., et al., SQuAD: 100,000+ questions for machine comp. of text. In EMNLP, pages 2383–2392, Austin, Texas. ACL (2016).

Rajpurkar P., et al., Know what you don't know: Unanswerable questions for squad. ACL (2018).

Chakravarthi, B. R et al., Improving wordnets for underresourced languages using machine translation, 9 th Global WordNet Conference (GWC 2018) Proceedings, page 78 (2018).

Agrawal, R., et al., No more beating about the bush: A step towards idiom handling for Indian language NLP. 11th LREC 2018 Proceedings, Miyazaki, Japan, May. ELRA (2018).

Devlin, J., et al., BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June. ACL (2019).

Yinhan Liu, et al., Roberta: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019).

Jiang, Q., et al., A challenge dataset and effective models for aspect-based sentiment analysis. 9th EMNLP-IJCNLP Proceedings, pages 6279–6284, Hong Kong, China, November (2019).

Chakravarthi, B. R., et al., Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription. 2nd Workshop on Technologies for MT of Low Resource Languages proceedings, pages 56–63, 20 August, Dublin, Ireland. EAML (2019c).

Chakravarthi, B. R, et al., Comparison of different orthographies for machine translation of under-resourced Dravidian languages. In 2nd LDK Conference 2019. Schloss Dagstuhl-Leibniz Zentrum fuer Informatik (2019a).

Chakravarthi, B. R., et al, WordNet gloss translation for under-resourced languages using multilingual neural machine translation. 2nd Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation proceedings, pages 1–7, Dublin, Ireland, 19 August, EAML (2019b).

Rani, P., Suryawanshi, S., Goswami, K., Chakravarthi, B. R., Fransen, T., & McCrae, J. P. (2020). A comparative study of different state-of-the-art hate speech detection methods for Hindi-English code-mixed data. In: Proceedings of the second workshop on trolling, aggression and cyberbullying. Marseille, France: European Language Resources Association (ELRA)

Chakravarthi, B. R., Anand Kumar, M., McCrae, J. P., Premjith, B., Soman, K., & Mandl, T. (2020a). Overview of the track on HASOC-offensive Language Identification-DravidianCodeMix. In: Working notes of the forum for information retrieval evaluation (FIRE 2020). CEUR Workshop Proceedings, CEUR-WS. org

Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2020b). A sentiment analysis dataset for code-mixed Malayalam-English. In: Proceedings of the 1st joint workshop of SLTU (spoken language UDTechnologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) (SLTU-CCURL 2020). Marseille, France: European Language Resources Association (ELRA).

Chakravarthi, B. R., Muralidaran, V., Priyadharshini, R., & McCrae, J. P. (2020c). Corpus creation for sentiment analysis in code-mixed Tamil-English text. In: Proceedings of the 1st joint workshop of SLTU (spoken language technologies for under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) (SLTU-CCURL 2020). Marseille, France: European Language Resources Association (ELRA)

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. arXiv:1901.07291

Sai, S., Sharma, Y., 2020. Siva@HASOC-Dravidian-CodeMix-FIRE-2020: Multilingual offensive speech detection in code-mixed and romanized text, in: CEUR Workshop Proceedings

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., et al., 2020. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. arXiv preprint arXiv:200306505

Liu, Y., et al., 2019b. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692

Pires, T., Schlinger, E., Garrette, D., 2019. How multilingual is multilingual bert? arXiv preprint arXiv:1906.01502

Kingma, Diederik P and Welling, Max. Auto-Encoding Variational Bayes. In: The 2nd International Conference on Learning Representations (ICLR) (2013)

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146

Samuel R. Bowman, Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating crosslingual sentence representations

Williams A., Nangia N., Bowman S.R., A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426, 2017

Lewis P., Denoyer L., Riedel S, Unsupervised Question Answering by Cloze Translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4896–4910, Florence, Italy. Association for Computational Linguistics, 2019

Wu J., et. al., Language models are unsupervised multitask learners. OpenAI Blog 1:9, (2019). Available online at: https://arxiv.org/abs/1908.05691

Matthew E. Peters, et. al., Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Lin-guistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Vol.1 (Long Papers), pp. 2227–2237