UMUTeam@TamilNLP-ACL2022: Emotional Analysis in Tamil

José Antonio García-Díaz and **Rafael Valencia-García***
Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain
{joseantonio.garcia8,valencia}@um.es

Miguel Ángel Rodríguez-García
Departamento de Ciencias de la Computación, Universidad Rey Juan Carlos,
28933 Madrid, Spain
miguel.rodriguez@urjc.es

Abstract

This working notes summarises the participation of the UMUTeam on the TamilNLP (ACL 2022) shared task concerning emotion analysis in Tamil. We participated in the two multi-classification challenges proposed with a neural network that combines linguistic features with different feature sets based on contextual and non-contextual sentence embeddings. Our proposal achieved the 1st result for the second subtask, with an f1-score of 15.1% discerning among 30 different emotions. However, our results for the first subtask were not recorded in the official leader board. Accordingly, we report our results for this subtask with the validation split, reaching a macro f1-score of 32.360%.

1 Introduction

In this work, we detail the participation of the UMUTeam in the shared-task Tamil NLP (ACL 2022), concerning Emotion Analysis (EA) in Tamil (Sampath et al., 2022). Emotion detection is a recent field of research included in the broader research area of sentiment analysis. Here, the target of emotion detection aims at detecting types of feelings in natural language like anger, fear, disgust, happiness, surprise and sadness (Iqbal et al., 2022). In literature, strategies can be found addressing emotion detection in quite different domains. For instance, Shelke et al., (Shelke et al., 2022) propose an architecture based on Leaky Relu activated Deep Neural Network (LRA-DNN) to address emotion analysis on social media. The architecture is comprised of four steps: (1) preprocessing to clean the data and change it representation in a more understandable format; (2) feature extraction step to extract the most relevant characteristics; (3) ranking step where extracted features are assigned to ranks that they are optimised by using a nature-inspired meta-heuristic optimisation algorithm; and (4) classification where the LRA-DNN is employed. Yong et al., (Yong et al., 2022) describe a BCBLAC model designed to tackle emotion analysis in a food review. Its name is due to its layer architecture: Bert Layer, CNN layer, BLSTM layer, Attention layer and CRF layer. Each layer represents a step that the input must go through to carry out the emotion classification process.

In this shared task, the organisers challenged the participants to extract one emotion per document from a collection of social media comments written in Tamil. The organisers provided the participants with three sets: development, training and test. It is worth mentioning that we use these splits as expected, that is, we train with the training set and validate with the development set. This shared task was divided into two minor subtasks. The first subtask distinguishes among 11 emotions whereas the second subtask with 30 emotions. The name of the emotions, and the number of instances per training and validation are depicted in Figure 1.

Our research group has experience dealing with EA tasks. For example, we participated in the EmoEvalEs shared task, proposed in IberLef 2021 (Plaza-del Arco et al., 2021) concerning EA in Spanish. This task consisted into a multi classification task with the Ekman basic emotions. We achieved the 6th position with an accuracy of 68.5990% (4.1667% below the best result) (García-Díaz et al., 2021c). In this shared-task we participated with similar methods to the ones described in (García-Díaz et al., 2021c). However, here we conduct a more advance hyperparameter tuning stage. Besides, we use this task to validate a subset of language-independent linguistic features extracted with a custom tool that is part of the doctoral thesis of one the members of the team. In fact, we had participated in different automatic document classification tasks in Spanish to validate these linguistic features. We have observed that these linguistic features contribute to improve state-of-the-art models

*Corresponding author

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 39 - 44

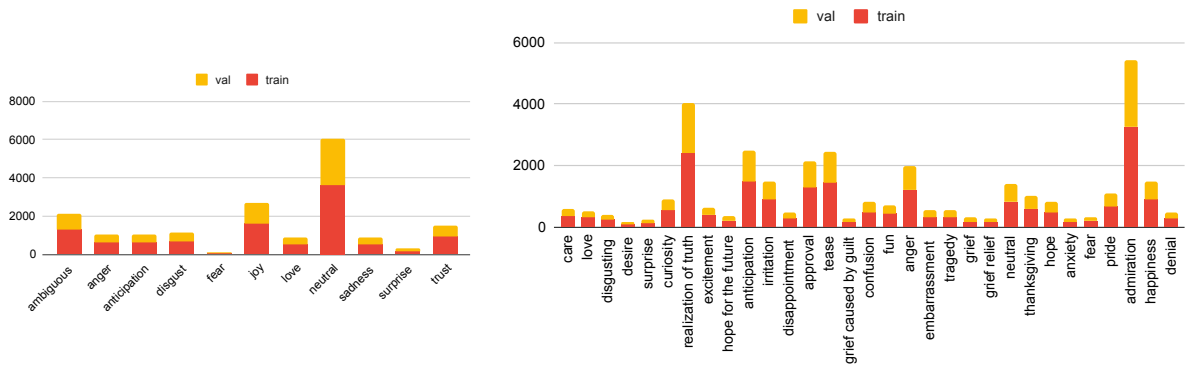

Figure 1: Label distribution for the first (left) and second (right) subtasks

based on Transformers. One of the secondary objectives of our participation is therefore to observe whether the subset of these linguistic features that are language-independent still improve the performance of automatic document classification in non-Latin languages. This subset of linguistic features are based on stylometry, which different metrics concerning word and sentence length as well as punctuation symbols. There are, in addition, features that capture emojis, hyperlinks, and social network jargon.

2 Methodology

In a nutshell, our participation consists in the development of a classifier based on neural network that uses four feature sets combined using a knowledge integration strategy. During the development stage, other methods for combining these features, such as ensemble learning, are evaluated.

Next, we describe the four feature sets in detail. The first feature set is **LF**, a subset of language-independent linguistic features extracted using the UMUTextStats tool (García-Díaz et al., 2021b; García-Díaz and Valencia-García, 2022). These features are stylometric features, PoS features based on the Tamil model of Stanza (Qi et al., 2020), and social media features that includes the detection of emojis. The second feature set is **SE**, that are non-contextual sentence embeddings from the Tamil pretrained model from fastText (Grave et al., 2018). The third and forth feature sets are **BF** and **RF**. These features are, respectively, sentence embeddings from multilingual BERT (Devlin et al., 2018) and multilingual RoBERTa (Conneau et al., 2019).

To obtain the sentence embeddings from BERT and RoBERTa, we fine-tuned them separately for each task with RayTune (Liaw et al., 2018). During this stage, 10 models with Tree of Parzen Estimators (TPE) (Bergstra et al., 2013) were trained to obtain the optimum values for the (1) weight decay, (2) batch size, (3) warm-up speed, (4) number of epochs, and (5) learning rate. TPE strategy selects the next hyperparameters using Bayesian reasoning and the expected improvement. Next, we extract the `[CLS]` token from the best models in a similar way as described in (Reimers and Gurevych, 2019).

Next, we train a neural network per feature set separately. We use these neural networks to build two classifiers based on ensemble learning. For this, we use Keras (TensorFlow) and RayTune for the hyperparameter stage. Besides, we train another neural network that combines all the feature sets at once using a knowledge integration strategy. For this, we fed each feature set in a separate hidden layer and then combine their outputs in the hidden layers.

The details of the hyperparameter optimisation stages are the following. As all feature sets are of a fixed size, we evaluate only MultiLayer Perceptrons (MLP) as the network architecture. These MLPs are divided into shallow and deep neural networks. This category is based on the number of hidden layers and the number of neurons per layer. Specifically, for the shallow neural networks we only try one or two hidden layers maximum. The number of neurons is the same in all layers. In deep neural networks, however, we try a larger number of hidden layers (between 3 and 8). Besides, the number of neurons per layer are arranged in different shapes (brick, triangle, diamond, rhombus, and short and long funnel). We also try several activation functions to connect the hidden layers as well as several learning rates and ratios of a dropout

mechanism. We also handle class imbalance evaluating larger batch sizes and class weights.

The best configuration for the knowledge integration strategy for subtask 1 is a deep neural network composed of 3 hidden layers, with 128 neurons stacked in a triangle shape. The batch size is 64, the dropout of .2, the learning rate is 0.001, and the activation function that connects the layers is a sigmoid. On the other hand, the best configuration for subtask 2 is a batch size of 32, no dropout, 4 hidden layers with 57 neurons stacked with a rhombus shape (the value of 57 is the max value of neurons per hidden layer), a learning rate of 0.001, and *selu* as activation function.

3 Results and discussion

Table 1 reports the results with the validation split for subtask 1 and 2. These results include each feature set separately, the knowledge integration strategy and two ensembles, one based on the mode of the predictions and another based on averaging the probabilities.

Subtask 1			
	precision	recall	f1-score
LF	17.84	15.35	13.70
SE	26.74	33.50	26.71
BF	29.25	29.26	28.84
RF	33.69	**35.29**	**33.74**
K.I.	**33.90**	32.85	32.36
mode	32.56	34.53	32.27
average	33.16	34.09	32.99
Subtask 2			
	precision	recall	f1-score
LF	8.56	6.73	5.40
SE	14.39	**16.30**	13.00
BF	13.67	14.12	13.38
RF	13.54	14.64	12.92
K.I.	13.97	14.29	13.33
mode	**15.10**	15.59	12.98
average	15.01	17.13	**15.12**

Table 1: Macro precision, recall and f1-score for the first and second subtask: LF stands for the Linguistic Features, SE stands for Sentence embeddings from fastText, BF and RF stands for Sentence embeddings from BERT and RoBERTa transformers, respectively. K.I. stands for knowledge integration strategy, and mode and average for the two ensemble strategies evaluated

As it can be observed from the first subtask (see Table 1 -top-), the best result for the models trained with only one feature set is achieved with the RF (XML RoBERTa). This result outperforms SE and BF. Besides, the performance of LF is more limited than the rest of the features based on embeddings. This is expected as the linguistic features (LF) is a small subset of features. The knowledge integration strategy (K.I.) achieves a macro average f1-score of 32.36. This f1-score is lower than the result achieved with RF used in isolation, which suggests that the combination of RF with other features within the same neural network downplays RF. We also check what is the macro f1-score of using other strategies for combining the features. We test two ensemble learning strategies, one based on the mode of the predictions, and another one based on averaging the probabilities of each class. The macro f1-score achieved is 32.274 for the mode, and 32.992 for averaging the predictions. These results are also lower than the ones achieved by RF in isolation.

As it can be observed from the first subtask (see Table 1 -bottom-), no larger difference between RF and BF is observed. In fact, RF achieves a lower score than BF and SE. The knowledge integration strategy reported a macro f1-score of 13.33%, which is better than the ensemble based on the mode but limited compared with the average of the predictions.

In view of these results, and as we only have one chance to submit our proposal, we decided to send the results with the knowledge integration strategy because, in our experience, it tends to produce better results with unseen test splits.

The classification report of the knowledge integration strategy for the first subtask is depicted in Table 2. Concerning the sentiments explored individually, we observed that *joy* and *ambiguous* are the emotions with higher score, with a f1-score of 57.63% and 57.38% respectively whereas *surprise* was the label with lower score, with a f1-score of 8.16%. These scores are related to the distribution of the labels, as documents labelled as *surprise* are underrepresented. However, documents labelled as *neutral*, which is the majority class, only achieves a f1-score of 47.73%. We calculate the confusion matrix (see Figure 2) to analyse this behaviour. Neutral documents are mismatched with the rest of the emotions. For instance, a 10% of the neutral documents are labelled as *joy*, and another 10% as *disgust*. This behaviour is not observed in the rest of the emotions. For example, documents labelled

as *love* are sometimes incorrectly labelled as *joy* (34%) or *trust* (13%). Moreover, the majority of wrong classifications with the emotions are related to the *neutral* class. In fact, a 42% of documents labelled as *surprise* are predicted as *neutral*.

	precision	recall	f1-score
ambiguous	58.65	56.17	57.38
anger	36.44	10.54	16.35
anticipation	27.39	29.50	28.41
disguist	22.37	34.60	27.17
fear	33.33	22.00	26.51
joy	56.69	58.59	57.63
love	18.10	24.28	20.74
neutral	49.49	46.08	47.73
sadness	34.44	43.94	38.61
surprise	6.94	9.92	8.16
trust	29.07	25.70	27.28
macro avg	33.90	32.85	32.36
weighted avg	43.05	41.74	41.81

Table 2: Classification report for the first subtask, with the validation split

Besides, in order to observe the correlation of the linguistic features with the class, we calculate the Information Gain. We observed that the most relevant linguistic features are related to positive emotions by the usage of certain emojis (0.03413). Regarding stylometric features, the most relevant ones are based on the average word length (0.03226) and concerning PoS features we found a important correlation between words that does not have defined grammatical gender.

The classification report for the second subtask is depicted in Table 3. The name of labels were translated using Google Translate. The macro f1-score is 13.329%. According to the individual emotions, the best result was achieved for *thanksgiving* (f1-score of 51.665%) and *admiration* (f1-score of 47.074%). The *neutral* documents achieved low performance (f1-score of 7.484%).

The results are also lower than the best of the feature sets in isolation: BF (macro f1-score of 13.38). In this case, the macro f1-score of combining the features using ensembles are 12.98% for the mode, and 15.12% for averaging the predictions of each model. As it can be observed, the result with the ensemble learning outperforms both the results achieved with BF and the combination of features into the same neural network. However, we decided to send the results using the same strat-

	precision	recall	f1-score
care	3.49	5.26	4.20
love	12.58	19.71	15.36
disgusting	12.73	17.83	14.85
desire	1.35	1.52	1.43
surprise	2.05	8.70	3.31
curiosity	15.89	16.99	16.42
realization of truth	26.77	22.33	24.35
excitement	7.17	8.95	7.96
hope for the future	6.18	11.59	8.06
anticipation	33.50	27.47	30.19
irritation	12.19	11.77	11.98
disappointment	4.32	10.99	6.20
approval	17.50	9.10	11.98
tease	28.05	26.26	27.13
grief caused by guilt	3.74	3.70	3.72
confusion	8.22	5.56	6.63
fun	5.57	10.00	7.15
anger	26.36	30.41	28.24
embarrassment	1.69	0.45	0.72
tragedy	23.40	5.14	8.43
grief	1.72	0.85	1.14
grief relief	0.78	0.92	0.84
neutral	13.24	5.22	7.48
thanksgiving	47.67	56.39	51.66
hope	10.08	8.13	9.00
anxiety	1.30	0.87	1.04
fear	2.44	13.64	4.13
pride	26.76	36.20	30.77
admiration	54.16	41.63	47.07
happiness	15.69	10.85	12.83
denial	6.48	14.43	8.95
macro avg	13.97	14.29	13.33
weighted avg	24.79	21.80	22.66

Table 3: Classification report for the second subtask

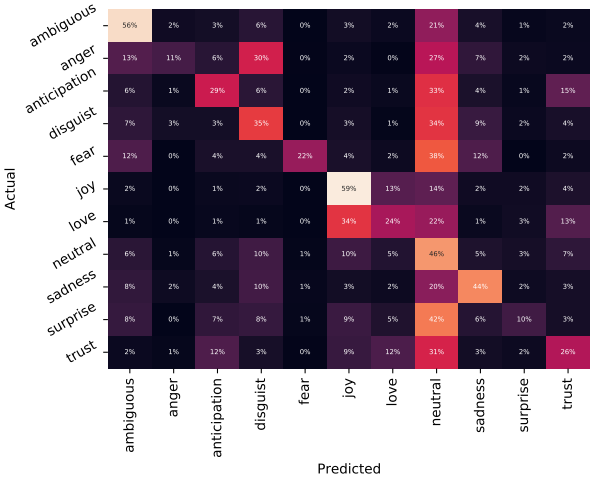

Figure 2: Confusion matrix for the first subtask

egy for both subtasks (as commented above, we could not receive any type of feedback using the CodaLab platform, which made the competition more challenging).

Next, we report the results for the official leader board. However, our participation was not considered for the first subtask. We suspect that the problem is related to a wrong format of the submission. It is worth mentioning that the results were not sending using the Codalab platform and we did not received feedback until the end of the evaluation phase.

Table 4 depicts the results for the second task, in which we achieve the first position, with a macro f1-score of 15.1% and improving the second best result (12.5%) in 0.026. Our system achieved the best precision and recall, being the most relevant the recall. This result is superior to the one achieved with the validation split. We assume, therefore, that the documents and their distribution in the validation and test sets are similar and that the performance of each label is similar.

team	precision	recall	f1-score
UMUTeam	15.0	17.1	15.1
GJG	14.2	14.4	12.5
Optimize_Prime	13.2	14.0	12.5
IIITSurat	15.6	9.9	9.0
Judith Jeyafreeda	9.4	6.8	5.7
GA	3.3	3.1	2.8
VCNVegetable	0.5	3.2	0.9

Table 4: Official results for the second task, sorted by rank. We include the macro averaged metrics of precision, recall and F1-score

4 Conclusions and further research lines

In this working notes we have described the participation of the UMUTeam in a shared task regarding emotion analysis in Tamil. We achieved the 1st position in a fine-grained emotion analysis classification in which 30 emotions can be defined. However, our results for the first multi-classification task were not reported due to an unknown error. We report our results for this task using the validation split. Our proposal to solve this problem was grounded on knowledge integration to combine linguistic features and different kind of sentence embeddings. As commented in the Introduction Section, we wanted to evaluate a subset of language-independent linguistic features in a non-Latin language. However, multilingual RoBERTa separately outperformed slightly the results of combining different feature sets with ensemble learning or knowledge integration.

As future work, we would like to extend the presented architecture by incorporating new feature extraction techniques to analyse their impact in precision. Furthermore, we will focus on interpretability techniques. Besides, regarding the application of emotions, we will evaluate the correlation of some linguistic features regarding anger and sadness with hate-speech in Spanish with the datasets published at (García-Díaz et al., 2022) and (García-Díaz et al., 2021a).

Acknowledgements

This work is part of the research project LaTe4PSP (PID2019-107652RB-I00) funded by MCIN/AEI/10.13039/501100011033. This work is also part of the research project PDC2021-121112-I00 funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. In addition, José Antonio García-Díaz is supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme.

References

James Bergstra, Daniel Yamins, and David Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

José Antonio García-Díaz, Mar Cánovas-García, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2021a. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506–518.

José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2021b. Psychographic traits identification based on political ideology: An

author analysis study on spanish politicians' tweets posted in 2020. *Future Generation Computer Systems*.

José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-Garcia. 2021c. Umuteam at emoevales 2021: Emosjon analysis for spanish based on explainable linguistic features and transformers.

José Antonio García-Díaz, Salud María Jiménez-Zafra, Miguel Angel García-Cumbreras, and Rafael Valencia-García. 2022. Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, pages 1–22.

José Antonio García-Díaz and Rafael Valencia-García. 2022. Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. *Complex & Intelligent Systems*, pages 1–14.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

MD Iqbal, Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H Sarker. 2022. Bemoc: A corpus for identifying emotion in bengali texts. *SN Computer Science*, 3(2):1–17.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

Flor Miriam Plaza-del Arco, Salud M Jiménez Zafra, Arturo Montejo Ráez, M Dolores Molina González, Luis Alfonso Ureña López, and María Teresa Martín Valdivia. 2021. Overview of the emoevales task on emotion detection for spanish at iberlef 2021.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Anbukkarasi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, Kogilavani Shanmugavadivel, Sajeetha Thavareesan, Sathiyaraj Thangasamy, Parameswari Krishnamurthy, Adeep Hande, Sean Benhur, and Santhiya Ponnusamy, Kishor Kumar Pandiyan. 2022. Findings of the shared task on Emotion Analysis in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Nilesh Shelke, Sushovan Chaudhury, Sudakshina Chakrabarti, Sunil L Bangare, G Yogapriya, and Pratibha Pandey. 2022. An efficient way of text-based emotion analysis from social media using lra-dnn. *Neuroscience Informatics*, page 100048.

Li Yong, Yang Xiaojun, Liu Yi, Liu Ruijun, and Jin Qingyu. 2022. A new emotion analysis fusion and complementary model based on online food reviews. *Computers & Electrical Engineering*, 98:107679.