# Exploring Language Independent Linguistic Features and Transformers in a Multi-label Emotion Detection Challenge in Urdu using Nastalīq Script

José Antonio García-Díaz[1], Manuel Valencia-García[1], Gema Alcaraz Mármol[2] and Rafael Valencia-García[1]

[1]*Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain*
[2]*Departamento de Filología Moderna, Universidad de Castilla-La Mancha, Spain*

Abstract
Emotion Analysis is a Natural Language Processing task whose objective is to obtain fine-grained emotions from a text. The understanding of emotions in written communication has applications in marketing, e-commerce and infodemiology among others. Besides, Emotion Analysis can be applied to identify threats that could represent a threat to citizens, from a Smart City perspective. In this working notes we describe the participation of the UMUTeam in the EmoThreat shared task, proposed at FIRE'2022 workshop. Out of the subtasks proposed, our team only participated in the main subtask, which consisted in a multi-label emotion classification based on Ekman's six basic emotions in documents written in Urdu using Nastalīq script. We achieved the second best result, from a total of 8 participants, achieving 66.9% of macro average F1-score. Our proposal combines in the same neural network four feature sets that include a subset of language independent linguistic features extracted from UMUTextStats, a non-contextual sentence embeddings from fastText and two contextual sentence embeddings from multilingual versions of BERT and RoBERTA.

**Keywords**
Feature Engineering, Deep-learning, Transformers, Linguistic Features, Natural Language Processing

## 1. Introduction

The proliferation of social media platforms has made it easier for people all over the world to communicate and share experiences. It has also provided benefits in international trade and improved public health policies, as social media posts can refer threats that potentially endanger citizens. Natural Language Processing (NLP) tools provide a useful way to process and to understand what the users want to express in an automatic manner. However, NLP tools have some challenges, highlight that some of the methods and state-of-the-art tools and datasets are based on English, and the fact that natural language is complex to understand, as it is highly subjective.

*FIRE 2022: Forum for Information Retrieval Evaluation, December 9–13, 2022, India*
joseantonio.garcia8@um.es (J. A. García-Díaz); manuelv@um.es (M. Valencia-García); Gema.Alcaraz@uclm.es (G. A. Mármol); valencia@um.es (R. Valencia-García)
0000-0002-3651-2660 (J. A. García-Díaz); 0000-0001-7703-3829 (G. A. Mármol); 0000-0003-2457-1791 (R. Valencia-García)

The organisers of the EmoThreat 2022 shared task [1] release a dataset for conducting multi-label emotion detection in written Urdu using Nastalīq script (written from right to left). Urdu language is spoken by more than 170 million people worldwide, highlighting India, Pakistan and Nepal. The underlying objective of this shared task is to understand public emotions from social networks applicable in NLP tools that can help to monitor events such as disasters, or to improve public policies in e-commerce and public health.

These working notes describe the participation of the UMUTeam at EmoThreat 2022 shared task [2], proposed in FIRE 2022 [3]. The main challenge in this shared task is a multi-label emotion classification task in Urdu [4]. The participants of the shared task are required to classify each document with one, or more of the Ekman's six basic emotions (plus one neutral emotion).

It is worth noting that our team has previous experience dealing with emotion classification. Specifically, we participate in the EmoEvalEs 2021 shared task [5], achieving the sixth position [6]. This shared task consisted in a multi-classification emotion detection with texts written in Spanish. However, this shared task allowed our team to continue validating subsets of language independent linguistic features tools that we have already applied in other languages such as Tamil [7].

The remainder of these working notes is organised as follows. First, Section 2 provides a review of related work focused in Urdu and emotion detection. Second, in Section 3, the developed pipeline for solving this task is described. Third, Section 4 includes the results achieved in the challenge and a comparison with the rest of runs submitted by our team and by the rest of the participants. Forth, Section 5 presents the findings obtained and it also includes some promising research lines.

## 2. Related work

The EmoThreat 2022 shared task is a continuation of a previously shared task [8]. In the previous edition of this shared task, the organisers proposed two challenges of abusive and threatening language detection in Urdu. The past shared task consisted only in binary classification tasks, one for detecting abusive documents (2400 documents for training, 1100 for testing) and another shared task for detecting threatening messages (6000 documents for training, 3950 documents for testing). The datasets were extracted from the micro-blogging platform Twitter. A total of 10 teams submitted their proposals for the abusive classification task and 9 for the threatening classification task. The best result was achieved with a F1-score value of 0.880 for Subtask A and 0.545 for Subtask B, using both run architectures based on Transformers.

Sentiment Analysis is another NLP field that has been explored in Urdu. Recent works such as the one described at [9] performed a multi-classification task from an Urdu dataset of 9312 reviews manually annotated compiled from user reviews about food, movies, apps, politics and sports. The dataset was annotated with three labels (positive, negative and neutral). The authors explored different baselines based on traditional machine-learning, deep-learning and models based on multilingual Transformers. Their experiments confirmed that multilingual BERT outperforms traditional models for Urdu, reaching an F1 score of 81.49%. In [10], the authors compile a dataset in Urdu for sentiment analysis and evaluate several traditional machine learning classifiers. The features were extracted using count-based techniques and pre-trained word embeddings from fastText. The authors found that the combination of features of these features outperformed the results achieved separately and compared with existing state-of-the-art approaches, reaching a F1-score of 82.05%. Another relevant work focused on Sentiment Analysis in Urdu is [11], in which the authors [11] evaluated several word embeddings using an architecture based on convolutional and recurrent neural networks, combined with traditional machine learning classifiers for the final classification. The authors evaluate their proposal with four corpora. Among the different architectures evaluated, the authors achieve their best results using a classifier based on Support Vector Machines and features based on Word2Vec based on Continuous Bag of Words.

## 3. Methodology

The first step of our pipeline is to explore the dataset and to create a custom validation split. The validation split is created using a stratified sample in a ratio of 80-20. In Table 1 it can be observed that there is an important imbalance among the emotions, being *fear*, *anger*, and *disgust* underrepresented.

**Table 1**
Dataset distribution of the training and custom validation splits

| sentiment | train | val | total |
|---|---|---|---|
| anger | 656 | 155 | 811 |
| disgust | 616 | 145 | 761 |
| fear | 495 | 114 | 609 |
| happiness | 841 | 205 | 1046 |
| neutral | 2412 | 602 | 3014 |
| sadness | 1760 | 430 | 2190 |
| surprise | 1246 | 304 | 1550 |
| total | 8026 | 1955 | 9981 |

As we deal with a multi-label classification challenge, we analyse the co-occurrence of emotions (see Figure 1). As it can be observed, *anger* and *disgust* are the two sentiments that usually appear together in the same tweet. It is also noticed that the *neutral* class is not used combined with other sentiments.

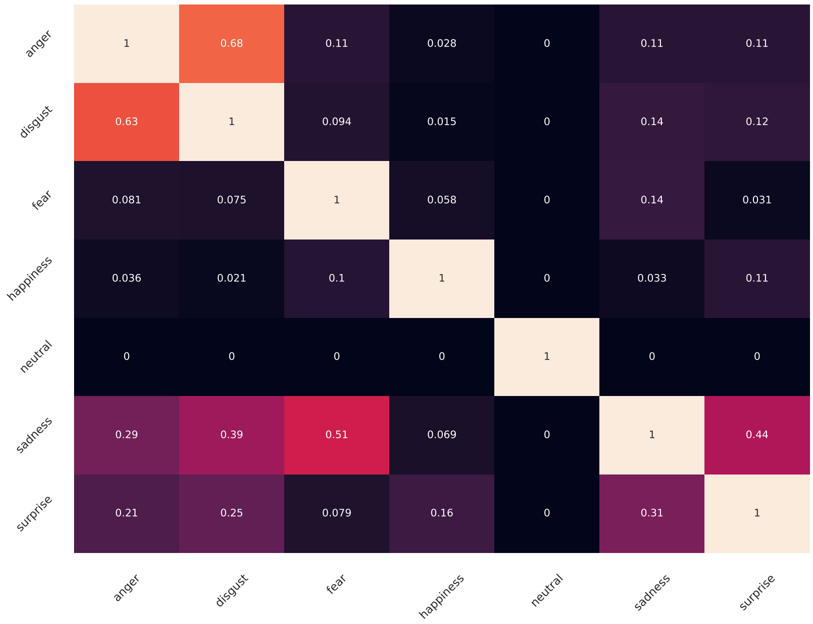
The next step in our pipeline is the feature extraction. Four different feature sets are involved in our participation. The first feature set is a subset of language independent linguistic features extracted with UMUTextStats **(LF)** [12, 13, 14, 15]. The second feature set are non-contextual sentence embeddings from FastText **(SE)** [16]. The third and forth feature sets correspond to multilingual contextual embeddings from BERT **(BF)** [17] and RoBERTa **(RF)** [18].

To obtain the contextual sentence embeddings from BF and RF we do hyperparameter tuning. A total of 20 transformers models (10 for BF, 10 for RF) were trained using the EmoThreat training split, and deciding which the best model is by using our custom validation split. From the best model, we extracted `[CLS]` token [19]. The hyperparameters involved in this process are 1) the weight decay, 2) the batch size, 3) the warm-up speed, 4) the number of epochs, and the 5) learning rate. The combination of these hyperparameters is performed using Tree of Parzen Estimators (TPE) [20].

Once all feature sets had been extracted, we evaluated two strategies for combining the strengths of each one. The first strategy is called Knowledge Integration **(KI)**, and consists in training a multi-input deep-learning model that combines all feature sets at once. The second strategy involves ensemble learning **(EL)**, which combines the predictions of models focused on one specific feature set. For this, we obtain a model for each feature set using hyperparameter tuning (described below) and then, we evaluate two ways to use ensemble learning. The first strategy is soft voting, which consists in calculating the mode of the predictions. The second strategy corresponds to average probabilities, which consists in averaging the probabilities predicted of each individual model to generate the final prediction.

Regardless the training of the KI or the ensemble learning, we perform a hyperparameters tuning stage. The hyperparameters involved are the shape of the network (that is, the number of neurons and the number of hidden layers), the dropout mechanism, the learning rate and several activation functions. The results of the hyperparameter tuning stage can be found in Table 2.

In all cases, the best result is achieved with shallow neural networks (that is, neural networks with only one or two hidden layers, and the same number of neurons in all layers). Besides, except for RF, all experiments achieved better results with a small dropout rate of .1. The learning rate varies, being 0.001 for SE, RF and KI. In case of the activation function, all experiments achieved better results with non-linear activation functions except LF.



| | anger | disgust | fear | happiness | neutral | sadness | surprise |
|---|---|---|---|---|---|---|---|
| anger | 1 | 0.68 | 0.11 | 0.028 | 0 | 0.11 | 0.11 |
| disgust | 0.63 | 1 | 0.094 | 0.015 | 0 | 0.14 | 0.12 |
| fear | 0.081 | 0.075 | 1 | 0.058 | 0 | 0.14 | 0.031 |
| happiness | 0.036 | 0.021 | 0.1 | 1 | 0 | 0.033 | 0.11 |
| neutral | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| sadness | 0.29 | 0.39 | 0.51 | 0.069 | 0 | 1 | 0.44 |
| surprise | 0.21 | 0.25 | 0.079 | 0.16 | 0 | 0.31 | 1 |

**Figure 1:** Label co-occurrence

## 4. Results and analysis

First, we report in Table 3 the macro average results achieved by each feature set (for the EL strategy) and the KI strategy. It can be observed that the best results achieved separately are obtained with RF. However, when combined with the rest of the feature sets, the recall is higher and the precision is lower. As it is expected, the results achieved by LF are limited, as they are based on stylometry and PoS features. It draw out attention the limited recall of the embeddings based on BERT compared with the embeddings based on RoBERTa. As we deal with classification tasks, we consider that this difference is not related to the tasks in which these models has been trained (Next Sentence Prediction and Masked Language Model), but with the tokenizer and the dataset used to learn the embeddings.

Next, we report in Table 4 the results per emotion achieved with the KI strategy using the custom validation split. It can be observed that the model reaches almost a perfect score concerning documents without attached emotions. In non-neutral documents, all emotions achieve similar scores. *Sadness* reaches the best f1-score and *happiness* gets very good precision but limited recall.

The results of the official leader board are reported in Table 5. The results are ranked using the Macro F1 score. The rest of the evaluated metrics are the multi-label accuracy, the Hamming loss and the micro and weighted versions of the F1-score. We achieve the second best position with our run based on KI.

The results of our three runs are depicted in Table 6. As it can be observed, the results achieved with ensemble learning are more limited in all metrics. The reason for this is that

**Table 2**
Best hyper-parameters for each feature set trained separately and combined using knowledge integration.

| Feature set | shape | # of layers | neurons | dropout | lr | activation |
|---|---|---|---|---|---|---|
| LF | brick | 1 | 48 | .1 | 0.010 | linear |
| SE | brick | 2 | 256 | .1 | 0.001 | relu |
| BF | brick | 2 | 256 | .1 | 0.010 | tanh |
| RF | brick | 1 | 16 | - | 0.001 | relu |
| KI | brick | 2 | 48 | .1 | 0.001 | sigmoid |

**Table 3**
Macro average precision, recall and f1-score of each feature set and the Knowledge Integration strategy using custom validation split.

| | precision | recall | f1-score |
|---|---|---|---|
| LF | 45.155 | 47.898 | 41.162 |
| SE | 64.728 | 57.551 | 60.646 |
| BF | 67.104 | 58.236 | 60.274 |
| RF | **71.821** | 64.585 | 67.422 |
| KI | 71.482 | **65.440** | **67.441** |

the majority of correct predictions are performed by the RF feature set and the contribution of the rest of the feature sets dismisses the performance of the model applying ensemble learning strategies.

## 4.1. Error Analyses

For the error analysis we get our best run and collect the wrong predictions with the test split. Next, we sort the multi-label output by euclidean distance in order to get the predictions with the higher number of wrong labels. We obtained that the wrong classifications represent the

**Table 4**
Classification report for the Knowledge Integration strategy using custom validation split.

| | precision | recall | f1-score |
|---|---|---|---|
| anger | 58.378 | 69.677 | 63.529 |
| disgust | 56.051 | 60.690 | 58.278 |
| fear | 60.000 | 60.526 | 60.262 |
| happiness | 80.734 | 42.927 | 56.051 |
| neutral | 100.000 | 99.003 | 99.499 |
| sadness | 74.346 | 66.047 | 69.951 |
| surprise | 70.866 | 59.211 | 64.516 |
| micro avg | 78.587 | 72.276 | 75.300 |
| macro avg | 71.482 | 65.440 | 67.441 |
| weighted avg | 78.915 | 72.276 | 74.807 |
| samples avg | 75.108 | 73.725 | 72.996 |

**Table 5**
Official leader-board

| Rank | Team | Accuracy | Weighted F1 | Micro F1 | Macro F1 | Hamming loss |
|---|---|---|---|---|---|---|
| 1 | FOSUNlpTeam | 63.6 | 75.9 | 75.9 | 68.7 | 8.80 |
| 2 | **UMUTeam** | 61.6 | 74.3 | 74.9 | 66.9 | 8.80 |
| 3 | hate-alert | 61.2 | 70.9 | 72.4 | 61.5 | 9.20 |
| 4 | MUCS | 58.2 | 69.6 | 69.2 | 60.3 | 11.30 |
| 5 | ERTIM | 59.3 | 69.9 | 72.0 | 59.9 | 9.18 |
| 7 | SakshiEmo2022 | 38.5 | 61.1 | 47.7 | 46.6 | 34.00 |
| 8 | Aces | 42.6 | 38.1 | 45.8 | 24.0 | 16.90 |

**Table 6**
Results per run

| Run | Accuracy | Weighted F1 | Micro F1 | Macro F1 | Hamming loss |
|---|---|---|---|---|---|
| 1 | 0.616 | 0.743 | 0.749 | 0.669 | 0.088 |
| 2 | 0.570 | 0.670 | 0.704 | 0.565 | 0.091 |
| 3 | 0.602 | 0.714 | 0.734 | 0.624 | 0.087 |

39.18% of total test split. 129 documents get one wrong label, 529 two wrong labels, 92 three wrong labels and 14 four wrong labels.

Next, we present the most notable failures. It is worth noting that the texts presented here are translated using Google Translator. The most distant classifications made by our system are those in which our system could not be able to identify any emotion. That is the case of: 1) *You interpret me very well. You hate Maulana Tariq Jameel Sahib. Fear Allah. Those holy persons should respect him.*, and 2) *Even if you express the pain in a happy way, the pain will still hurt.*. For the first sentence, we consider that the problem is that the application does not have enough context to understand the sentence. For the second sentence, we consider that the text does not express any emotion but a refrain. The case of 3) *The pain and sadness of Imran Niazi on the death of such a close friend of Imran Niazi is not seen, or even Imran Niazi* is just the opposite. This document was rated as *anger*, *sadness*, and *surprise*, but the annotators did not find any emotion in the document. Besides, we identified other documents related to SARS-Covid 2019 diseases. That is the case of 4) *Those who ask for permission to open shops are not afraid of Corona. Watch the program Live with Nasrullah Malik only New*, and 5) *Sami Ibrahim sir, what are you most afraid of Corona till now? Imran Ahmed Khan Niazi still scares me the most.* Besides, there are other errors with short texts. That is the case of *I hate this game.* In this case, our system correctly predicted the *anger* emotion, but missclassified *disgust* with *sadness*, which can be considered a minor mistake.

## 5. Conclusions

We achieved the second position in a multi-label classification task in Urdu (66.9% of macro F1-score), in which our pipeline is based on the combination of a subset of independent linguistic features and transformers. Our best result combines the features using a knowledge integration strategy; however, the runs submitted with ensemble learning achieved limited results, losing several positions in the official ranking. Although we are very happy with our participation, as we have evaluated our tools with non-Latin languages, we could not participate in the second subtask of the competition due to lack of time. The source code is available at: https://github.com/Smolky/umuteam-emothreat-2022

As promising future research lines, we would include nested cross validation to prevent the hyperparameter tinning stages from being biased to the custom validation split and we will apply data augmentation to increase the number of instances and reduce the effects of class imbalance. We also explore the reliability of using transformers focused on Urdu rather than multilingual. Besides, we will include features concerning figurative language [21], as its identification may increase the generalisation of emotion analysis detectors. Another research line is to apply emotion analysis to authors profiling tasks. In this sense, we are planning to extend the PoliticES 2022 shared task [22] to compile tweets from politicians and journalist and to extract emotions per author profile. For this, we will use the UMUCorpusClassifier tool [23].

## Acknowledgments

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 projects AIInFunds (PDC2021-121112-I00) funded by MCIN/AEI/10.13039/501100011033, by the European Union NextGenerationEU/PRTR, LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, and by "Programa para la Recualificación del Sistema Universitario Español 2021-2023". In addition, José Antonio García-Díaz is supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme.

## References

[1] S. Butt, M. Amjad, F. Balouchzahi, N. Ashraf, R. Sharma, G. Sidorov, A. Gelbukh, Overview of EmoThreat: Emotions and Threat Detection in Urdu at FIRE 2022, in: CEUR Workshop Proceedings, 2022.

[2] N. Ashraf, L. Khan, S. Butt, H.-T. Chang, G. Sidorov, A. Gelbukh, Multi-label emotion classification of urdu tweets, PeerJ Computer Science 8 (2022) e896.

[3] S. Butt, M. Amjad, F. Balouchzahi, N. Ashraf, R. Sharma, G. Sidorov, A. Gelbukh, EmoThreat@FIRE2022: Shared Track on Emotions and Threat Detection in Urdu, in: Forum for Information Retrieval Evaluation, FIRE 2022, Association for Computing Machinery, New York, NY, USA, 2022.

[4] I. Ameer, N. Ashraf, G. Sidorov, H. Gómez Adorno, Multi-label emotion classification using content-based features in twitter, Computación y Sistemas 24 (2020) 1159–1164.

[5] F. M. Plaza-del Arco, S. M. Jiménez-Zafra, A. Montejo-Ráez, M. D. Molina-González, L. A. Ureña-López, M. T. Martín-Valdivia, Overview of the emoevales task on emotion detection for spanish at iberlef 2021, Procesamiento del Lenguaje Natural 67 (2021) 155–161.

[6] J. A. García-Díaz, R. C. Palacios, R. Valencia-García, Umuteam at emoevales 2021: Emotion analysis for spanish based on explainable linguistic features and transformers, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021, volume 2943 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 59–71. URL: http://ceur-ws.org/Vol-2943/emoeval_paper6.pdf.

[7] J. García-Díaz, M. Á. R. García, R. Valencia-García, Umuteam@ tamilnlp-acl2022: Emotional analysis in tamil, in: Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, 2022, pp. 39–44.

[8] M. Amjad, A. Zhila, G. Sidorov, A. Labunets, S. Butt, H. I. Amjad, O. Vitman, A. Gelbukh, Urduthreat@ fire2021: Shared track on abusive threat identification in urdu, in: Forum for Information Retrieval Evaluation, 2021, pp. 9–11.

[9] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, Multi-class sentiment analysis of urdu text using multilingual bert, Scientific Reports 12 (2022) 1–17.

[10] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, A. Gelbukh, Urdu sentiment analysis with deep learning methods, IEEE Access 9 (2021) 97803–97812.

[11] L. Khan, A. Amjad, K. M. Afaq, H.-T. Chang, Deep sentiment analysis using cnn-lstm architecture of english and roman urdu text shared in social media, Applied Sciences 12 (2022) 2694.

[12] J. A. García-Díaz, P. J. Vivancos-Vicente, Á. Almela, R. Valencia-García, Umutextstats: A linguistic feature extraction tool for spanish, in: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 6035–6044. URL: https://aclanthology.org/2022.lrec-1.649.

[13] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020, Future Generation Computer Systems 130 (2022) 59–74.

[14] J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, R. Valencia-García, Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers, Complex & Intelligent Systems (2022) 1–22.

[15] J. A. García-Díaz, R. Valencia-García, Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers, Complex & Intelligent Systems 8 (2022) 1723–1736.

[16] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, CoRR abs/1802.06893 (2018). URL: http://arxiv.org/abs/1802.06893. arXiv:1802.06893.

[17] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[18] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[19] N. Reimers, I. Gurevych, 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, 2019, pp. 3982–3992. URL: https://arxiv.org/abs/1908.10084.

[20] J. Bergstra, D. Yamins, D. Cox, Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures, in: International conference on machine learning, PMLR, 2013, pp. 115–123.

[21] M. del Pilar Salas-Zárate, G. Alor-Hernández, J. L. Sánchez-Cervantes, M. A. Paredes-Valverde, J. L. García-Alcaraz, R. Valencia-García, Review of english literature on figurative language applied to social networks, Knowledge Information Systems 62 (2020) 2105–2137. URL: https://doi.org/10.1007/s10115-019-01425-3. doi:10.1007/s10115-019-01425-3.

[22] J. A. García-Díaz, S. M. J. Zafra, M. T. M. Valdivia, F. García-Sánchez, L. A. U. López, R. Valencia-García, Overview of politices 2022: Spanish author profiling for political ideology, Proces. del Leng. Natural 69 (2022) 265–272. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6446.

[23] J. A. García-Díaz, Á. Almela, G. Alcaraz-Mármol, R. Valencia-García, Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks, Proces. del Leng. Natural 65 (2020) 139–142. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6292.