UMUTeam at REST-MEX 2022: Polarity Prediction using Knowledge Integration of Linguistic Features and Sentence Embeddings based on Transformers

José Antonio García-Díaz[1,*], Miguel Ángel Rodríguez-García[2,*], Francisco García-Sánchez[1,*] and Rafael Valencia-García[1,*]

[1] *Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain*
[2] *Departamento de Ciencias de la Computación, Universidad Rey Juan Carlos, 28933 Madrid, Spain*

Abstract

Tourism is a crucial activity for several countries in which it is the core of economic, social, and cultural activities. However, during disasters such as outbreaks of infectious diseases, the activities related to tourists are limited or even have to stop completely. The second edition of the REST-MEX (IberLEF 2022) is focused on knowing if the tasks related to Natural Language Processing can help mitigate this situation. In this shared task, the organisers challenge the participants to create a recommendation system, a polarity prediction system based on Sentiment Analysis, and a risk detection system. Our team only participated in the Sentiment Analysis Track, achieving the first position. This track consists of determining the polarity on a scale between 1 and 5 and determining the activity being reviewed. For solving both subtasks, we build a neural network that combines the strengths of linguistic features and three feature sets based on sentence embeddings using a Knowledge Integration strategy.

Keywords

Sentiment Analysis, Feature Engineering, Knowledge Integration, Ensemble Learning

CEUR-WS.org/Vol-3202/restmex-paper13.pdf

1. Introduction

These working notes describe the participation of the UMUTeam in the REST-MEX 2022 shared task [1], focused on improving tourism activities through Natural Language Processing tasks. Tourism promotes social, cultural, and economic relationships, and it is essential in daily life, both for personal life, environment, and business relationships. Tourism activities are vital in several countries, such as Mexico, representing nearly 9% of its national GDP, generating around 4.5 million direct jobs. However, the effects related to the COVID 2019 pandemic [2] affected several sectors in Mexico, including touristic products and services [3].

This shared task aims to prove that Natural Language Processing (NLP) tasks can be applied to help restore tourism by monitoring social networks and detecting problems as soon as possible. For example, NLP can be applied to identify the polarity of tourists' opinions. Specifically, three

IberLEF 2022, September 2022, A Coruña, Spain
joseantonio.garcia8@um.es (J. A. García-Díaz); miguel.rodriguez@urjc.es (M. Rodríguez-García); frgarcia@um.es (F. García-Sánchez); valencia@um.es (R. Valencia-García)
0000-0002-3651-2660 (J. A. García-Díaz); 0000-0001-6244-6532 (M. Rodríguez-García); 0000-0003-2667-5359 (F. García-Sánchez); valencia@um.es (R. Valencia-García)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

tracks related to tourism are proposed: (1) developing a Recommender System for tourism, considering affinity between users' profiles and places' descriptions; (2) a Sentiment Analysis task on tourism, where, in the last years, it has gained academic and commercial relevance, especially on the English language. However, this task is focused on Mexican-Spanish without dialects, and it has been mainly designed to analyse sentences and assign values between 1 and 5 that express their polarity; (3) the Epidemiological semaphore is a system which consists of 4 colours, and its target is to determine which activities are allowed according to the severity of the COVID pandemic [4].

The dataset was compiled from TripAdvisor between 2002 and 2021. Each opinion was labelled as an integer between [1, 5], where 1 represents the most negative polarity, and 5 is the most positive. The dataset contains 43,150 documents, which 12,938 of them are the test set. For the rest, we picked up 7553 documents for custom validation. The label distribution of such dataset is the following: 547 opinions were labelled with 1 point, 730 with 2 points, 2121 with 3 points, 5878 with 4 points, and 20936 with 5 points. For the second subtask, 16565 opinions were collected from hotels, 8450 from restaurants, and 5197 from sights.

2. Methodology

In a nutshell, we conduct the Sentiment Analysis track as two separate challenges. On the one hand, the polarity classification problem is handled as a regression problem. On the other hand, the target of the opinion is resolved as a multi-classification problem.

The first step in our pipeline is text cleaning. We create a cleaned version of each text, removing social media jargon, emojis, and correcting misspellings. However, we kept the original version of the documents to extract particular features related to correction and style.

The second step is the extraction of several feature sets. The first feature set is composed of 389 linguistic features (LF) extracted with the UMUTextStats tool [5, 6, 7]. The linguistic features are organised as (1) phonetics, (2) morphosyntax, (3) correction and style, (4) semantics, (5) pragmatics and figurative language, (6) stylometry, (7) lexis, (8) psycho-linguistic processes, (9), and (10) social media jargon. The second feature set are non-contextual sentence embeddings (SE) from the Spanish model FastText [8]. The third and fourth feature sets are contextual sentence embeddings based on two Transformers: BERT (BF) and RoBERTa (RF) [9, 10].

In the case of BERT and RoBERTa, we conduct a fine-tuning approach combined with hyper-parameter optimisation. For this, we use RayTune [11] to evaluate 10 BERT and 10 RoBERTa models. The selection of the best parameters is based on Tree of Parzen Estimators (TPE) [12], which is a strategy that selects the next hyper-parameter combination using Bayesian reasoning and the expected improvement. The evaluated parameters are (1) weight decay, (2) the batch size, (3) the warm-up speed, (4) the number of epochs, and (5) the learning rate. For the best BERT and RoBERTa model, we extract their sentence embeddings based on the `[CLS]` token [13].

For the regression subtask, the neural network models are trained using RMSE as a loss function. The dataset does not contain outliers, and we do not want to distinguish between positive and negative errors. For the classification subtask, we based our loss function on binary cross-entropy.

Table 1
Hyper-parameters of each feature set. The hyperparameters are the shape of the neural network, the number of hidden layers and neurons, the dropout rate, the learning rate and the activation function

	shape	layers	neurons	dropout	learning rate	activation
			Subtask 1			
LF	lfunnel	7	128	.2	0.001	selu
SE	brick	2	64	.1	0.001	sigmoid
BF	brick	1	256	.2	0.001	sigmoid
RF	3angle	7	37	-	0.001	elu
KI	lfunnel	4	16	.2	0.001	elu
			Subtask 2			
LF	brick	2	256	.2	0.001	relu
SE	brick	2	16	.2	0.010	relu
BF	brick	2	48	.3	0.001	relu
RF	diamond	3	37	.3	0.001	tanh
KI	brick	1	8	-	0.001	tanh

We conduct other hyper-optimisation stages for each feature set. All neural networks are Multi-Layer Perceptrons (MLP) as we deal only with fixed-size feature sets. The main parameters of the MLP are the shape, the number of layers, and the number of neurons in the first layer. If the number of hidden layers is small, all the hidden layers have the same number of neurons. In contrast, with large numbers of hidden layers (between 3 and 8), we evaluate to put a different number of neurons per layer. We assess the following shapes: brick, triangle, diamond, rhombus, and short and long funnel. Apart from the MLP architecture, we evaluate several activation functions, learning rates and dropout mechanisms. Table 1 reports the best hyper-parameter combination for each feature set and the Knowledge Integration (KI) strategy for both parts of the task.

As can be observed from Table 1, the best results for the polarity prediction are achieved with deep-learning models in the case of LF, RF, and KI. SE and BF, however, worked better with shallow neural networks. The best learning rate, in all cases, is 0.001, and the activation function varies from one feature set to another, being sigmoid better for SE and BF, selu for LF, and elu for RF and KI. For the touristic asset classification being reviewed (subtask 2). We observe that simpler models based on shallow neural networks performed better for LF, SE, BF and KI, and only for RF the best results are obtained with a deep-learning model of three hidden layers.

3. Results and discussion

As commented earlier, the Sentiment Analysis track consists of a polarity classification and asset classification. A polarity is a natural number between 1 and 5, and it is evaluated with Mean Absolute Error (MAE). There are three possible labels for the asset classification: Attractive, Hotel, and Restaurant. The overall score of the Sentiment Analysis track is the average of the inverse of MAE and the Macro F1.

As commented earlier, we faced both problems separately. The results achieved with the

Table 2
Results for the first task with our custom validation split. We report the Explained Variance (EV), the Mean Squared Log Error, the coefficient of determination (R2), the Mean Absolute Error (MAE), the Mean Squared Error (MSE) and the Root Mean Squared Error (RMSE)

	EV	MSLR	R2	MAE	MSE	RMSE
LF	0.469	0.023	0.468	0.420	0.411	0.641
SE	0.563	0.019	0.563	0.387	0.338	0.581
BF	0.696	0.013	0.695	0.301	0.236	0.485
RF	0.713	0.012	0.713	**0.291**	0.222	0.471
KI	**0.715**	**0.012**	**0.714**	0.327	**0.221**	**0.471**
Ensemble	0.677	0.014	0.677	0.330	0.250	0.500

Table 3
Results for the second task with our custom validation split

	precision	recall	f1-score
Attractive	99.843	99.454	99.648
Hotel	99.295	99.343	99.319
Restaurant	98.520	98.657	98.588
macro avg	99.219	99.151	99.185
weighted avg	99.167	99.166	99.166

custom validation split for the regression subtask are depicted in Table 2.

As it can be observed from Table 2, the LF achieves limited results compared with sentence embeddings (SE, BF, RF). When combining the strengths of each feature set, we observed that the KI strategy performs better in regards to every metric except for MAE, in which RF achieves better results (0.291 vs 0.327).

Next, Table 3 reports the classification report with the KI strategy for solving the classification subtask. We can observe that the performance of this model is almost perfect, which suggests the classification task is relatively trivial; that is, the text contains explicit references that state whether a comment is directed to a hotel, a restaurant, or a tourist attraction.

The official leader board is depicted in Table 4. We achieved the first position in the Sentiment Analysis track. Our score is 0.892, the average between the Polarity classification (0.258, MAE) and review target classification (99.010, Macro F1-Score). The average results of the rest of the participants are 82.484%, with a standard deviation of 0.99. We achieve this result with the knowledge integration strategy. With our second run, based on ensemble learning based on averaging the predictions, we achieved a score of 0.885.

4. Conclusions and further work

These working notes summarise the participation of the UMUTeam in one of the three challenging tasks thrown by the shared task REST-MEX at IberLEF 2022. We have participated only on the Sentiment Analysis Track, where our system has obtained the best performance in the final ranking. The challenge has been faced from two different perspectives: a regression problem in the case of the assignment of polarity; and a multi-classification problem in identifying the

opinion target. To accomplish the task, we have defined a pipeline mainly composed of three stages: (1) text cleaning, where the documents are pre-processed to correct misspellings and remove nonsense words such as emojis and hashtags, among others; (2) feature extraction, where a combination of a different nature is utilised to extract linguistic features and different types of sentence embeddings; (3) classification model, a neural networks modules is ubicated at the end of the pipeline to accomplish the classification. Between various combinations established, the knowledge integration strategy achieved the best overall result in the Sentiment Analysis Track, obtaining a final score of 0.892 that combines the MAE of the polarity classification and a macro F1-score of 98.964 in the review target classification task.

We observed that the majority of incorrect predictions for subtask 1 are not more than one point apart. For example, a wrong classification of our system is about the Mummies of Guanajuato[1]. The writer of the review assigned 5 stars in TripAdvisor; however, our system rate the review with 3 stars. This review had several references to death. Besides, writer can use different narrative devices to make their reviews interesting. Specifically, this author states that initially the visit to the museum gives me doubts. Besides, the author also alludes to other unfavorable reviews from other visitors. In fact, there are very interesting aspects to learn

[1]https://en.wikipedia.org/wiki/Mummiesc_of_Guanajuato

Table 4
Official leader-board

Team	Score	MAE (Polarity)	M. F1 (Attraction)
UMU-Team-Run-1	**0.892**	**0.258**	**0.990**
UC3M-Run1	0.891	0.261	0.988
CIMAT MTY-GTO-Run1	0.890	0.264	0.989
MCE_Team-Run2	0.889	0.267	0.989
MCE_Team-Run1	0.887	0.269	0.986
UMU-Team-Run-2	**0.886**	**0.279**	**0.989**
GPI_CIMAT-Run1	0.885	0.267	0.982
CIMAT2020_beto-Run1	0.883	0.270	0.978
DCI-UG-Run1	0.875	0.270	0.963
UCI-UC-CUJAE-Run2	0.872	0.305	0.978
UCI-UC-CUJAE-Run1	0.869	0.305	0.972
CIMAT2020-Run2	0.869	0.315	0.978
DCI-UG-Run2	0.866	0.300	0.963
ESCOM-IPN-IIA_run2	0.860	0.322	0.963
GPI_CIMAT-Run1	0.844	0.288	0.912
ESCOM-IPN-LCD_run2	0.840	0.353	0.941
ESCOM-IPN-IIA_run1	0.834	0.345	0.925
UPTC_UDLAP-Run1	0.827	0.448	0.964
SENA Team	0.803	0.471	0.926
DevsExMachina-Run2	0.704	0.636	0.796
DevsExMachina-Run1	0.667	0.971	0.826
ESCOM-IPN-LCD_run1	0.596	0.856	0.652
UPTC_UDLAP-Run2	0.542	0.548	0.438
Majority Class (baseline)	0.457	0.476	0.236

from this review. First, the final score assigned by our system should pay more attention to the conclusions. Second, some negative feelings, such as fear, do not have to have a negative connotation in the context of certain attractions or cultural events. Third, our system should distinguish more clearly what events allude to distant or hypothetical events from real events. Forth, in reviews it seems to be more common to spend more time describing problems and drawbacks, even if the final conclusions are positive.

As future work, we will focus on improving the training of the neural networks. One drawback of our approach is that we divide the training split in training and validation, so the resulting neural networks are biased to this random split. In this sense, we are developing methods to perform cross-validation in reasonable time. Another improvement is the evaluation of multi-task learning approaches in order to observe if it is beneficial to train both subtasks at the same time. Besides, the usage of sentence embeddings and linguistic features in texts of medium size could ignore some relevant facts. To solve this issue, we will evaluate to calculate the sentence embeddings by sentence and then averaging them weighting different parts of the whole paragraph, such as the topic, body sentences, bridge sentences and the concluding sentences.

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, 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] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cárdenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Rodríguez-González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism, Procesamiento del Lenguaje Natural 67 (2021). doi:https://doi.org/10.26342/2021-67-14.

[2] M. A. Álvarez-Carmona, R. Aranda, A. Y. Rodrıguez-González, L. Pellegrin, H. Carlos, Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news, Journal of Information Sciences (2022). doi:https://doi.org/10.1177/01655515221100952.

[3] G. Ozbay, M. Sariisik, V. Ceylan, M. Çakmak, A comparative evaluation between the impact of previous outbreaks and covid-19 on the tourism industry, International Hospitality Review (2021). doi:https://doi.org/10.1108/IHR-05-2020-0015.

[4] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Rodríguez-González, D. Fajardo-Delgado, R. Guerrero-Rodríguez, L. Bustio-Martínez, Overview of rest-mex at iberlef 2022:

Recommendation system, sentiment analysis and covid semaphore prediction for mexican tourist texts, Procesamiento del Lenguaje Natural 69 (2022).

[5] 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 (2022) 1–14. doi:https://doi.org/10.1007/s40747-021-00625-1.

[6] 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. doi:https://doi.org/10.1016/j.future.2021.12.011.

[7] 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. doi:https://doi.org/10.1007/s40747-022-00693-x.

[8] 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.

[9] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained bert model and evaluation data, PML4DC at ICLR 2020 (2020).

[10] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022) 39–60. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405.

[11] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, I. Stoica, Tune: A research platform for distributed model selection and training, arXiv preprint arXiv:1807.05118 (2018).

[12] 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. URL: https://proceedings.mlr.press/v28/bergstra13.html.

[13] 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. URL: https://arxiv.org/abs/1908.10084.