UMUTeam at DA-VINCIS 2022: Aggressive and Violent Classification using Knowledge Integration and Ensemble Learning

José Antonio García-Díaz[1], Salud María Jiménez-Zafra[2], Miguel Ángel Rodríguez-García[3] and Rafael Valencia-García[1]

[1]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain
[2]Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain
[3]Computer Science Department, Universidad Rey Juan Carlos, 28933 Madrid, Spain

Abstract

This paper presents the contribution of the UMUTeam for the DA-VINCIS shared task organized at IberLEF 2022, as part of the SEPLN conference. The objective of the task is to identify violent incidents in Twitter, for which two subtaks are proposed: i) violent event identification, and ii) violent event categorization. We have addressed both subtasks exploring different strategies for combining linguistic features and embeddings from Transformers. Our team has been ranked fourth in subtask 1 and sixth in subtask 2, with an F1-score of 76.4 and 44.8, respectively.

Keywords

Aggressive and Violent Classification, Feature Engineering, UMUTextStats, Negation Processing, Natural Language Processing

1. Introduction

Social media has become an important source for acquiring opinions and information about events [1]. In the context of defense and security, violent events have a high impact, as governments are responsible for ensuring the security of their population [2]. In this sense, social networks could be monitored to quickly detect violent events based on real-time posts made by users. For this, the development of automatic solutions based on Natural Language Processing is crucial.

The DA-VINCIS shared task [3], *Detection of Aggressive and Violent Incidents from Social Media in Spanish*, organized at IberLEF 2022, arises with the aim of promoting the development of automatic models to determine whether a news item obtained from Twitter describes a violent incident or not. Specifically, it focuses on the processing of Spanish tweets and it proposes two subtasks: *Violent event identification*, and *Violent event category recognition.*

IberLEF 2022, September 2022, A Coruña, Spain.
joseantonio.garcia8@um.es (J. A. García-Díaz); sjzafra@ujaen.es (S. M. Jiménez-Zafra); miguel.rodriguez@urjc.es (M. Á. Rodríguez-García); valencia@um.es (R. Valencia-García)
0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-3274-8825 (S. M. Jiménez-Zafra); 0000-0001-6244-6532 (M. Á. Rodríguez-García); 0000-0003-2457-1791 (R. Valencia-García)

This paper describes the participation of the UMUTeam in both subtasks of the DA-VINCIS shared task. The strategy followed by our team for the identification and classification of aggressive and violent events is based on the combination of Transformers with linguistic features, extracted from our UMUTextStats tool [4, 5] and from our negation processing system [6, 7]. In particular, we evaluate knowledge integration and ensemble learning as combination strategies.

The rest of the paper is organized as follows. First, Section 2 presents the task and the data provided in the competition. Second, the methodology carried out for the development of our system is described in Section 3. Then, the results obtained are shown in Section 4. Finally, the conclusions obtained after the participation in the task are summarized in Section 5.

2. Task description

The aim of the DA-VINCIS task is to identify the presence of violent incidents in Twitter and classify them in a given set of crime categories. For this, two subtasks are proposed:

- *Subtask 1: Violent event identification*, on determining whether a given tweet is related to a violent incident or not.
- *Subtask 2: Violent event category recognition*, on classifying the type of crime in one of the given categories: i) accident (involuntary damage), ii) homicide (deprivation of life), iii) kidnapping (deprivation of liberty), iv) non-violent-incident (no crime), or v) robbery (taking property unlawfully).

For this purpose, a dataset consisting of 5,000 Spanish tweets related to violent incidents is provided. The statistics, divided for each subtask, are depicted in Table 1. As it can be observed, the dataset is almost balanced for the binary classification problem (violent, non-violent), but it is imbalanced for the multi-class setting (accident, homicide, kidnapping, non-violent-incident, robbery). It should be mentioned that the organizers provided two sets, training and test, so we selected from training a custom split for validation. It was created by stratified sampling in order to maintain a balance between labels.

Regarding the evaluation measures, the F1 measure of the violent class was used to evaluate the participating systems in subtask 1, while in subtask 2 the Macro-F1 was the measure selected.

3. Methodology

Our methodology can be summarised as follows. We start by dividing the DA-VINCIS dataset into training and validation. Next, we obtain a cleaned version of the dataset. For this, we remove punctuation marks, hyperlinks, and emojis. We fix misspellings and expand acronyms and abbreviations. Next, we extract four feature sets. These features include linguistic features (LF) related to phonetics, morphosyntax, correction and style, semantics, pragmatics, figurative language, stylometry, lexis, psycho linguistic processes, social media jargon, and fine-grained negation features. They are extracted from our UMUTextStats tool [4, 5] and our negation processing system [6, 7]. The other feature sets are based on sentence embeddings. A non-contextual sentence embeddings from the Spanish model of FastText [8] (SE), and two contextual

Table 1
Dataset distribution for subtask 1 and 2.

	train	val	test	total
Subtask 1				
non-violent	1460	365	-	1825
violent	1270	317	-	1587
Subtask 2				
accident	906	231	-	1137
homicide	221	44	-	265
kidnapping	36	11	-	47
non-violent-incident	1460	365	-	1825
robbery	143	41	-	184

embeddings from BETO [9], the Spanish version of BERT (BF), and MarIa, the Spanish version of RoBERTa [10] (RF). Finally, we train several neural networks using hyperparameter selection. For this experiment, we select the best neural network (according to the validation dataset) for each feature set.

The neural networks are combined using two strategies. One the one hand, Knowledge Integration (KI), which consists of training from scratch a neural network with an input per feature set (see Figure 1 for a diagram with the architecture of this strategy). In this neural network, every feature set is connected separately to several hidden layers. The final hidden layers of every feature set are concatenated in a new hidden layer and connected to the final output layer. On the other hand, Ensemble Learning (EL), which consists of combining the performance of the best neural network from each feature set. There are several strategies for combining the results of several neural networks. We evaluate: i) hard voting, ii) soft voting, iii) averaging probabilities, and iv) highest probability. The first strategy, hard voting, is the mode of the predictions of each classifier. The second strategy, soft voting, is the weighted mode of the predictions of each classifier. The weights are calculated on basis on the F1-score achieved with the custom validation split. The third strategy, averaging probabilities, consists of averaging the probabilities predicted for each classifier. The last strategy, highest probability, consists of selecting the final label on basis of the model that predicts the final output with higher value.

Table 2 shows the best hyperparameters for subtasks 1 and 2. The experiments involve all the feature sets used separately (LF, SE, BF, and RF) and combined using KI, as this strategy consists of training a neural network from scratch. In case of EL, the results are combined using the best neural network for each feature set. For subtask 1 it can be observed that LF and SE are shallow neural networks composed by few hidden layers and neurons. However, sentence embeddings from BERT and RoBERTa work better with complex neural networks. This result draw our attention as our previous experience suggested that contextual sentence embeddings after fine tuning require few neurons to achieve their best results. All experiments achieve their best results with a dropout mechanism. The ratio, however, changes from one experiment to another. Concerning the learning rate, it keeps constant for subtask 2, with a ratio of 0.010. For

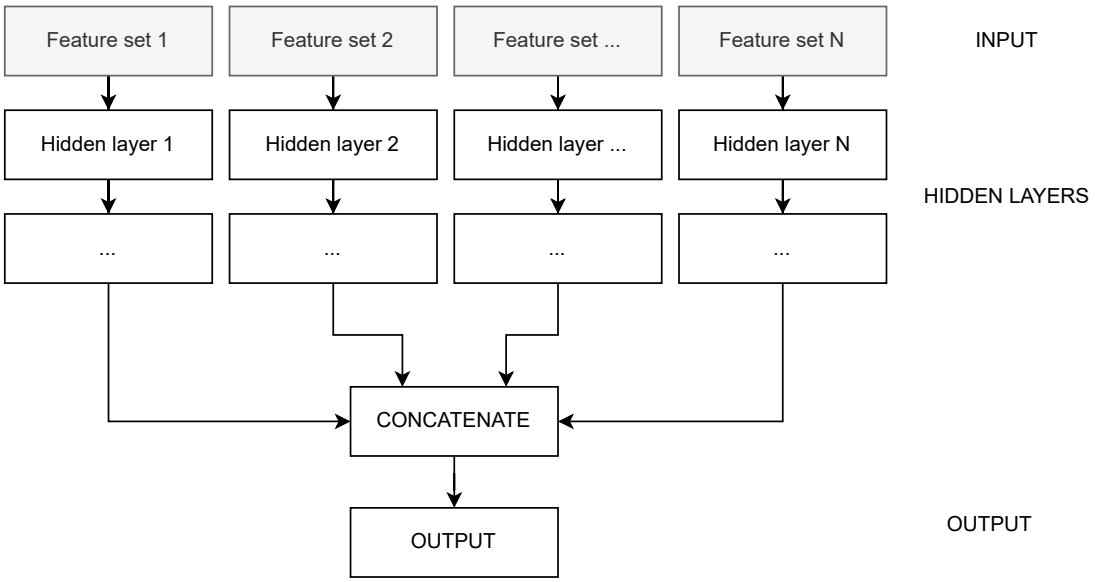

Figure 1: Architecture of the Knowledge Integration strategy.

subtask 1, the ratio is small for LF, SE and KI and bigger for contextual sentence embeddings (RF and BF). Finally, concerning the activation function, it can be observed that for subtask 1, tanh achieves better results for all the feature sets trained separately and selu when combined using KI. The results for subtask 2 are different, being linear the best activation function for LF and SE, sigmoid for RF, tanh for BF, and elu for KI.

Table 2
Best hyper-parameters for subtask 1 and subtask 2 for each feature set trained separately and combined using KI.

Feature set	shape	hidden layers	neurons	dropout	lr	activation
			Subtask 1			
LF	brick	2	2	.2	0.001	tanh
SE	brick	1	8	.2	0.001	tanh
BF	triangle	5	512	.2	0.010	tanh
RF	brick	4	256	.2	0.010	tanh
KI	diamond	3	2	.3	0.001	selu
			Subtask 2			
LF	brick	1	256	.3	0.010	linear
SE	brick	2	256	.3	0.010	linear
BF	brick	2	16	.1	0.010	tanh
RF	brick	2	128	.1	0.010	sigmoid
KI	triangle	5	512	.1	0.010	elu

4. Results

For both subtasks we sent five runs in total. The first run consists of the integration of all the feature sets using KI. The second, third, and fourth runs are based on EL using soft voting, the average of the predictions and a ensemble based on the highest probability. The fifth run consists of a KI strategy excluding the negation features.

Table 3 depicts the results achieved with each run in the first subtask. As it can be observed, our best result is achieved with Soft voting (run2), followed by KI (run5). Considering the impact of the negation features in the KI strategy (run 1 vs run 5), these features increase the precision of the system but decrease the recall. It is worth noting that the most limited result is achieved with the highest probability strategy (run4), both in terms of precision and recall. This result draws our attention, as this strategy usually reported the better precision in past experiments.

Table 3
Macro precision, recall, and f1-score for the first subtask with our custom validation split.

run	description	precision	recall	f1-score
1	KI	80.986	81.022	81.002
2	EL - Soft voting	**81.138**	**81.138**	**81.138**
3	EL - Averaging	80.134	80.034	80.076
4	EL - highest prob.	79.487	78.863	78.107
5	KI - without negation	80.275	80.399	80.303

As it can be observed from Table 4, the best result with the custom validation split is achieved with the soft-voting strategy and ensemble learning. In this subtask, the contribution of the negation features (run1) limited the precision of the results (41.527% vs 44.630%) but increases the recall (42.696% vs 38.004%). The run4, consisted in ensemble learning based on highest probability strategy, achieved the most limited precision (35.767%) but the best recall (53.770%).

Table 4
Macro precision, recall, and f1-score for the second subtask with our custom validation split.

run	description	precision	recall	f1-score
1	KI	41.527	42.696	41.540
2	EL - Soft voting	**57.248**	43.800	**47.174**
3	EL - Averaging	42.836	40.347	41.281
4	EL - highest prob.	35.767	**53.770**	41.701
5	KI - without negation	44.631	38.004	40.597

We report the results of the soft voting strategy as the best run for each trait in Table 5. This model could not identify any of the instances of the `kidnapping` class. Besides, other traits that achieve limited results are `homicide`, and `robbery`. As these traits are minority in the dataset, the micro average F1-score is competitive: 73.282%.

Table 6 contains the results for the first subtask. We achieved the forth position, with an F1-score of 76.4%. This result is achieved with an ensemble learning with soft voting (run2).

Table 5
Macro precision, recall, and f1-score for each trait of the second subtask with our custom validation split using the soft-voting strategy.

	precision	recall	f1-score
accident	79.111	77.056	78.070
homicide	75.000	20.455	32.143
kidnapping	0.000	0.000	0.000
non-violent-incident	78.796	82.466	80.589
robbery	53.333	39.024	45.070
micro avg	77.658	72.832	75.168
macro avg	57.248	43.800	47.174
weighted avg	75.899	72.832	73.282
samples avg	73.387	73.827	73.509

However, the results achieved by all participants are similar. The average F1-score is 75.418% with a standard deviation of 0.014.

Our next best result is achieved with the ensemble learning based on averaging probabilities (run3) with an F1-score of 76.152%, followed by the knowledge integration strategy: 75.963% of F1-score with all features excluding negation features (run5), and 75.748% with all features (run1). Our most limited result is achieved with the highest probability strategy, with an F1-score of 73.971%. However, this strategy achieved the best precision in the overall task: 89.12%.

Table 6
Official leader board for subtask 1.

Ranking	Team	F1-Score	Recall	Precision
1	danielvallejo237	77.589	75.037	80.320
2	Vicomtech	77.321	73.730	81.280
3	ITAINNOVA	76.512	75.154	77.920
4	UMUTeam	**76.401**	**75.389**	**77.440**
5	sdamian	75.616	75.079	76.160
6	Bernardo	75.483	73.054	78.080
7	Abu	74.802	74.097	75.520
8	atnafu	74.550	73.006	76.160
9	tahoangthang	74.437	74.798	74.080
10	JuanCalderon	74.281	76.351	72.320
11	sustaitangel	72.608	74.248	71.040

Table 7 contains the results for the second subtask. As it can be observed, we achieved the sixth position with an F1-score of 44.844%. This result is obtained with our second run, that consisted of a ensemble learning with soft voting. This run got better precision (53.684%) than recall (40.7863%). Our result is lower than the average of the results, that is an F1-score of 47.601% with a standard deviation of 0.05.

With the rest of the runs, we achieved an F1-score of 44.8444% with the ensemble learning

based on highest probability (run2). In fact, this run achieved a very promising precision of 53.684%. Our next best result is obtained with the ensemble learning based on averaging probabilities (run3), with an F1-score of 41.006%. The knowledge integration strategies achieved our most limited results: an F1-score of 39.025% (run5) without negation features and 38.520% (run1) considering all features.

Table 7
Official leader board for subtask 2.

Ranking	Team	F1-Score	Recall	Precision
1	Kelven	55.428	56.423	55.003
2	Vicomtech	52.856	54.592	51.749
3	ITAINNOVA	50.460	50.374	50.926
4	atnafu	49.030	52.099	46.747
5	danielvallejo237	47.331	42.195	65.508
6	UMUTeam	**44.844**	**40.786**	**53.684**
7	sustaitangel	43.362	42.418	45.948
8	Abu	37.501	37.761	37.691

5. Conclusions

In these working notes, we have detailed the participation of the UMUTeam in the DA-VINCIS shared task, which objective is the identification and categorization of violent events on social networks. Both challenges have been addressed exploring two strategies for combining linguistic features and embeddings from Transformers. Our team ranked fourth in subtask 1 (F1-score of 76.4%) and sixth in subtask 2 (F1-score of 44.8%).

From here, there are different ways to improve our work. First, our results are biased to our custom validation split. Therefore, we will evaluate better strategies for model selection, such as nested cross validation. Second, the performance achieved in the second subtask is limited as there are some traits in which any instance has been correctly classified. We suspect that this limitation is caused by the fewer instances of these traits. Therefore, we will evaluate data-augmentation techniques in order to overcome this drawback. Third, in future works we will focus on the interpretability of the results, in order to observe the relation between the linguistic features and the embeddings from Transformers.

Acknowledgments

This work was supported by Project LaTe4PSP (PID2019-107652RB-I00) funded by MCIN/AEI/10.13039/501100011033, Project AlInFunds (PDC2021-121112-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project LIVING-LANG (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe, and Big Hug project (P20_00956, PAIDI 2020) and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Govern-

mentR, and by "Programa para la Recualificación del Sistema Universitario Español 2021-2023". In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the industrial doctorate programme, and Salud María Jiménez-Zafra has been partially supported by a grant from Fondo Social Europeo and Administración de la Junta de Andalucía (DOC_01073).

References

[1] S. Stieglitz, M. Mirbabaie, B. Ross, C. Neuberger, Social media analytics – Challenges in topic discovery, data collection, and data preparation, International Journal of Information Management 39 (2018) 156–168. doi:https://doi.org/10.1016/j.ijinfomgt.2017.12.002.

[2] E. Kotzé, B. A. Senekal, W. Daelemans, Automatic classification of social media reports on violent incidents in South Africa using machine learning, South African Journal of Science 116 (2020) 1–8.

[3] L. J. Arellano, H. Jair Escalante, L. Villaseñor-Pineda, M. Montes y Gómez, F. Sánchez-Vega, Overview of DA-VINCIS at IberLEF 2022: Detection of Aggressive and Violent Incidents from Social Media in Spanish, Procesamiento del Lenguaje Natural 69 (2022).

[4] 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.

[5] 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.

[6] S. M. Jiménez-Zafra, R. Morante, E. Blanco, M. T. Martín-Valdivia, L. A. Ureña-López, Detecting negation cues and scopes in Spanish, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 6902–6911.

[7] S. M. Jiménez-Zafra, M. Taulé, M. T. Martín-Valdivia, L. A. Ureña-López, M. A. Martí, SFU Review SP-NEG: a Spanish corpus annotated with negation for sentiment analysis. a typology of negation patterns, Language Resources and Evaluation 52 (2018) 533–569.

[8] É. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning Word Vectors for 157 Languages, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018, pp. 3483–3487.

[9] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained BERT model and evaluation data, PML4DC at ICLR 2020 (2020) 1–10.

[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.

A. Online Resources

The source code is available via GitHub.