UMUTeam at SemEval-2022 Task 6: Evaluating Transformers for detecting Sarcasm in English and Arabic

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

Abstract

In this manuscript we detail the participation of the UMUTeam in the iSarcasm shared task (SemEval-2022). This shared task is related to the identification of sarcasm in English and Arabic documents. Our team achieve in the first challenge, a binary classification task, a F1 score of the sarcastic class of 17.97 for English and 31.75 for Arabic. For the second challenge, a multi-label classification, our results are not recorded due to an unknown problem. Therefore, we report the results of each sarcastic mechanism with the validation split. For our proposal, several neural networks that combine language-independent linguistic features with pre-trained embeddings are trained. The embeddings are based on different schemes, such as word and sentence embeddings, and contextual and non-contextual embeddings. Besides, we evaluate different techniques for the integration of the feature sets, such as ensemble learning and knowledge integration. In general, our best results are achieved using the knowledge integration strategy.

1 Introduction

Sarcasm is a form of rhetorical device based on biting humour to *disarm* an opponent during a dialog (Wilson, 2006). On the Web and, specially in social networks and opinion forums, sarcasm is very popular because it is funny to read and helps to stimulate the viral phenomenon of social media content (Peng et al., 2019). As sarcasm usually relies on figurative language and wordplay, in which words diverts from their conventional meaning, sarcastic statements hinder the ability of automatic classification tasks to perform sentiment analysis, hate-speech detection, or author analysis among other tasks.

From a Natural Language Processing (NLP) perspective, sarcasm, among other forms of figurative speech, such as irony or satire, has been explored in (del Pilar Salas-Zárate et al., 2020). In this work, the authors explored what stands out the most discriminant features for satire and irony detection. The authors identified a total of 25 feature sets. Apart from lexicon-based features and word-n-grams features, such as unigrams or bigrams, the authors identified style features, sentiment and emotional features, pragmatic features, and punctuation features, to name a few. Our proposal for solving sarcasm detection includes language-independent feature sets extracted with a custom tool, UMUTextStats (García-Díaz et al., 2021; García-Díaz and Valencia-García, 2022; García-Díaz et al., 2022a,b).

In this work, we describe the participation of the UMUTeam at SemEval 2022 task 6, concerning sarcasm identification in Arabic and English (Abu Farha et al., 2022). In this edition, three challenges are proposed. The first one is a binary classification problem to determine whether a document is sarcastic or not. This challenge is for English and Arabic. The second challenge is available only in English, and consists in a multi-label classification task discerning among different types of ironic speech. Finally, the last challenge consists of the identification of a sarcastic document between itself and a non-sarcastic rephrase, but with the same meaning. Our team attemted to participate in the first and second challenge. We achieve a F1 score of the sarcastic class of 17.97% for English, and 31.75% for Arabic in the first challenge. However, our results are not considered in the official leader board for the second challenge due to an unknown error. Therefore, we report the results for the second challenge with the validation split of dataset.

2 Dataset

The dataset proposed at iSarcasm 2022 was annotated by the authors themselves. Besides, each author was asked to rephrase their sarcastic texts without the usage of sarcasm. Finally, some linguistic

*Corresponding author

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1012 - 1017

Split	English	Arabic
training	2774	2800
val	694	700
test	1400	1400
total	4868	4900

Table 1: Corpus statistics by split and language for the first challenge in English and Arabic

Trait	Training	Validation
irony	130	25
overstatement	35	5
rhetorical question	81	20
sarcasm	565	148
satire	19	6
understatement	9	1

Table 2: Corpus statistics per sarcastic mechanism (subtask 2, English)

experts were asked to perform the multi-label annotations based on the following ironic speech labels: sarcasm, irony, satire, understatement, overstatement, and rhetorical question (Leggitt and Gibbs, 2000).

The statistics of the dataset concerning the first challenge are shown in Table 1. There is a significant imbalance between the labels. The relationship between the sarcasm and non-sarcasm texts is a 1:3 for English and 1:6 for Arabic. The statistics concerning the second challenge are shown in Table 2. There are 713 documents annotated as sarcasm, 155 as irony, 101 as rhetorical questions, 40 as overstatement, 25 as satire, and 10 as understatement.

From the training set, we select a 20% of instances to build the validation set using stratified sampling, in order to keep the balance.

3 System architecture

Figure 1 depicts the architecture of our proposal. Basically, we build two systems: one for English and one for Arabic, so we could apply different preprocessing techniques and apply language-specific pretrained embeddings. In a nutshell, this pipeline can be described as follows. First, is the preprocessing step module. For both languages, we ensure that the dataset does not contains hyperlinks, hashtags, quotations or emojis. Plus, for the English dataset we expand acronyms. Second, is the data-splitter, to divide the iSarcasm dataset into training and validation. As there was a strong imbalance in the dataset, we keep this imbalance in both splits. Third, is the feature extraction module, for extracting the language-independent linguistic features and the sentence embeddings. Forth, is the training of several neural networks using hyperparameter selection. Finally, is the feature integration module, in which we evaluate ensemble learning and knowledge integration in order to combine the results of each neural network.

Next, the feature extraction module is described. The first feature set is a subset of language-independent linguistic features (LF) from UMU-TextStats. This feature sets includes Part-of-Speech (PoS) features and stylometric features concerning several linguistic metrics such as Type Token Ratio (TTR), punctuation symbols and corpus length. The second and third feature sets are, respectively, non-contextual word and sentence embeddings from FastText (Mikolov et al., 2018). For this, we use the Arabic and English pretrained models. The word embeddings allow to evaluate convolutional and recurrent neural network architectures, apart from multi-layer perceptrons that suitable for feature sets of fixed size. The forth feature set is contextual sentence embeddings. We use BERT (Devlin et al., 2018) for English, and Arabic BERT (Safaya et al., 2020) for Arabic. To extract these embeddings, we applied a similar method as described at (Reimers and Gurevych, 2019). Before extracting the sentence embeddings, we used RayTune (Bergstra et al., 2013) to fine tune BERT and Arabic BERT. For this, we used Tree of Parzen Estimators (TPE) to select the best hyperparameters from a total of 10 trials. The hyperparameters evaluated are: (1) the weight decay (between 0 and 0.3); (2) two training batch sizes: 8 and 16 (we were limited to the GPU); (3) four warm-up steps: 0, 250, 500, 1000; (4) the number of training epochs, between 1 and 5; and (5) a learning rate between 1e–5 and 5e–5.

Once the feature sets are extracted, the next step in the pipeline is the training of the neural networks. We train a neural network per feature set and a neural network combining all the feature sets, using a knowledge integration strategy. Each training is performed with an hyperparameter optimisation stage. This evaluation includes 20 shallow neural networks, in which one or two hidden layers are stacked and that contains the same number of neurons. For the shallow neural networks we evaluate the following activation functions: (`linear`,

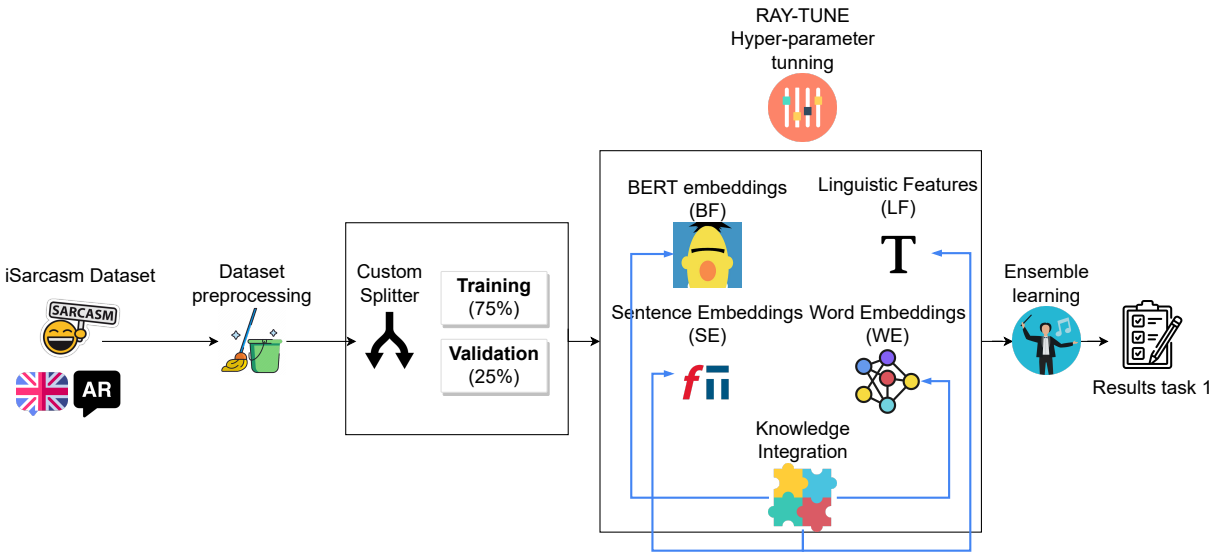

Figure 1: System architecture for solving the iSarcasm 2022 shared tasks

`ReLU`, `sigmoid`, and `tanh`). We also evaluate 5 deep-learning networks, composed from 3 to 8 hidden layers, in which the number of neurons per layer are arranged different shapes (brick, triangle, diamond, rhombus, and funnel). The activation functions evaluated for the deep-learning networks are the `sigmoid`, `tanh`, `SELU` and `ELU`. We also adjust the learning rate to evaluate `10e-03` or `10e-04`. As commented above, the word embeddings from fastText allow us to evaluate 10 convolutional and 10 bidirectional recurrent neural networks. As the dataset was heavily imbalanced, we evaluate large batch sizes, so in all the experiments, we evaluate large batch sizes: `128`, `256`, and `512`. We apply these large batch sizes for ensuring all batches contains sufficient number of instances of both classes. In addition, we apply regularisation by using a dropout mechanism (`[False, .1, .2, .3]`).

Besides, we evaluate two ensemble learning strategies to combine all the features. One method consisting of hard voting the labels predicted by each model (mode) and another method consisting of averaging the probabilities of the label of each model (mean).

It is worth mentioning that one of the tasks of English language is a multi-label problem. To solve it, we trained several binary classification models, one per ironic speech label.

4 Results and validation

During the development stage, we evaluate our models with a custom validation split. The results for the first task are depicted in Table 3, both for English and Arabic. We can observe that there is a strong difference between the results of the English and the Arabic datasets.

First, in case of English, we can observe that the results for identifying sarcasm (F1-pos) are limited, reaching a F1-score of 45.82% with the knowledge integration strategy. We focus on this metric because the validation split is unbalanced, containing 173 sarcastic documents and 521 non-sarcastic documents. The results of the LF are limited in case of English. LF are based mostly on stylometric and PoS features, which are not enough for the correct identification of sarcasm in English. Moreover, the results concerning the sarcasm label achieved with the non-contextual word embeddings (WE) are similar to the ones achieved with the LF (40.10% vs 40.08%), both slightly worse than non-contextual sentence embeddings (SE, 42.27%). The results achieved with transformers (BF) are the best results achieved with the feature sets evaluated in isolation (F1-pos of 45.10%). When combining the results, we observe that we achieve slightly superior results by the knowledgte integration approach, improving to a F1-pos of 45.82%. However, the results are more limited with the the ensemble strategies, achieving the lower results of 38.06% (mode) and 37.96% (mean). The identification of the non-sarcasm label (F1-neg) is more similar (and even slightly superior) compared with the rest of the strategies.

Second, in case of Arabic, we observe astonishing results for all feature sets. We reviewed the

	English			Arabic		
Strategy	**F1-neg**	**F1-pos**	**F1-score**	**F1-neg**	**F1-pos**	**F1-score**
LF	67.92	40.08	54.00	98.14	90.00	94.07
SE	73.21	42.27	57.74	97.99	88.35	93.17
WE	76.68	40.10	58.39	98.07	89.00	93.53
BF	82.40	45.10	63.75	99.08	94.69	96.88
LF-SE-WE-BF (K.I.)	83.57	**45.82**	**64.69**	99.08	94.69	96.88
LF-SE-WE-BF (mode)	**83.71**	38.06	60.89	**99.17**	**95.05**	**97.11**
LF-SE-WE-BF (mean)	78.84	37.96	58.40	96.71	81.86	89.28

Table 3: Results for the custom validation split in the first challenge for English and Arabic. We show F1-neg for non-sarcasm, F1-pos for sarcasm, and the macro F1-score (F1-score) for the neural networks evaluated with a feature set, and the combinations of features by using knowledge integration and two ensemble learning strategies.

dataset in order to find duplicates but we could not identify a relevant number of them. The identification of a large number of duplicates could indicate that some of the instances of the training split were present in our custom dataset. In this case and due to our lack of understanding of Arabic, we could not identify in the validation split the high performance achieved. After the competition, the ground truth labels of the test set were released for the development of the working notes. We observed that the results with the test split are more limited. For example, the F1-score falls from 94.07% with LF with the custom validation split to 33.82% with the official test split. The drop in the results cannot be explained by class imbalance, as we performed a stratified split in order to build the validation split. In fact, our validation split has 102 sarcasm documents whereas the rest (598) are non-sarcasm. The official test contains 1400 documents, 200 labelled as sarcasm whereas the rest were labelled as non-sarcasm. The results achieved with the rest of the feature sets SE, WE, and BF are in the same line, achieving the best result with a macro F1-score of 44.85% with the knowledge integration strategy.

Next, we report the results achieved for each trait separately in Table 4. This table reports the overall macro F1-score. The limited results achieved are caused by the strong imbalance among the validation split. From a total of 694 samples of the validation split of the English dataset, there are 25 samples of irony, 5 of overstatement, 20 rhetorical questions, 148 based on pure sarcasm, 6 based on satire, and 1 for understatement. These results indicate that one of the main drawbacks of our proposal is related to handle class imbalance, as the majority of the neural networks developed does not behave better than a random classifier. Moreover, some surprising results are achieved for the rhetorical questions and understatements, with the ensemble strategy of averaging the probabilities (mean).

Concerning the official results, we achieved very limited results for the binary classification challenge in English, achieving position 41 with a F1-score of the sarcastic class of 17.97% (see Table 5). The best result was achieved by user *stce* with a F1-score over the sarcastic label of 60.52%, which indicates that our system is far from the best results. The average F1-score of the sarcastic label of the rest of the participants is 32% with a standard deviation of 11.24%.

The results achieved with the Arabic dataset for the binary classification were better, with a F1-score of the sarcastic label of 31.75%, reaching position 22 of a total of 32 participants. (see Table 6). In this case, the best result was achieved by user *Abdelkader* with an F1-score over the sarcastic label of 56.32%. The average F1-score of the sarcastic label of the rest of the participants is 35% with a standard deviation of 10.04%.

5 Conclusions and further research lines

In this working notes we have described the participation of the UMUTeam in the iSarcasm shared task of SemEval 2022. In this shared task, the participants were required to solve a binary and a multi-label classification task regarding sarcasm identification in English and Arabic. We achieved a F1 score of the sarcastic class of 17.97% for English, and 31.75% for Arabic in the first challenge.

After sending the official results, we received the annotated test set. Although we are happy with our participation in this shared task, as we have evaluated some of our methods, such as a subset of language-independent feature sets in Arabic, we

Strategy	F1-trait-1	F1-trait-2	F1-trait-3	F1-trait-4	F1-trait-5	F1-trait-6
LF	52.39	66.52	68.18	56.30	55.46	49.96
SE	52.85	64.10	60.34	58.23	64.10	49.96
WE	**57.50**	52.79	63.96	56.41	60.86	49.96
BF	56.96	**69.78**	67.25	**62.84**	56.71	49.96
LF-SE-WE-BF (K.I.)	56.47	60.86	**71.10**	62.60	**62.28**	49.96
LF-SE-WE-BF (mode)	57.50	66.52	64.12	61.96	49.78	**49.96**
LF-SE-WE-BF (mean)	51.35	50.12	6.50	54.01	42.15	0.14

Table 4: Macro F1-score of the custom validation split in the second challenge. The traits are, number from 1 to 6, irony, overstatement, rhetorical question, sarcasm, satire, and understatement

Rank	User/Team	F1-sarcastic
1	stce	60.52
2	emma	52.95
3	saroyehun	52.95
41	**UMUTeam**	17.97
42	Matan	16.84
43	abhayshukla9	15.53

Table 5: Results for the first challenge (English)

Rank	User/Team	F1-sarcastic
1	Abdelkader	56.32
2	Aya	50.76
3	rematchka	47.67
22	**UMUTeam**	31.75
23	Pat275	30.13
43	Matan	29.51

Table 6: Results for the first challenge (Arabic)

are aware that our results are limited. Our preliminary experiments with the official annotated test suggests that our major weakness is the class imbalance. As we already include some techniques to address this problem, as weighting the classes and evaluating larger batch sizes, we will explore methods for performing data-augmentation and try to increase the performance of our models.

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

Ibrahim Abu Farha, Silviu Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

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.

María del Pilar Salas-Zárate, Giner Alor-Hernández, José Luis Sánchez-Cervantes, Mario Andrés Paredes-Valverde, Jorge Luis García-Alcaraz, and Rafael Valencia-García. 2020. Review of english literature on figurative language applied to social networks. *Knowledge and Information Systems*, 62(6):2105–2137.

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. 2021. 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. 2022a. Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. *Future Generation Computer Systems*, 130:59–74.

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

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.

John S Leggitt and Raymond W Gibbs. 2000. Emotional reactions to verbal irony. *Discourse processes*, 29(1):1–24.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Wei Peng, Achini Adikari, Damminda Alahakoon, and John Gero. 2019. Discovering the influence of sarcasm in social media responses. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1331.

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.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Deirdre Wilson. 2006. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10):1722–1743.