Graph-Based Profile Condensation for Users Profiling

Notebook for PAN at CLEF 2022

Roberto Labadie-Tamayo[1], Daniel Castro-Castro[2]

[1]Computer Science Department, University of Oriente, Santiago de Cuba, Cuba
[2]Information Retrieval Lab, Computer Science Department, University of A Coruña, Spain

Abstract

For social media profiling tasks, it is crucial the way information regarding user accounts is modeled and aggregated, especially when exploring micro-blogging platforms such as Twitter. In this paper, we describe our system for participating in the "Profiling Irony and Stereotype Spreaders on Twitter" task shared on PAN 2022, which analyses the user account with a graph-based modelation. The approach relies on a β-similar graph construction, whose nodes represent tweets within the account, encoded as a combination of irony and stereotyping-related features. The graph's information is propagated through a Spacial Attention-based Graph Neural Network and prototype nodes are selected and aggregated to classify the profile with a Feed-forward Neural Net.

We also describe a second model supported by a profile reduction whose ideas would be introduced to the graph construction under the assumption that given an ironic profile, non-ironic tweets may introduce some noise at the classification stage and vice versa.

Experimental results show this graph modelation solves issues detected in previous related proposals, also the profile reduction yields a profile kernel containing valuable and non-noisy information for classifying the user accounts.

Keywords

Spacial Graph Neural Network, Attention Mechanisms, Semantic Similarity, Profile Kernels, Prototypes

1. Introduction

Social media has provided users with an unprecedented and multi-modal means to interact with each other and express themselves. The language employed in these platforms typically exhibits a natural flow of speech and grammatical structure involving creative and figurative devices such as irony and sarcasm, which add complexity to the automatic analysis of the information posted by users.

Specifically, when using irony, people utter something different from what is wanted to express (i.e., typically the opposite of the true appraisal of circumstances), but it is expected the listener/reader to recognize the overt dissimulation by seeing through the counterfactual nature of the utterance and eventually detect the intended meaning of what is said [1].

It is very common to find out on social media a bunch of content employing this device to spread hateful and toxic expressions into a particular process, entity, or even groups of persons with some demographic or psychological characteristics by stereotyping them. The latter shows

CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
rlabadiet@gmail.com (R. Labadie-Tamayo); daniel.castro3@udc.es (D. Castro-Castro)
0000-0001-9102-7601 (D. Castro-Castro)

why it's imperative to discriminate whether some content is ironic in order to filter what is delivered to a user or censure the one harmful and socially inappropriate. Nevertheless, if we take into account that in this kind of media the receptors of written messages are deprived of additional knowledge from gestures, prosody features, visual content, and sometimes situational environment, which are crucial for understanding such a figurative form of communication as irony is, we can certainly assume that complexity increases for automatic methods. These methods need to consider at least socio-cultural rules to properly understand the real meaning behind funny but hateful material.

Twitter as one of the most popular microblogging platforms on the internet, has not escaped from the raising of misinformation and hate speech spreading phenomena, constituting an important source of toxicity. Following this line Author Profiling (AP) tasks have been proposed to identify psychosocial characteristics of Twitter users [2, 3, 4] that could help to detoxify this scenario.

Last year at PAN 2021, was proposed the task: "Profiling Hate Speech Spreaders on Twitter" [4] where systems were developed to determine whether a Twitter account tends to spread hateful speech. A diversity of models reasonably able to profile authors were proposed by participants, employing techniques ranging from traditional Machine Learning approaches to more novel methods involving Convolutional, Recurrent, and Graph-based Neural Networks [5, 6, 7].

This year a more complex task in terms of understanding and associativity with contextual knowledge has been proposed at PAN 2022 [8]: "Profiling Irony and Stereotype Spreaders on Twitter" (IROSTEREO) [9], aiming at profiling ironic authors on Twitter, paying special attention to those which mask stereotyping content by employing irony.

Many works have been proposed for predicting automatically whether a textual message is ironic. These methods rely mainly in the use of Recurrent Neural Networks [10] or more recently, fine-tuned Transformer Models [11] and conventional Machine Learning models like Logistic Regression [12] or Support Vector Machines (SVM) [13]. Stereotypical language detection has also been well-studied [14, 15, 16] yielding proposals aligned with the use of architectures employed for irony detection, such as [17] with Transformer models.

Regarding the Author Profiling field, the way of aggregating the information of isolated tweets/texts from the author account has also been studied, especially in the last edition of PAN were employed successfully different methods such as sequential representations, single-dense vectors, and graph-based representations.

In this work, we describe our system for participating in the "Profiling Irony and Stereotype Spreaders on Twitter" task at PAN 2022 and explore with an incremental modeling complexity how to structure the information obtained from a Twitter user account for classifying it regarding the proposed task. Our general architecture combines Deep Learning techniques with traditional ones from Machine Learning, particularly, it consists of a Sentence Encoder Module, based on a transformer architecture for obtaining abstract tweet representations. These encodings are employed for modeling and classifying the user's profile. The source code of our approach is available on GitHub[1].

The paper is organized as follows: In Section 2 we briefly describe the task proposed and datasets employed. Section 3 presents the system's architecture and provides details about its modules.

[1] https://github.com/labadier/IROSTEREO

Section 4 describes the experiments and the achieved results. Finally, we present our conclusions and provide some directions that we plan to explore in future works.

2. Task and Datasets

This year the Profiling Task at PAN, in contrast with its last editions, is deployed in a monolingual perspective, analyzing just the English language. It aims at classifying a Twitter user account as an Irony and Stereotypes Spreader or not, given 200 posts from its feed. Also, as a subtask, participants are asked to determine whether the use of stereotypical language was intended to hurt or support the targeted social group. The dataset provided by the organizers [18], as usual, only contains annotations at the profile level and is balanced concerning the positive and negative classes distribution for the first subtask. For the second subtask, the hurting class is underrepresented.

We introduce additional data from the training set proposed in Task 3A at SemEval-2018: "Irony Detection in English Tweets" [19], annotated regarding the presence of irony and also uniformly distributed for positive and negative classes. This training set is composed of 3817 data points, with 1901 examples of ironic messages.

The training set from SemEval-2019 Task 5: Detection of Hate Speech Against Immigrants and Women in Twitter (HateEval)[20] was also taken into account. This dataset is annotated regarding the hateful speech towards specific social groups, the presence of an aggressive language, and its range, indicating whether the hate targets a person or a group. It contains 9000 tweets for the English language, with 3783 annotated as hateful. In this work, it is of our interest just the latter information from the dataset.

3. System Description

As in previous works [6, 7] our system relies on a modular architecture, conditioned by the use of data representations learned through fine-tuned state-of-the-art neural language models; while introducing a more sophisticated modelation and aggregation for the Twitter feed and whose improvement target is to bypass as much as we can the hyperparameters from Deep Impostor Method in [7]. We also propose a non-graph-based system to compare with this modelation.

From a general standpoint, given a user account, our system at first approach the tweet encoding to afterward aggregate them and classify the profile in IROSTEREO spreader or not. This procedure also applies to the stance detection subtask.

3.1. Data Representation

Three pre-trained Transformers Language Models (LMs) [21] were employed to encode in a dense way the semantic information contained in a tweet. All three employ the BERT-base [22] configuration but are pre-trained or fine-tuned into a different data domain.

The first model, a simple BERT-base, pre-trained on a dataset consisting of 11,038 unpublished

books from BookCorpus[2] and texts from English Wikipedia, is employed as a ***raw*** encoder, unbiased into any downstream NLP task but just the Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).
A second encoding for tweets was learned by Twitter-roBERTa-base for Irony Detection [23] (***irony*** encoding). This model was fine-tuned into the Irony Detection task, by employing the training dataset from Task 3A at SemEval-2018: "Irony Detection in English Tweets". With this representation, we aim to capture abstract features that the model relies on for classifying whether a tweet is ironic, which is a key matter for the studied AP task.
Finally, we also include the encodings learned by the pre-trained LM HateXplain [24] fine-tuned with the data from HatEval over the downstream task of classifying whether a tweet comes hatefully undertoned (***h-stereo*** encoding). Taking into account this dataset has the particularity that hate is spread toward immigrants and women, and in the case of sarcastic language generally, the speaker aims to hurt [25], we hypothesize that these encodings capture relations regarding stereotyping and toxicity of posts.
For each case, we made use of the HuggingFace Transformers library [3] and when the fine-tuning was carried out, we stacked to the Transformer Model (TM) an intermediate layer as a bottleneck for extracting a representation of the input message in a latent space. This intermediate layer is fed with the first token of the output sequence from the TM. Then, this encoding is passed to an output neuron which makes the prediction for the targeted task.
As in [7], in the tuning phase, we employed a gradual unfreezing-discriminative setting, proposed in the Universal Language Model Fine-tuning (ULMFiT) [26] by employing a different learning rate for each encoder layer in the TM, increasing it while the neural network gets deeper.

3.2. Graph-based Profile Modeling

In [7] we proposed modeling the profile in terms of tweets as a graph-based structure assuming the facts:

(i) In social media accounts, sentiments towards a group of people, topic, or simply some speech target are spread in a scattered manner e.g., the hateful content of an idea can be constructed by relating the information from different posts.

(ii) The profiles within the data were not structured following any temporal order, which is also aligned with our actual data.

This allowed us to share the information from one tweet (node) to the others while its individual information was being transformed to express how it belonged in its context, by employing a Spectral Graph Convolutional Neural Network (SGN). In this graph, every node was a tweet, and each of them was connected to the others.
Nevertheless, since a trivial complete-graph structuring was employed, no semantic relations were considered when creating the edges between nodes. On the other hand, as the aggregation

[2] https://yknzhu.wixsite.com/mbweb
[3] https://huggingface.co/transformers

function we used a simple normalized and un-weighted sum of the neighbors' information, defined by the node-wise operator:

$$x'_i = ReLU\left(\Theta \sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{1}{\sqrt{\hat{d}_j \hat{d}_i}} x_j\right) \tag{1}$$

Where x_i represents the encoding of the i^{th} node, d_i the degree of the i^{th} node, $\mathcal{N}(i)$ is the set of neighbor nodes of i and Θ is the matrix of learnable parameters (our filter) which learn the spectral relations.

Graph Structuring

For our approach, given a profile, we propose to link a pair of nodes according to semantic similarities captured by our *raw* encoder described in Section 3.1 in a β-similar graph. Afterward, communities are isolated by the Girvan-Newman clustering algorithm [27] for the detection and analysis of community structure.
Girvan-Newman algorithm relies on the iterative elimination of edges with the highest number of shortest paths between nodes passing through them, i.e., their betweenness [28]. With this connection reduction, we aim to group tweets related by topic and/or style within the same connected component of the graph.
Once nodes are grouped, their vector representation is replaced by the aggregation of irony-stereotyping features learned by *irony* and *h-stereo*. Over this representation, the non-necessarily connected graph is then fed into a non-Spectral Graph Convolutional Net.

In the prior process, we first model the graph with *raw* node encodings because we want to avoid grouping tweets according to biased representations into irony or stereotype information. Instead, we want to relate tweets according to their topic and semantic information regardless of any communicative device and style. All of this, considering how information is transformed by the Graph ConvNet and how prototypes are selected from topic-connected components.

Convolutional Operator

For spectral approaches like the one employed in [7], the learned filters Θ depend on the Laplacian eigenbasis, which depends on the graph structure as can be observed in the matrix-wise formulation of the message-passing function from [29]. Thus, a model trained on a specific structure can not be directly applied to a graph with a different structure, causing the high variance observed in [7] for this encoding-classification method.
In our approach, we employed the Spacial Attention-based operator from [30]. Also, this operator in the aggregation phase, rather than making an indiscriminate summation over neighbors' information, allows to compute the hidden representations of each node, by attending over its

neighbors following a self-attention strategy and giving them relative importance; that is:

$$x'_i = \sum_{j \in \mathcal{N}(i) \cup \{i\}} a_{ij} \Theta x_j \tag{2}$$

Where $\mathcal{N}(i)$ is the set of nodes belonging to the neighborhood of i, Θ computes higher-level features of the nodes encoding x_j and a_{ij} corresponds to the weighting factor for node j taking into account the query node i. This relative importance is computed as:

$$a_{ij} = softmax(\mathrm{a}^T g(\Theta[x_i || x_j])) = \frac{exp(\mathrm{a}^T g(\Theta[x_i || x_j]))}{exp\Big(\sum_{k \in \mathcal{N}(i) \cup \{i\}} \mathrm{a}^T g(\Theta[x_i || x_k])\Big)} \tag{3}$$

Here, a^T corresponds to a single dense-layer self-attention mechanism and g to the non-linear activation function, in this case, *LeakyRelu*.

Components Aggregation and Classification

When information within nodes has been shared and transformed according to their context, we aggregate all the connected components' information to approach the graph (profile) classification, for this, we explored two paradigms.

The first one was to construct for each connected component with a greedy fashion an approximation of a set of prototypes, where for any path of length 3 (e_1, e_2, e_3), there is one and just one node belonging to this set. This construction avoids considering two nodes with separation degrees lower than 3. Taking into account that when one message-passing step is carried out, the nodes get informed about all their neighbors, by selecting such a set of prototypes we cover all the information in the original components.
The graph representation is determined by the normalized summation over its prototypes and it is considered the learned profile modeling.

The second criterion relies on conducting the minimum message-passing iterations needed to ensure that any node has knowledge about the information of all the other nodes within its component. Then, under this all-nodes-informed assumption, a random prototype for each component is selected for modeling the profile as in the first approach.
Assuming the expected degree of a node within a connected component C is $\bar{\delta}$ and based on the social media degrees of separation [31], the maximum expected distance between two nodes is given by $t = \lceil \log_{\bar{\delta}} |C| \rceil$, and hence the minimum amount of needed iterations.

For both methods, the obtained profile representation is then fed into two sequentially-connected dense layers in charge to predict the probability of a profile being an irony and stereotypes spreader.

3.3. Two inner prototypes for a profile

As an alternative to the graph-based modeling, we propose to condense the profile information by selecting a kernel of posts with highly-shared knowledge. This reduction in the number of useful tweets to represent the profiles is carried out by considering that not necessarily every post from an irony and stereotypes spreader contains relevant information regarding being ironic or not. Hence, we aim to inhibit the probably noisy impact of non-representative tweets.

Given a profile, at first, each tweet is encoded by employing the *sn-xlm-roberta-base-snli-mnli-anli-xnli* pre-trained Language Model [4]. This model, as well as the above mentioned TMs, is based on a RoBERTa pre-training strategy but fine-tuned through a Siamese Neural Network over Natural Language Inference (NLI) task [32].
Afterward, it is defined for a message the knowledge-sharing score as the average for its cosine similarity with the rest of the profile's tweets. Taking into account these scores, we make two disjoint subsets of tweets for each profile, one containing the half (100 tweets) most knowledge-sharing (i), whereas the other contains its complement (ii).
The first subset is then composed of tweets associated with the main ideas or content addressed by the user, the second one can be seen as the set of tweets less representative of the profile.
Finally the embedding of the profile results from the sum of each tweet encoding from the subset (i).
For this modeling, we classify an unknown profile as an irony and stereotypes spreader or not using the following steps:

1. For each profile in the training set (IROSTEREO spreader or not), obtain its embedding as described above (k_e^i).
2. Obtain the embedding for the unknown profile (u_e).
3. Compare u_e with each of the embeddings k_e^i and assign the unknown profile to the class of the most similar embedding (1-NN classification).

For this strategy (in 3^{rd} step), we employed cosine similarity, and we also tried to construct the embedding with elements from (ii) rather than (i).

4. Experiments and Results

For evaluating the conducted experiments over the prior proposals, we employed a cross-validation approach with 5-folds, taking into account the original distribution for each introduced dataset. Also, we used the evaluation metrics proposed by the profiling task organizers, the accuracy and F1-Macro for first and second subtask respectively.
As naturally, we first fine-tuned the LMs seeking a good performance in the introduced intermediate tasks and hence latent representations able to capture ironic and/or stereotyping features to construct the final modelation of the whole profile.
For the stacked dense layer between the TM and the softmax classification layer refereed in Section 3.1, we employed the ReLU non-linear activation function and optimized the models'

[4] https://huggingface.co/symanto/sn-xlm-roberta-base-snli-mnli-anli-xnli/discussions

parameters with the RMSprop algorithm [33]. Over this configuration, we tuned the batch size, starting learning rate with values sampled in the range [1e-5, 5e-5] and the learning rate decay, resulting in the most stable results by employing an initial learning rate of 2e-5 and equal decay with a batch size of 64 examples over 8 epochs, as shown in Table 1.

Table 1
Data Representation Models Results for Intermediate Tasks.

Task	Model	Accuracy
Irony	irony-encoder	**0.979**
	bertweet	0.940
	bertweet-mtl	0.610
h-stereo	h-stereo-encoder	**0.965**
	bertweet	0.941
	bertweet-mtl	0.730

Here, bertweet refers to a model pre-trained in general-topic tweets [34] and fine-tuned into the corresponding task and data. It is introduced as a base model for a multitask approach (bertweet-mtl) to unify the ironic and stereotyping features learning.
As can be observed, in both feature-learners the higher accuracy was reached by models trained and fine-tuned within the downstream task domain, whereas the worst performance is obtained by the multi-task approach. Hence, the more effective combination of features for each tweet results from aggregating the ironic and stereotyping representations learned individually.

At the profile modeling stage, we tuned the β-value for constructing the β-similar graph. The most stable constructions were obtained for $\beta = 0.97$, with expected node degree $\bar{\delta} = 18$ and expected component size $\bar{c} = 53$, above this value almost every node remained isolated, which is equivalent to at aggregation and classification stage, making a summation over the tweets representation. Below 0.97, the components became considerably big, yielding a modelation similar to not making any grouping over style-content representations given by the LM.
Also, intuitively[5] we can assume that function $f(\beta) = \bar{\delta}$, as well as $g(\beta) = \bar{c}$, is monotonically decreasing because of the way β-similar graphs are constructed, therefore the results obtained within a sampled and relaxed neighborhood such as [0.9, 0.99] can be extended to the whole f-domain (g-domain).
Employing this construction and after tuning the batch size, learning rate as well as the number of hidden units for the condensation layer after the Graph Neural Net message passing, we explored avoiding the community refining phase by the Girvan-Newman algorithm on strategies described in Section 3.2.
Here, *GreedySharing* refers to the first aggregation method based on greedily selecting prototypes after message passing. *LogSharing*, refers to the second strategy, which is based on approaching as many sharings as needed to ensure every node to know about each other (based on the expected node degree). Also, the * stands for the no application of the Girvan-Newman algorithm. For all the configurations, the best results were achieved by employing a batch size

[5]Two cases must be considered when β increases: (i) the edges remain the same, (ii) they decrease at least by 1.

Table 2
Community Refining Results.

Modelation	Accuracy
GreedySharing	**0.845**
GreedySharing*	0.833
LogSharing	**0.714**
LogSharing*	0.711

of 16, a learning rate of $2e-3$, and 32 neurons for the intermediate dense layer, and as we can see, isolating communities and sharing their prototypes information, outperforms the accuracy in both methods, especially for the greedy-based approach.

Finally, to compare these architectures' results, jointly with the method described in Section 3.3 we bring the ideas previously proposed in [7]. The latter experimental models were fed with the latent representation learned by the intermediate layer of our Spacial Graph Network (SGN).

Table 3
Architectures Cross-validation Results.

Modelation	Accuracy
GreedySharing	0.845
kernel+	0.897
kernel-	0.864
Previous Works	
LSTM	0.821
SGN	0.761
GreedySharing+SVM	0.783
GreedySharing+DIM	0.775

In Table 3 *kernel+* and *kernel-* refers to methodologies described in Section 3.3, where tweets with higher or lower knowledge-sharing score are taken into account for modeling the profile respectively. *LSTM* modelation, instead considering the aggregation of *irony* and *h-stereo* features as nodes encodings, it analyzed the profile as a sequence. *GreedySharing+SVM* and *GreedySharing+DIM* employ the latent representation of the GreedySharing-SGN wit a SVM and the Deep Impostor Method [7] respectively.

From here, we can see how employing information commonly shared across the tweets within a profile outperforms our greedy prototypes aggregation through the Spacial Graph Neural Network. Nevertheless, this approach outperforms the deep-aggregation made by the Spectral Network and the LSTM sequential analysis.

For facing second subtask, we employed the best-performed models configurations on irony and stereotyping spreader detection. These are kernel+ and GreadySharing strategies, achieving

0.587 and 0.581 of F1-Macro resectivelly.
Regarding the official competition results, in the unblinded submission our Graph-based modelation achieved 81.1% of accuracy under the test dataset for the first subtask, showing an important variance-issue reduction w.r.t. the method proposed in last year's competition. Hence, increasing the model size would be a reasonable way to squeeze the biased scenario shown by cross-validation results taking into account the proposed kernel-based method. For the second subtask, we were allowed to make two submissions achieving 0.4886 and 0.4685 of F1-Macro for kernel+ and GreadySharing strategies respectively.

5. Conclusions and Future Works

In this paper, we described two systems for addressing the task shared on PAN 2022: "Profiling Irony and Stereotype Spreaders on Twitter". The first approach relies on a β-similar graph construction for each profile, whose nodes represent the tweets within the account, encoded as a combination of irony and stereotyping-related features. The graph's information is propagated through a Spacial Attention-based Graph Neural Network and prototype nodes are selected and aggregated to classify the profile with a Feed-forward Neural Net.
The second approach is supported by a profile reduction, under the assumption that given an ironic profile, non-ironic tweets may introduce some noise at the classification stage. Firstly, the tweets are encoded and compared for determining a knowledge-sharing score. Afterward, the highest scored tweets are selected as prototypes and aggregated, taking into account that containing highly employed contextual/stylistic abstract features indicates a particular behavior of users. Finally, an unknown profile is classified with the label of its more similar aggregation from the training set.
Experimental results show our graph-based model outperforms the combination of its latent representation with another meta-classifier, especially the Deep Impostor Method (DIM) proposed in [7]. The latter solves two issues; (i) sensible hyper-parameter selection for DIM, and (ii) the high variance observed with previously graph-based architecture. On the other hand, the approach based on profile reduction achieved even better results, which indicates that effectively some elements may result in noise when modeling and classifying the profile.
Since the new construction of the graph and its analysis paradigm (spacial instead of spectral) achieved competitive results with the kernel-based method, besides β-similar graphs still looking as a natural way to represent profile information, we plan to introduce the reduction schema described in Section 3.3. Also, for constructing the initial graph instead the *raw* encoder, employing a more similarity-oriented pre-trained LM as *sn-xlm-roberta-base-snli-mnli-anli-xnli* from Section 3.3, would yield a more robust graph structure. Finally, since the optimizing problem is now the bias reduction instead of the variance of the model, introducing a bigger architecture, possibly by using a more complex aggregation function, would help to this end.

Acknowledgments

This work was supported by projects PLEC2021-007662 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación,

Transformación y Resiliencia, Unión Europea-Next Generation EU) and RTI2018-093336-B-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación). The second author also thanks the financial support supplied by the Consellería de Cultura, Educación e Universidade (GPC ED431B 2022/33)

References

[1] N. Groeben, B. Scheele, Produktion von ironie und witz, 2003. URL: https://madoc.bib.uni-mannheim.de/7446/.

[2] F. Rangel, P. Rosso, Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/.

[3] F. Rangel, A. Giachanou, B. Ghanem, P. Rosso, Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/.

[4] F. Rangel, P. Rosso, G. L. D. L. P. Sarracén, E. Fersini, B. Chulvi, Profiling Hate Speech Spreaders on Twitter Task at PAN 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[5] R. Dias, I. Paraboni, Combined CNN+RNN Bot and Gender Profiling, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/.

[6] R. Labadie-Tamayo, D. Castro-Castro, R. Ortega-Bueno, Fusing Stylistic Features with Deep-learning Methods for Profiling Fake News Spreader—Notebook for PAN at CLEF 2020, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/.

[7] R. Labadie Tamayo, D. Castro Castro, R. Ortega Bueno, Deep Modeling of Latent Representations for Twitter Profiles on Hate Speech Spreaders Identification Task—Notebook for PAN at CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2936/paper-177.pdf.

[8] J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection, in: M. D. E.-F. S. C. M. G. P. A. H. M. P. G. F. N. F. Alberto Barron-Cedeno, Giovanni Da San Martino (Ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), volume 13390 of *Lecture Notes in Computer Science*, Springer, 2022.

[9] O.-B. Reynier, C. Berta, R. Francisco, R. Paolo, F. Elisabetta, Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO) at PAN 2022, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.

[10] C. Wu, F. Wu, S. Wu, J. Liu, Z. Yuan, Y. Huang, THU_NGN at SemEval-2018 task

3: Tweet irony detection with densely connected LSTM and multi-task learning, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 51–56. URL: https://aclanthology.org/S18-1006. doi:10.18653/v1/S18-1006.

[11] R. A. Potamias, G. Siolas, A. Stafylopatis, A transformer-based approach to irony and sarcasm detection, CoRR abs/1911.10401 (2019). URL: http://arxiv.org/abs/1911.10401. arXiv:1911.10401.

[12] O. Rohanian, S. Taslimipoor, R. Evans, R. Mitkov, WLV at SemEval-2018 task 3: Dissecting tweets in search of irony, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 553–559. URL: https://aclanthology.org/S18-1090. doi:10.18653/v1/S18-1090.

[13] H. Rangwani, D. Kulshreshtha, A. Kumar Singh, NLPRL-IITBHU at SemEval-2018 task 3: Combining linguistic features and emoji pre-trained CNN for irony detection in tweets, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 638–642. URL: https://aclanthology.org/S18-1104. doi:10.18653/v1/S18-1104.

[14] J. Cryan, S. Tang, X. Zhang, M. Metzger, H. Zheng, B. Y. Zhao, Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods, Association for Computing Machinery, New York, NY, USA, 2020, p. 1–11. URL: https://doi.org/10.1145/3313831.3376488.

[15] S. Matthews, J. Hudzina, D. Sepehr, Gender and racial stereotype detection in legal opinion word embeddings, 2022. URL: https://arxiv.org/abs/2203.13369. doi:10.48550/ARXIV.2203.13369.

[16] J. Sánchez-Junquera, B. Chulvi, P. Rosso, S. P. Ponzetto, How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants, Applied Sciences 11 (2021) 3610. URL: http://dx.doi.org/10.3390/app11083610. doi:10.3390/app11083610.

[17] J. S.-J. y Paolo Rosso y Manuel Montes-y-Gómez y Berta Chulvi, Masking and bert-based models for stereotype identication, Procesamiento del Lenguaje Natural 67 (2021) 83–94. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6379.

[18] R. O. BUENO, B. CHULVI, F. RANGEL, P. ROSSO, E. FERSINI, PAN 22 Author Profiling: Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO), 2022. URL: https://doi.org/10.5281/zenodo.6514916. doi:10.5281/zenodo.6514916.

[19] C. Van Hee, E. Lefever, V. Hoste, SemEval-2018 task 3: Irony detection in English tweets, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 39–50. URL: https://aclanthology.org/S18-1005. doi:10.18653/v1/S18-1005.

[20] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://www.aclweb.org/anthology/S19-2007. doi:10.18653/v1/S19-2007.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, CoRR abs/1706.03762 (2017). URL: http:

//arxiv.org/abs/1706.03762. arXiv:1706.03762.
[22] 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.
[23] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, L. Neves, TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification, in: Proceedings of Findings of EMNLP, 2020.
[24] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, arXiv preprint arXiv:2012.10289 (2020).
[25] S. Frenda, A. T. Cignarella, V. Basile, C. Bosco, V. Patti, P. Rosso, The unbearable hurtfulness of sarcasm, Expert Systems with Applications 193 (2022) 116398. URL: https://www.sciencedirect.com/science/article/pii/S0957417421016870. doi:https://doi.org/10.1016/j.eswa.2021.116398.
[26] J. Howard, S. Ruder, Universal Language Model Fine-tuning for Text Classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 328–339. URL: https://www.aclweb.org/anthology/P18-1031. doi:10.18653/v1/P18-1031.
[27] M. Girvan, M. E. Newman, Community structure in social and biological networks, Proceedings of the national academy of sciences 99 (2002) 7821–7826.
[28] L. C. Freeman, A set of measures of centrality based on betweenness, Sociometry 40 (1977) 35–41. URL: http://www.jstor.org/stable/3033543.
[29] T. N. Kipf, M. Welling, Semi-Supervised Classification with Graph Convolutional Networks, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017. URL: https://openreview.net/forum?id=SJU4ayYgl.
[30] S. Brody, U. Alon, E. Yahav, How attentive are graph attention networks?, 2021. URL: https://arxiv.org/abs/2105.14491. doi:10.48550/ARXIV.2105.14491.
[31] S. Bhagat, M. Burke, C. Diuk, S. Edunov, I. O. Filiz, Three and a half degrees of separation - meta research, 2016. URL: https://research.facebook.com/blog/2016/02/three-and-a-half-degrees-of-separation/.
[32] B. MacCartney, C. D. Manning, Modeling semantic containment and exclusion in natural language inference, in: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Coling 2008 Organizing Committee, Manchester, UK, 2008, pp. 521–528. URL: https://aclanthology.org/C08-1066.
[33] G. Hinton, N. Srivastava, K. Swersky, Lecture 6a overview of mini–batch gradient descent, Coursera Lecture slides https://class. coursera. org/neuralnets-2012-001/lecture,[Online (2012).
[34] D. Q. Nguyen, T. Vu, A. T. Nguyen, BERTweet: A pre-trained language model for English Tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.