MB-Courage @ EXIST: GCN Classification for Sexism Identification in Social Networks*

Rodrigo Wilkens and Dimitri Ognibene

University of Milano-Bicocca, Italy
{rodrigo.souzawilkens,dimitri.ognibene}@unimib.it

Abstract. We describe our approach (MB-Courage team) in the Sexism Identification in Social Networks Shared Task (EXIST). We submitted three runs for each task, two of them based on Graph Convolutional Neural Networks (GCN) exploring different edge creation strategies and one combining graph embeddings from different GCN through ensemble methods. In addition, we explored different GCN models and text-to-graph strategies. We identified that in Task 2 the models take advantage of the syntactic relationship between words encoded in the graph, while it did not strongly impact Task 1. Moreover, the models generalized the task while maintaining similar (in some cases better) results in the social network that was not used in training. On average, our best models performed similarly across languages and social media, ranking 37th (out of 72 runs) for Task 1 and 40th (out of 63) for Task 2.

Keywords: Graph Neural Network · MeanPooling · set2set · EXIST.

1 Introduction

Social media (SM) advent has been described as nothing less than a shift in the communication paradigm [1], or in other words, *the freedom to publish* marks the birth of a new era altogether [2]. There is obviously ample evidence of SM use's positive effects that go beyond just-in-time connectivity with a network of friends and like-minded people. However, far from creating a global space for mutual understanding, truthful and objective information, the large-scale growth of SM has also fostered negative social phenomena [24]. Therefore, threats on Social Media have become extensively studied, and Hate Speech (HS) is a frequent topic. Classification of text generated by social media, and Twitter, in particular [22], poses several significant challenges, between which their informality, noisiness, and limited size, leading first to a lack of features for classification, negatively affecting results, and second to a lack of context and ambiguity. Due to the tasks' difficulty and the necessity of generating different responses, some works discriminate the type and target of the hate, including sexism (e.g.,

* *IberLEF 2021, September 2021, Málaga, Spain.*

[35, 36, 34, 13, 9, 21, 25, 7, 19, 5, 29]). This line of research may support a measure of collective well-being on social media and social media governance strategies that aim at improving it [24].

Recently, the *sEXism Identification in Social neTworks* shared task (EXIST) [28] in the IberLEF 2021 [23] asked for systems capable of classifying sexism on microblogs. This shared task is divided into two classifying tasks: (1) the presence of sexist content and (2) the type of sexism. Moreover, this shared task provides posts written in both English and Spanish from Twitter (train and test) and gab.com (only test).

In this paper, we describe the participation of the MB-Courage team at the EXIST shared task. In this, we aimed to explore Graph Convolutional Neural Networks (GCN) from the perspective of a Graph Classification task. In other words, we focus on a GCN that models each document as an independent graph, which enables the inclusion of new documents without the need to retrain the model. Moreover, we examine different approaches for word association and encoding. The models are available at `https://github.com/rswilkens/courage-at-exist`. In specific, this paper is organized as follows. We start presenting initiatives for sexism classification, then describing GCN models employed in this work. We explore three different GCN models: MeanPool, SAGpool$_\text{h}$ and set2set. Section 3 presents the word encoding, text-to-graph strategies and the methodology to choose the submitted model. A discussion of the performance of the different models is presented in Section 4. Finally, in Section 5, we summarize our finds.

2 Related Work

There exist a broad number of possible approaches for carrying out HS classification. For example, Canós [6] resorted to Support Vector Machine (SVM) and TF-IDF. In complementary work, Liu et al. [18] employed SVM, Random Forests and Gradient Boosted Trees in an embedding space created by doc2vec [16] as well as a soft vote approach combining classifiers. Similarly, Shushkevich and Cardiff [30] used a blended model [33] combining Naïve Bayes and SVM. Gambino and Pirrone [10] proposed combining embeddings from fastText [4] and PoS tagging embedding as input features. Hoffmann and Kruschwitz [11] explored an ensemble approach of three SVMs trained with transformer document embeddings, document pool embeddings, and TF-IDF. Rodríguez-Sánchez et al. [27] compared different word encodings (i.e., TF-IDF and word embedding) and models (i.e., LR, SVM, and Bi-LSTM), including multilingual BERT (mBERT) fine-tuned to their dataset. In these works, the text is generally encoded, discarding the relations between the words. Exceptions, for example, are the RNN and LSTM models that encode the sequential information in their internal state, but this representation relies on learning to account for elements at a longer distance. Moreover, despite some methods that use contextualized embeddings from BERT, only a summary vector (which contains limited word information) is used for the classification [8].

GCN, convolutional networks that operate on graphs, can explicitly model the relationships between words by representing words as nodes and their relations as edges in the graph. Thus, in this work, we explore the GCN as a solution for identifying sexism on social networks.

2.1 Graph Neural Networks

Applying deep learning models to structured data such as graphs has been proposed in recent years. In particular, studies have focused on generalizing convolutional neural networks to graph data, which includes redefining the convolution and the downsampling (pooling) operations for graphs [17]. In a broad sense, graphical networks may be seen as a combination of simple map-reduce operations on graphs, corresponding to transformation and several aggregation operations on graphs [12]. The aggregation (also named as message passing) aims to aggregate multiple messages between a node and its context and reduce them into one element. The graph pooling aims to aggregate elements in a graph, reducing them into high-order graph-level representations. GCN mainly differs from other neural networks in the forward step that connects nodes by considering an adjacency matrix.

In terms of document classification, Kipf and Welling [14] compared the performance of GCN with different classification algorithms. Later, Yao et al. [37] proposed the TextGCN by extending the Kipf and Welling work by enriching the adjacent matrix. Although the interesting results achieved by these works, their models need to be retrained for every new document since the corpus is represented as a single graph and the model uses transductive learning. While this is not a general issue, it poses a problem for post-classification, given the speed at which new data is created on social media. In this work, we face the GCN models looking to overcome this limitation. In the remain of this section, we describe the three GCN architectures explored in this work.

The **MeanPool** is a naïve graph pooling model, which obtains graph representations by concatenating the mean pooling and max pooling results of GCNs, then the pooled graph feeds a classification layer [12]. At the most basic level, it aggregates the nodes' neighbors after the multiplication of the node features by the weights, then sums the bias and performs the activation function. This process is realized twice. Then, the node features are averaged, resulting in a vector representation with the same dimension of the node features for each graph.

Self-Attention GraphPooling (**SAGpool$_\mathbf{h}$**) [17] exploits the self-attention mechanism to distinguish nodes that should be dropped. The architecture comprises three blocks (convolutional layer, pooling layer and readout layer) applied sequentially. The outputs of these blocks feed a classification layer. The convolutional layer in SAGpool$_\mathrm{h}$ comprises a GCN and a self-attention mask (or intra-attention), aiming to focus on relevant input features [31].

Vinyals et al. [32], motivated by the impact of the sequence that the data is presented to the models, proposed **set2set** as an extension of the seq2seq framework. In set2set, the network's next state is the concatenation of LSTM and an attention readout. The latter is defined as the multiplication of a memory

vector by softmax of a scalar function (e.g., dot product) between the memory and the LSTM state. In this work, we use the set2set implementation of Hu et al. [12] in which the input of the set2set module is two stacked GCN.

3 Methodology

We trained all the models following a stratified 10-fold cross-validation approach (10% of each fold as validation). We firstly trained the GCN discussed in Section 2.1, using the parser and ngram text-to-graph approaches (see Section 3.1). In addition, inspired by the performance of ensemble methods in HS (e.g., [17, 33, 11]), we apply them, aiming to explore model complementarity. In this step, we extract the GCN output embedding (classification layer input) and train an XGBoost. All these processes resulted in 7 models for each fold and task (i.e., SAGpool_h, set2set and MeanPool using parser or ngram, and XGBoost). Figure 1 illustrates this process, highlighting the main steps and the different models.

3.1 Preprocessing and text to graph

We start our pipeline by cleaning and tokenizing the text before training the models. The cleaning step tokenizes the text, standardizes symbols, and replaces URL and emojis by the domain and the emoji textual representation based on the post's language. The text is then annotated with part-of-speech, morphological feats, dependency relation (tag and attachment), and NER tag using the stanza parser [26] aiming at a rich syntactic representation. We encoded each annotation in a one-hot vector and concatenate these five vectors. Aiming at a semantic representation, we also encoded the text with mBERT embeddings. For that, we take advantage of the *feature-extraction* pipeline from huggingface (`https://huggingface.co`), and truncate all sentences with more than 300 words long due to computational limitation. This threshold is particularly relevant for the posts from GAB in the test set. These encoding processes result in a vector of 1172 dimensions for each token (404 dimensions from syntactic annotation and 768 from mBERT).

Our text-to-graph approach takes the tokens' vectors as nodes, and we explore two different strategies for the edges. The **parser** strategy, inspired by syntacticGCN [20, 3], links nodes using the dependency attachment, while the **n-gram** strategy associates all words in a context window creating mesh-like graphs. Furthermore, we set the edge weight in both strategies as the cosine similarity between mBERT features at the node level.

3.2 Submission Selection

The models quickly converge to their best score in both tasks (Figure 2). For the cross-validation, we opted to use the best epoch of each model for each fold, taking the vote of the folds as the final classification. The XGBoost is trained using the 15th epoch of the GCNs for simplicity. Our preliminary tests (Table 1)

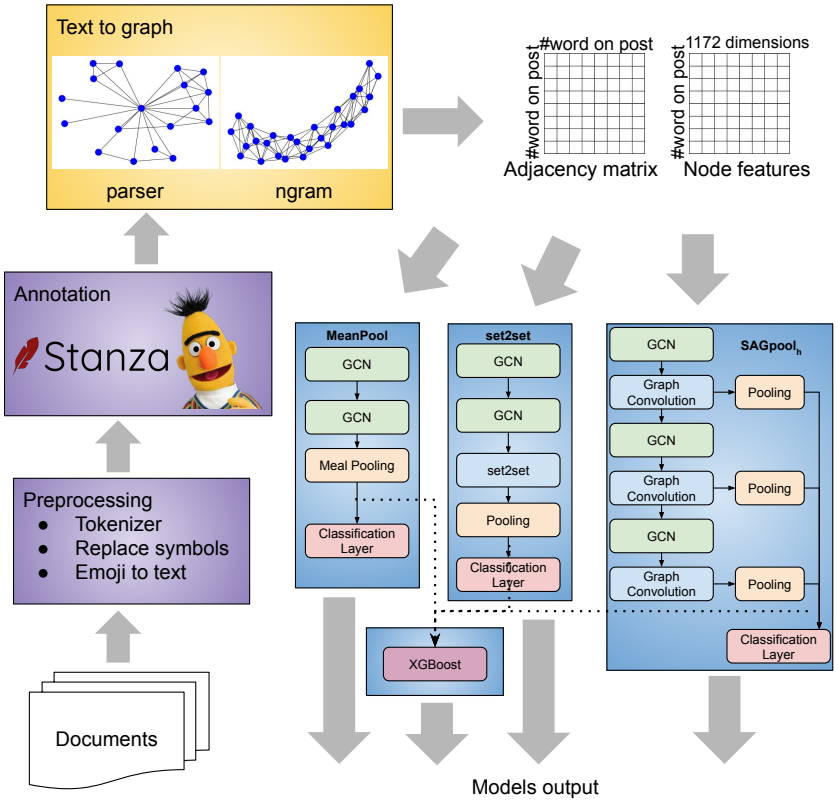


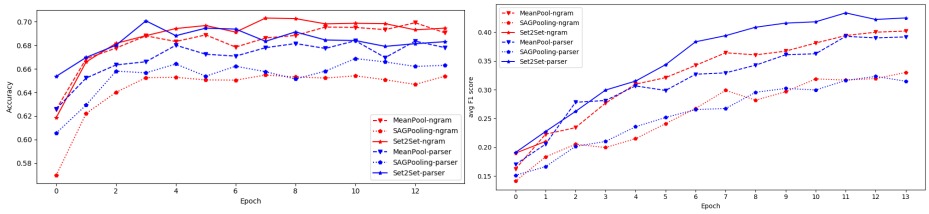
Fig. 1: System pipeline with the main steps (preprocessing, annotation and text to graph) and the 4 model explored, three GCN (MeanPool, SAGpool$_h$ and set2set) and one ensemble (XGBoost)

pointed XGBoost, which combines the architectures and the edge approaches, as the best model, in a second level, $set2set_{ngram}$ (Task 1) and $set2set_{ngram\dagger}$, and the $set2set_{parser}$ (Task 1 and 2) in a third place. We submitted the parser one for both tasks and the ngram (Task 1) and ngram† (Task 2). Table 1 shows the results of the different models (GNN and ensemble) using the training set.

4 Results

The official evaluation ranked XGBoost in the 53rd (Task 1) and 49th (Task 2) position, $set2set_{ngram}$ in 37th, $set2set_{ngram\dagger}$ in 44th, and $set2set_{parser}$ in 40th for Task 1 and 2.

The consistently poor performance of XGBoost is surprising considering it combines the set2set models. To understand this result, we evaluated the scores of the XGBoost's inputs (i.e., the output of the GCN), observing that the models



(a) Average accuracy in Task 1

(b) Average macro F1-score in Task 2

Fig. 2: Average result for each 10-fold cross-validation model

Table 1: Average (and standard deviation) scores per task, specifying the architectures and the strategies for the graph's edge. Best results are in bold, stars indicate submitted models, and the cross marks a mBERT only setup

Architecture	**Edge approach**	**Task 1**	**Task 2**
set2set	parser	0.707 (0.04)*	0.429 (0.04)*
set2set	ngram	0.713 (0.04)*	0.434 (0.04)
set2set	ngram†	0.71 (0.04)	0.453 (0.03)*
SAGpool_h	parser	0.683 (0.02)	0.342 (0.03)
SAGpool_h	ngram	0.674 (0.02)	0.358 (0.03)
MeanPool	parser	0.698 (0.02)	0.418 (0.02)
MeanPool	ngram	0.711 (0.01)	0.418 (0.02)
XGBoost	both	**0.823 (0.01)***	**0.669 (0.04)***

score worse than expected (e.g., $set2set_{ngram}$ scored 3.3% below the official score and $set2set_{parser}$ scored 4.5%). Looking carefully at the folds' learning curves, we noticed that the models quickly memorize the data. Hence the models used to train the XGBoost are biased towards the training set.

Moving forward in studying the set2set models, we evaluate the influence of the language and the source of the post in our results (see Table 2). Looking at the performance of the different languages in the same social media source, we perceive a substantial variation in the results when the source is GAB, but this is not the case when it is Twiter (except for $set2set_{ngram}$ in Task 2). This is expected since the models were trained only using tweets. For example, this is marked in the difference of 0.09 and 0.06 points of F-score in $set2set_{ngram}$, respectively, for English and Spanish for Task 2 as well as 0.07 and 0.04 in $set2set_{parser}$. Table 2 also highlights that $set2set_{ngram}$ performs better in Task 1 while $set2set_{parser}$ is better in Task 2.

Looking deeper at the models' performance, we calculate their F1 score for each class (Figure 3). The first observation of this evaluation is that our models perform similarly in Task 1, except GAB_{EN}. Moreover, the negative class also score similarly in both Task 1 and 2. However, that is not true for the positive classes in Task 2. In general, the models learned better *ideaological-inequality* class than the other four positive classes. In GAB_{ES}, they score better than the

negative class, which contains more instances, and even the binary classification. The models could not learn the *misogyny-non-sexual-violence* class correctly, presenting consistently poor scores in all studied cases. Moreover, we observed that $set2set_{parser}$ performs better than $set2set_{ngram}$ in all classes except for *objectification*.

Table 2: Results for Task 1 (accuracy) and 2 (macro F1) of the three submitted runs. Bold results indicate the best score for a task, and italic ones are the worst

		set2set					
		ngram	**ngram†**	**parser**		**XGBoost**	
Language	**Source**	**Task 1**	**Task 2**	**Task 1**	**Task 2**	**Task 1**	**Task 2**
Both	*Both*	0.714	0.449	0.708	0.459	0.680	0.421
EN	*GAB*	**0.745**	0.347	0.727	0.392	0.705	*0.297*
ES	*GAB*	0.665	0.419	0.687	0.423	*0.622*	0.340
EN	*Twitter*	0.711	0.438	0.699	0.464	0.689	0.440
ES	*Twitter*	0.723	**0.479**	0.717	0.461	0.682	0.440

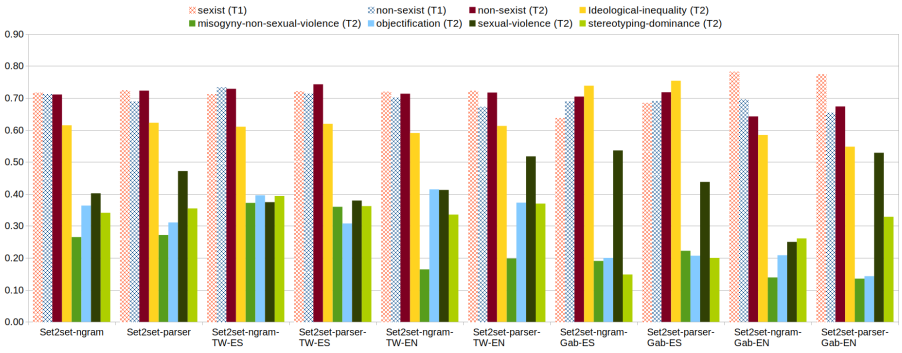


Fig. 3: Macro average F-score per class for each model and class, also discriminating the performance by source and language

5 Conclusions

This paper reported the participation of the MB-Courage team at the EXIST shared task. Our participation focused on using Graph Convolutional Neural Networks (GCN) as a solution for sexism identification. In this work, we explored three different GCN models (set2set, SAGpool_h and MeanPool) and their combination through ensemble methods. We also explored two different word

association strategies. Concerning the GCN models, we identified a poor performance of the $\text{SAGpool}_{\text{h}}$. This is probably due to an incompatibility between the limited text size and the attention strategy. On the other hand, the set2set models achieved our best scores, pointing to the importance of the word sequence in this task. As a general note about the models, we identified that they quickly learn the task, but at the same time, they also quickly memorize the training set. Assessing our best models' performance, we noticed a similar performance in sexism identification (Task 1) for both text-to-graph approaches, but that is not the case for identifying the type of sexism. This is probably due to a lack of language knowledge in our models since mBERT poorly encodes the language compared to language-specific versions of BERT [15]. As for the nodes association strategies explored in this work, we observed better scores using the n-gram association in Task 1, while in task 2 the best score came from the parser-based word association. Again, this is probably due to a need for more language information for identifying the type of sexism. Moreover, we see no significant improvement when we explicitly use syntactic annotations as features. This may have been caused by a mismatch of (sparse and dense) features or by BERT already encoding the same information.

The results obtained in this work point out that GCN may be a good solution for identifying threats on social media. Moreover, these models can easily model other social media aspects (e.g., users' relationships). However, in terms of NLP, it is not completely clear how these models perform for other threats, such as hate speech, and longer or less structured texts.

Acknowledgements This work has been developed in the framework of the project COURAGE - A social media companion safeguarding and educating students (no. 95567), funded by the Volkswagen Foundation in the topic Artificial Intelligence and the Society of the Future.

Bibliography

[1] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley (1999)

[2] Baeza-Yates, R., Ribeiro-Neto, B. (eds.): Modern Information Retrieval. Addison-Wesley, 2nd edn. (2010)

[3] Bastings, J., Titov, I., Aziz, W., Marcheggiani, D., Sima'an, K.: Graph convolutional encoders for syntax-aware neural machine translation. arXiv preprint arXiv:1704.04675 (2017)

[4] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)

[5] Bosco, C., Felice, D., Poletto, F., Sanguinetti, M., Maurizio, T.: Overview of the evalita 2018 hate speech detection task. In: EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. vol. 2263, pp. 1–9. CEUR (2018)

[6] Canós, J.S.: Misogyny identification through svm at ibereval 2018. In: Evaluation of Human Language Technologies for Iberian Languages: IberEval 2018 (IberEval@SEPLN). pp. 229–233 (2018)

[7] Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the 11th International AAAI Conference on Web and Social Media. pp. 512–515. ICWSM '17 (2017)

[8] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[9] Fersini, E., Rosso, P., Anzovino, M.: Overview of the task on automatic misogyny identification at ibereval 2018. Evaluation of Human Language Technologies for Iberian Languages: IberEval 2018 (IberEval@SEPLN) **2150**, 214–228 (2018)

[10] Gambino, G., Pirrone, R.: Chilab@ haspeede 2: Enhancing hate speech detection with part-of-speech tagging. In: 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2020 (2020)

[11] Hoffmann, J., Kruschwitz, U.: Ur nlp@ haspeede 2 at evalita 2020: Towards robust hate speech detection with contextual embeddings. In: 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2020 (2020)

[12] Hu, J., Qian, S., Fang, Q., Wang, Y., Zhao, Q., Zhang, H., Xu, C.: Efficient graph deep learning in tensorflow with tf_geometric. arXiv preprint arXiv:2101.11552 (2021)

[13] Jha, A., Mamidi, R.: When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In: Proceedings of

the Second Workshop on NLP and Computational Social Science. pp. 7–16 (2017)
[14] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
[15] Lavergne, E., Saini, R., Kovács, G., Murphy, K.: Thenorth@ haspeede 2: Bert-based language model fine-tuning for italian hate speech detection. In: 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2020. vol. 2765. CEUR-WS (2020)
[16] Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International conference on machine learning. pp. 1188–1196. PMLR (2014)
[17] Lee, J., Lee, I., Kang, J.: Self-attention graph pooling. In: International Conference on Machine Learning. pp. 3734–3743. PMLR (2019)
[18] Liu, H., Chiroma, F., Cocea, M.: Identification and classification of misogynous tweets using multi-classifier fusion. In: Evaluation of Human Language Technologies for Iberian Languages: IberEval 2018 (IberEval@SEPLN). pp. 268–273. CEUR Workshop Proceedings (2018)
[19] Mandl, T., Modha, S., Mandlia, C., Patel, D., Patel, A., Dave, M.: Hasoc-hate speech and offensive content identification in indo-european languages (2019)
[20] Marcheggiani, D., Titov, I.: Encoding sentences with graph convolutional networks for semantic role labeling. arXiv preprint arXiv:1703.04826 (2017)
[21] Mathur, P., Sawhney, R., Ayyar, M., Shah, R.: Did you offend me? classification of offensive tweets in hinglish language. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). pp. 138–148. Association for Computational Linguistics (2018), http://aclweb.org/anthology/W18-5118
[22] Michelson, M., Macskassy, S.A.: Discovering users' topics of interest on twitter: a first look. In: Proceedings of the fourth workshop on Analytics for noisy unstructured text data. pp. 73–80 (2010)
[23] Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Ángel Álvarez Carmona, M., Álvarez Mellado, E., de Albornoz, J.C., Chiruzzo, L., Freitas, L., Adorno, H.G., Gutiérrez, Y., Zafra, S.M.J., Lima, S., de Arco, F.M.P., Taulé, M. (eds.): Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings, 2021 (2021)
[24] Ognibene, D., Taibi, D., Kruschwitz, U., Wilkens, R.S., Hernandez-Leo, D., Theophilou, E., Scifo, L., Lobo, R.A., Lomonaco, F., Eimler, S., et al.: Challenging social media threats using collective well-being aware recommendation algorithms and an educational virtual companion. arXiv preprint arXiv:2102.04211 (2021)
[25] de Pelle, R.P., Moreira, V.P.: Offensive comments in the brazilian web: a dataset and baseline results (2017)
[26] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082 (2020)

[27] Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L.: Automatic classification of sexism in social networks: An empirical study on twitter data. IEEE Access **8**, 219563–219576 (2020)
[28] Rodríguez-Sánchez, F., de Albornoz, J.C., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., Donoso, T.: Overview of exist 2021: sexism identification in social networks. Procesamiento del Lenguaje Natural **67**(0) (2021)
[29] Sanguinetti, M., Comandini, G., Di Nuovo, E., Frenda, S., Stranisci, M., Bosco, C., Caselli, T., Patti, V., Russo, I., Pisa, I.: Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. In: Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online. CEUR. org (2020)
[30] Shushkevich, E., Cardiff, J.: Classifying misogynistic tweets using a blended model: The ami shared task in ibereval 2018. In: Evaluation of Human Language Technologies for Iberian Languages: IberEval 2018 (IberEval@SEPLN). pp. 255–259 (2018)
[31] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
[32] Vinyals, O., Bengio, S., Kudlur, M.: Order matters: Sequence to sequence for sets. arXiv preprint arXiv:1511.06391 (2015)
[33] Wang, S.I., Manning, C.D.: Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 90–94 (2012)
[34] Waseem, Z.: Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In: Proceedings of the First Workshop on NLP and Computational Social Science. pp. 138–142. Association for Computational Linguistics, Austin, Texas (November 2016), http://aclweb.org/anthology/W16-5618
[35] Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL student research workshop. pp. 88–93 (2016)
[36] Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL Student Research Workshop. pp. 88–93. Association for Computational Linguistics, San Diego, California (June 2016), http://www.aclweb.org/anthology/N16-2013
[37] Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 7370–7377 (2019)