IU-Percival: Linear Models for Sexism Detection

Notebook for the EXIST Lab at CLEF 2023

Elizabeth Gabel[1], Holly Redman[1], Daniel Swanson[1] and Sandra Kübler[1]

[1]*Indiana University, Bloomington, IN, USA*

Abstract

In this work, we detail the approach taken by the IU-Percival team in the EXIST 2023 shared task on sexism detection in English and Spanish tweets. Using syntactic n-grams generated by Universal Dependencies parsing as features, we train four classifiers: SVM, random forest, multi-layer perceptron, and single-layer perceptron. While we find that these four classifiers perform similarly, we focus our efforts on the single-layer perceptron as it not only performs slightly better than the rest but also boasts a much quicker training time. Our results for the development data indicate that our approach improves on previous non-deep learning approaches, and provide some support for continued examination of Universal Dependencies' application to Sexism Detection.

Keywords

sexism detection, classification, perceptron, syntactic n-grams

1. Introduction

This paper describes the contribution of IU-Percival to Task 1 of EXIST 2023 [1, 2]. The objective of this shared task is to promote the development of systems that are capable of automatically detecting sexist comments in social media in both English and Spanish. Task 1 is a binary task that involves labeling tweets as "sexist" or "non-sexist."

Sexist comments and various other forms of harmful language are rampant on social media, creating a monumental challenge for content moderation efforts. Therefore, the creation and refinement of efficient systems to automatically detect such messaging is essential. These systems must be able to identify both explicit and implicit sexism, as well as be able to contend with the many irregularities that exist in social media data, such as code-switching within comments, dialectal variations, misspellings, abbreviations, and the perpetually evolving colloquialisms that proliferate in online spaces.

Our work deviates from the transformer-based methods that have predominated the field in recent years. We focus instead on pre-transformer methods, leveraging Universal Dependencies (UD) [3] parsing to generate syntactic features. We train SVM, random forest, single-layer perceptron, and multi-layer perceptron classifiers. We find that these traditional methods

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

eligabel@iu.edu (E. Gabel); hredman@iu.edu (H. Redman); dangswan@iu.edu (D. Swanson); skuebler@indiana.edu (S. Kübler)

http://dangswan.com/ (D. Swanson); https://cl.indiana.edu/~skuebler/ (S. Kübler)

0000-0002-9847-8111 (D. Swanson); 0000-0003-0885-5436 (S. Kübler)

perform reasonably well on the task. This approach was developed during a course on machine learning in NLP.

The rest of this work is structured as follows: section 2 introduces related work, focusing on sexism detection with non-transformer approaches; section 3 describes our research methods, including pre-processing steps, feature selection, and our choice of classifiers; section 4 presents our results; section 5 provides a discussion of our findings; and section 6 shares concluding remarks and discusses plans for future work.

2. Related Work

The EXIST shared task, starting in 2021, remains one of the primary sources of research into sexism detection in tweets. The majority of submissions to EXIST have focused on transformer-based methods, with little attention paid to the potential of more traditional machine learning methods.

Rodríguez-Sánchez et al. [4] explored a variety of machine learning methods for the task of sexism detection. They compared the performance of logistic regression, SVM, random forest, bi-LSTMs, and mBERT on sexism detection in Spanish tweets. They found the neural models to be slightly better than the non-neural machine learning algorithms at detecting sexism in the dataset, although random forest achieved the highest precision. The bi-LSTM models were on par with mBERT in terms of F1, accuracy, precision, and recall.

Rizvi and Jamatia [5] participated in the 2022 EXIST shared task [6]. They experimented with logistic regression, Naive Bayes, and SVM systems and found that the logistic regression model worked best for both Spanish and English on both tasks. They used TF-IDF unigram and bigram representations as features for all three models. While their submission ultimately ranked 17th out of 19 submissions in the competition, with an official F1-score of 70.65% overall, their approach showed promise among the few submissions that did not implement pretrained transformer-based models.

Moldovan et al. [7] addressed the issue of sexism in Romanian. They used logistic regression, SVM, random forests, Ro-BERT, and mBERT to classify Romanian tweets as sexist or non-sexist. They used BOW-based representations, TF-IDF word representations, and sentence representations generated by mBERT and Ro-BERT as features for the non-neural models. The best performance was achieved with a fine-tuned Ro-BERT model; however, the best recall for non-sexist tweets was achieved by the random forest classifier using TF-IDF-based word representations, by a significant margin.

Related to the topic of sexism detection is abusive language detection. Steimel et al. [8] investigated abusive language detection in English and German tweets using topic modeling and a number of neural and non-neural classifiers. They found that SGBoost performed best on the English data while SVMs performed best on the German data. They also found that different sampling methods to address class imbalance led to drastically different outcomes regarding the two data sets. Their work provides evidence that the best classifier and techniques for one language cannot be assumed to perform well for other languages, even if the data sets share similarities. Thus, it is important to experiment with a variety of methods when handling multilingual data.

Table 1
Distribution of labels in the training data by language. Numerical labels indicate the number of annotators out of six who labeled a given tweet as containing sexist language.

Label	English	Spanish
0	768	666
1	545	526
2	420	442
3	390	466
4	423	487
5	392	580
6	322	493
Majority no-sexist	1 733	1 634
Tie	390	466
Majority sexist	1 137	1 560

3. Methods

3.1. Data

The dataset was provided by the organizers of EXIST 2023. It is comprised of a selection of Spanish and English tweets, annotated for sexism. The training set contains tweets collected between the 1st of September 2021 to 28th of February of 2022, the development set contains tweets from the 1st to 31st of May of 2022, and the test set contains tweets collected from the 1st of August 2022 to 30th of September.

In order to avoid author bias, the final selection contains one tweet per author. Additionally, all tweets containing less than five words were removed[1].

The training data contains 3 260 English tweets (34.88% sexist) and 3 660 Spanish tweets (42.62% sexist). Table 1 shows the distribution of English and Spanish tweets along with their in-language sexist to non-sexist distribution for the training data.

Since the tweets are not provided with a single gold label but rather with a set six labels given by different annotators, it is necessary to deal with ties. The official evaluation script of EXIST 2023 for the hard-hard evaluation simply discards ties. We, on the other hand, have chosen to include them in the sexist category for training, in order to boost the minority class.

3.2. Pre-Processing

We extracted the syntactic features using UDPipe2 [9], specifically using the available pre-trained models for the largest English and Spanish treebanks: English-EWT [10] and Spanish-AnCora [11]. The English EWT contains text from four different web genres; the AnCora treebank is based on news texts. Unfortunately, there does not exist a Spanish UD treebank based on social media data.

[1]More detailed information about the dataset collection can be found at the EXIST 2023 website: http://nlp.uned.es/exist2023/.

Table 2
Total number of syntactic n-grams used from each parser and means of feature selection.

	English	Spanish	Combined
Full	319 923	320 471	591 469
Trimmed	11 983	12 405	20 254
Selected	500	500	500

Since UDPipe2 already performs a significant amount of pre-processing internally, ours was fairly minimal. Links were replaced with URL and user mentions were replaced by USER in English and NOMBRE in Spanish. Additionally, spaces were added after periods to correct for some errors in sentence segmentation.

Many of the tweets in the data set, particularly those labeled as Spanish, contained code-switching. Additionally, some tweets labeled as Spanish contained only English text. Because of this, we opted to pass every tweet through both the English parser and the Spanish parser.

3.3. Features

After parsing, we extracted syntactic unigrams, bigrams, and trigrams. Unigrams consist of a lower-cased lemma and a part of speech tag, such as `user` NOUN. Bigrams consist of a word, its parent word, and the relation between them, such as `user` NOUN `nsubj ignora` VERB or just the unigram and relation if the word is the root of the sentence (hence `ignora` VERB `root`). Trigrams, similarly, concatenate a word, its parent, and its grandparent.

We tested three methods of selecting training features. The first method, "Full", uses all syntactic n-grams present in the training data. The second, "Trimmed", uses only n-grams which are present in both the training set and the development set. Finally "Selected" uses χ^2 to choose the 500 most informative features. The total number of features for each parser and selection method is listed in Table 2.

3.4. Model

We trained an SVM, a random forest classifier, a multi-layer perceptron, and single-layer perceptron trained with stochastic gradient descent on the Full feature set from both parsers combined, from the scikit-learn toolkit [12]. Our initial experimentation produced the scores listed in Table 5. The four architectures all gave roughly equivalent performance, so we chose to focus on the single-layer perceptron, which slightly outperformed the others and also had a substantially shorter training time.

For the final model we performed hyperparemeter tuning over the learning rate and the random seed. We used logistic regression, l2 regularization, and a constant learning rate schedule.

Table 3
Official Rankings from EXIST 2023 Task 1. Rankings by evaluation metric and run, underlining indicates best performing IU-Percival run in each language

Run	Lang	Hard-Hard	Hard-Soft	Soft-Soft
IU-Percival_1	All	45	48	39
	English	47	47	<u>39</u>
	Spanish	49	51	42
IU-Percival_2	All	46	46	<u>38</u>
	English	50	54	40
	Spanish	47	46	<u>37</u>
IU-Percival_3	All	51	52	40
	English	58	60	42
	Spanish	45	47	38

Table 4
Official Results from EXIST 2023 Task 1. Columns labeled "Norm" indicate ICM scores rescaled so that the best performing submission had 1.00 and the worst had 0.00. Bolded values are the best performing in their column.

		Hard-Hard			Hard-Soft		Soft-Soft	
Run	Lang	ICM-Hard	Norm	F1	ICM-Soft	Norm	ICM-Soft	Norm
IU-Percival_1	All	0.3024	0.5587	69.71	-0.4612	0.4217	-0.4612	0.4217
	English	**0.3655**	0.6264	68.57	-0.4952	0.4792	-0.4952	0.4792
	Spanish	0.2273	0.4885	70.50	-0.5085	0.3629	-0.5085	0.3629
IU-Percival_2	All	0.2964	0.5549	69.81	-0.4435	0.4246	-0.4435	0.4246
	English	0.2998	0.5865	66.84	-0.6435	0.4578	-0.6435	0.4578
	Spanish	0.2737	0.5192	**71.91**	**-0.3556**	0.3898	**-0.3556**	0.3898
IU-Percival_3	All	0.2675	0.5365	69.07	-0.5491	0.4075	-0.5491	0.4075
	English	0.2363	0.5479	65.72	-0.8610	0.4264	-0.8610	0.4264
	Spanish	0.2827	0.5252	71.68	-0.3572	0.3895	-0.3572	0.3895

4. Results

4.1. Official Results

We show the official rankings of our three submitted models on the test set in Table 3 and the official results in Table 4. Our first model, IU-Percival_1, consists of two single-layer perceptrons (one for English and one for Spanish) which were trained on all syntactic unigrams, bigrams, and trigrams. The second model, IU-Percival_2, is comprised of one single-layer perceptron for both languages and was trained on the same syntactic features. The final model, IU-Percival_3, is one single-layer perceptron trained on the same syntactic features, but excluding all n-grams which do not also occur in the dev set provided for the task.

We report ICM scores [13], along with the F1-score for the positive class in the Hard-Hard evaluation.

In terms of rankings, our multilingual model using all features, IU-Percival_2, provided

Table 5
F1 scores on the development set (with tied annotations included) for the internal experiments.

Classifier	F1
Multi-Layer Perceptron	73.93
Random Forest	73.15
Single-Layer Perceptron	**74.12**
SVM	73.53

our best ranking for All (both languages) and Spanish, while the combination of monolingual models, IU-Percival_1, ranked the best of our models for English. Our best rankings were under the Soft-Soft evaluation for all languages and models.

Our results, on the other hand, show that our submitted models all performed best in terms of ICM values under the Hard-Hard evaluation, in particular the ICM-Hard Normalized (although our non-normalized ICM-Hard values also outperformed the non-normalized ICM-Soft). Our best overall model in terms of ICM is IU-Percival_1 for English, although this is not our best ranked model. This difference can partially be explained by the lower number of submissions for the Soft-Soft evaluations.

For all models, the classifier performed worse on Spanish data than both All data and only English data. Although IU-Percival_3 had better results for Spanish for both non-normalized Hard and Soft evaluations (0.2827 ICM-Hard compared to 0.2363 English and 0.2675 All), following normalization, Spanish once again had the lowest results. In fact, following normalization, our classifier performed best on English for all models and evaluations.

Interestingly, our best scores for All and English come from IU-Percival_1, or the run with all n-grams, while our best score for Spanish comes from IU-Percival_3, which filters n-grams not occurring in the developmental set. Additionally, while the evaluation metric for the task is ICM, it is still of note that the highest F1 scores for each run (70.50, 71.91, 71.68) were all for Spanish, while the lowest F1 scores (68.57, 66.84, 65.72) were for English.

4.2. Evaluation on Development Set

We performed more extensive experiments on the development set. For these experiments, we split the official training set into 80% for training and 20% for validation, and we tested on the official development set. This was done to avoid an optimistic evaluation since one of the feature selection methods would otherwise have been operated on the same data set on which we also tested.

We first tested all classifiers considered for the task. The results are shown in Table 5. They show that the single-layer perceptron outperforms the other classifiers in terms of macro-averaged F1. For this reason, we used this classifier for the remaining experiments.

We next investigated whether to use the features from the English or Spanish parser since we parse each tweet with both parsers, or a combination of both. Additionally, we compared the Trimmed feature selection method to the full feature set. The results of these experiments are shown in Table 6. In all but one case, the best performing F1 scores corresponded with the best performing ICM scores.

Table 6
Results of using syntactic features from the English and Spanish parser, a combination of both, in combination with feature selection. Asterisks indicate instances where the best F1 score and ICM score are from different runs.

		English Data		Spanish Data		Combined Data	
Parser	Features	F1	ICM	F1	ICM	F1	ICM
English	Full	**73.45**	**0.3067**	70.24	0.2333	71.48	0.2633
	Trimmed	73.36	0.3058	70.58	0.2433	71.50	0.2645
Spanish	Full	72.02	0.2630	70.27	0.2339	70.67*	0.2409*
	Trimmed	70.49	0.2147	68.70	0.1838	70.55	0.2341
Both	Full	72.90	0.2912	**71.30**	**0.2684**	**72.41**	**0.2930**
	Trimmed	72.80	0.2903	69.76	0.2189	71.85	0.2756
Both	Merged					**72.62**	**0.2999**

Across the board, the best performing classifier for English data is the one trained on English features; for the Spanish data, the best model uses the combined English and Spanish features. Our best performing system on the English data alone had an ICM of 0.3067 and the best for the Spanish data alone got 0.2684. Evaluating these two together on the full development set gave an ICM of 0.2999.

It is interesting to note that every setting achieves a lower score when trained and tested only on Spanish data than when trained and tested only on English data.

5. Discussion

For EXIST 2022, Rizvi and Jamatia [5] (who also used linear models) ranked 38th of 45 submissions in Task 1, and of those who submitted papers ranked 17th out of 19. Their best model had an F1 score of 70.65 on Combined data, 70.84 on only English data, and 70.24 on only Spanish data.

Because we do not use the same classifier as Rizvi and Jamatia, nor do we have the same train or test sets, it is impossible to draw a direct comparison between their results and ours. However, the fact that our system reached comparable scores on similar datasets indicates the potential value of syntactic features extracted through Universal Dependencies for the problem of sexism detection.

Additionally, although our classifier performed worse on Spanish data and Combined data (in keeping with the findings of Steimel et al. [8], which indicate that classifiers do not perform universally well on different languages), the presence of both Spanish and English features extracted using UDPipe improved the performance for Spanish and Combined results.

Since we elected to pass every tweet through both the Spanish parser and the English parser, all tweets, regardless of original language, were used to extract both Spanish and English syntactic features. The universal decline in our classifier's performance on Spanish data may be due to a few different reasons. First, it is possible that the single-layer perceptron performs generally worse on Spanish data than on English data. Regarding the lower performance by the classifier trained using only Spanish features, it is also possible that our decision to use

Spanish-AnCora– due to the lack of Spanish treebanks for online language– resulted in less accurate features than those extracted using English-EWT (which, while not including tweets, does include more casual language and language pulled from some social media). The dataset is fairly well balanced and has more Spanish tweets than English tweets, so issues with the data distribution are an unlikely cause. Regardless, the best models for Spanish data are those that use features from both English and Spanish, indicating that the inclusion of features from multiple languages is valuable in cases where code-switching is prevalent in the data, and that using Universal Dependencies to extract additional syntactic features may increase classifier performance for Multilingual Sexism Detection. While the Spanish data set contains many instances of Spanish-English code-switching, the English data set does not, which explains why Spanish features do not improve performance on English data. Future research using Universal Dependencies syntactic features for similar purposes will need to consider these factors when making decisions about which treebanks are the most appropriate to use for a given dataset.

Here we would like to provide some comments on our perspective as students participating in a shared task, as this work also served as a final project for a course on machine learning for computational linguistics. We had all read literature and completed assignments based on various past shared tasks but had not, until now, gone through the process ourselves. Unsurprisingly, we encountered some challenges along the way. One challenge was simply deciding on a model and feature set, having had limited personal experience with knowing what might work well for this type of task. We also decided against using transformer models even though that is what the vast majority of other participants of this task had done. This decision made our work more difficult in some aspects because there was less of a precedent, and it forced us to be creative in our feature selection. We also had to temper our expectations regarding our eventual placement on the leaderboard. From the beginning, our aim was not to achieve state-of-the-art performance but to investigate less-studied avenues.

Another source of difficulty for us was collaboration on the practical side, coordinating coding and running experiments with different types of expertise in the group. Finally, we realized rather late in our experimentation process that we had made a serious methodological error; we had created a feature set based on the presence of certain features in both the train and dev set, which we had then used to test on the dev set. Luckily, we were able to rectify this before submission. Overall, we are grateful to have had this opportunity to participate in EXIST and engage with this important field of research.

6. Conclusion and Future Work

By training four classifiers, we have shown that decent results can be obtained for sexism detection in English and Spanish tweets using machine learning approaches that are not based on transformers. Surprisingly, our best performing system for both datasets has been a single-layer perceptron, which has an extremely simple architecture and is fast to train.

Additionally, we have examined the potential of training on syntactic features extracted using Universal Dependencies for the problem of sexism detection, and provided a baseline for future work examining this approach.

This study has several limitations which provide opportunities for future research. These

include the treebanks available to us in both English and Spanish integrated with UDPipe2. While there is a tweet treebank for English, it is not an official Universal Dependencies treebank, and there is no treebank for Spanish which includes internet/online language. Additionally, we encountered some errors in parsing, which could be resolved with additional pre-processing, particularly for the Spanish data.

Initially, our pre-processing involved replacing usernames in both the Spanish and English data with the token USER. However, upon examination of the data, we found that the parser we used for Spanish incorrectly tagged USER as a verb. Changing usernames to NOMBRE instead improved the accuracy of the parser. Further examination of the parsed data may reveal other insights that could be utilized to improve parsing accuracy, which would result in better syntactic features.

Our initial attempts at feature selection with scikit-learn were not fruitful, and due to time limitations we were unable to run sufficient feature selection experiments. Therefore, we plan to continue refining our feature selection methods by adjusting the number of features and utilizing different feature selection methods.

Another avenue for future investigation is an ensemble method combining multiple classifiers. At this time, we do not know enough about the specific strengths and weaknesses that each of our trained classifiers has, but determining this and combining classifiers could result in a more robust system.

Acknowledgments

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

[1] L. Plaza, J. Carrillo-de-Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Thessaloniki, Greece, 2023.

[2] L. Plaza, J. Carrillo-de-Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization (Extended Overview), in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[3] J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, D. Zeman, Universal Dependencies v2: An evergrowing multilingual treebank collection, in: Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC), Marseille, France, 2020, pp. 4034–4043. URL: https://aclanthology.org/2020.lrec-1.497.

[4] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576. doi:10.1109/ACCESS.2020.3042604.

[5] A. Rizvi, A. Jamatia, NIT-Agartala-NLP-Team at EXIST 2022: Sexism identification in social networks, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF), volume Vol-3202, A Coruña, Spain, 2022. URL: http://ceur-ws.org/Vol-3202/. doi:urn:nbn:de:0074-3202-9.

[6] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2022: sEXism Identification in Social neTworks, Procesamiento del Lenguaje Natural 69 (2022) 229–240. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6443.

[7] A. Moldovan, K. Csürös, A.-m. Bucur, L. Bercuci, Users hate blondes: Detecting sexism in user comments on online Romanian news, in: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Seattle, Washington (Hybrid), 2022, pp. 230–230. URL: https://aclanthology.org/2022.woah-1.21. doi:10.18653/v1/2022.woah-1.21.

[8] K. Steimel, D. Dakota, Y. Chen, S. Kübler, Investigating multilingual abusive language detection: A cautionary tale, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), Varna, Bulgaria, 2019, pp. 1151–1160. URL: https://aclanthology.org/R19-1132. doi:10.26615/978-954-452-056-4_132.

[9] M. Straka, UDPipe 2.0 prototype at CoNLL 2018 UD shared task, in: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 197–207. URL: https://www.aclweb.org/anthology/K18-2020. doi:10.18653/v1/K18-2020.

[10] N. Silveira, T. Dozat, M.-C. de Marneffe, S. Bowman, M. Connor, J. Bauer, C. D. Manning, A gold standard dependency corpus for English, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), 2014.

[11] M. A. M. Mariona Taulé, M. Recasens, Ancora: Multilevel annotated corpora for catalan and spanish, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 2008. Http://www.lrec-conf.org/proceedings/lrec2008/.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[13] E. Amigó, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022, pp. 5809–5819.