Improving Aggressiveness Detection using a Data Augmentation Technique based on a Diffusion Language Model

Antonio D. Reyes-Ramírez[α], **Mario Ezra Aragón**[β],
Fernando Sánchez-Vega[α γ], **A. Pastor López-Monroy**[α]
[α] Mathematics Research Center (CIMAT), Gto, Mexico
[β] Centro Singular de Investigación en Tecnoloxias Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain
[γ] Consejo Nacional de Humanidades, Ciencias y Tecnologías, México
{antonio.reyes, fernando.sanchez, pastor.lopez}@cimat.mx, ezra.aragon@usc.es

Abstract

Cyberbullying has grown in recent years, primarily attributed to the proliferation of social media users. This phenomenon manifests in various forms, such as hate speech and offensive language, increasing the necessity of effective detection models to tackle this problem. Most approaches focus on supervised algorithms, which have an essential drawback—they heavily depend on the availability of ample training data. This paper attempts to tackle this insufficient data problem using data augmentation (DA) techniques. We propose a novel data augmentation technique based on a Diffusion Language Model (DLA). We compare our proposed method against well-known DA techniques, such as contextual augmentation and Easy Data Augmentation (EDA). Our findings reveal a slight but promising improvement, leading to more robust results with very low variance. Additionally, we provide a comprehensive qualitative analysis using classification errors and complementary analysis, shedding light on the nuances of our approach.

1 Introduction

Social networks have fundamentally transformed human communication. Initially conceived as platforms for sharing ideas, experiences, and opinions, popular networks like Facebook, Twitter, Reddit, and others emerged. However, these platforms have also become arenas for intolerance, hateful comments, aggression, and harassment. Consequently, detecting hate speech has become a significant concern for researchers in natural language processing (NLP) due to its harmful societal impact, affecting the interactions within online communities (Burnap and Williams, 2015). The intolerance and aggression displayed by certain users harm the experiences of other individuals or entire online groups.

As the frequency of online interactions continues to rise, the necessity for automated systems to detect and handle abusive language becomes increasingly critical (Nobata et al., 2016). Currently, many approaches view this challenge as a supervised classification task, encountering difficulties such as requiring extensive labeled datasets to train the models. However, creating these new labeled data is often costly and demands significant human resources. To address this obstacle, an alternative solution involves using data augmentation techniques, which entails generating synthetic data from existing datasets. This approach was initially proposed for computer vision tasks and has been adapted for text processing. However, many existing methods provide little diversity in the data generated. For example, techniques like Easy Data Augmentation (Wei and Zou, 2019a), contextual augmentation (Kumar et al., 2020), (Kobayashi, 2018), and back-translation (Sennrich et al., 2015) make only a small amount of changes to the original data.

We introduce an innovative data augmentation approach leveraging a diffusion language model to tackle these challenges. We propose to use DiffuSeq (Gong et al., 2022), a non-autoregressive model employing a sequence-to-sequence framework, with the added capability of conditional generation based on input sequences. This unique setup enables us to generate samples conditioned on their respective classes from the original dataset. Compared to traditional methods, our diffuser is sure to generate conditional and more diverse text. We compare our proposed technique and widely used data augmentation methods like contextual augmentation (Devlin et al., 2019) and EDA (Wei and Zou, 2019b). The key contributions of this research are summarized as follows:

- A comparative analysis of the data augmentation methods. Presenting the advantages of using diffusers in text data augmentation tasks.
- A qualitative analysis of errors in classifica-

Proceedings of the The 8th Workshop on Online Abuse and Harms (WOAH), pages 171–177
June 20, 2024 ©2024 Association for Computational Linguistics

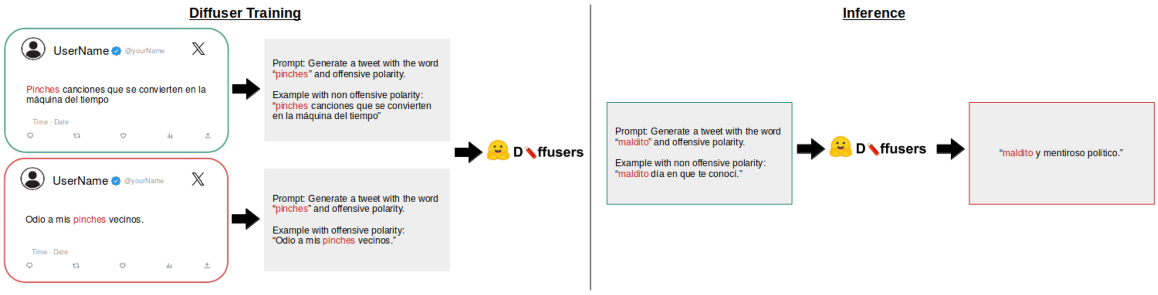

Figure 1: Training and inference process for our diffusion language model. We display synthetic examples in Spanish.

tion to try and understand the limitations of our approach.

2 Related work

This section presents an overview of the approaches prop task of hate speech detection. Most research on identifying abusive language tackles the problem as text categorization (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018), wherein posts, comments, or documents are assigned to predefined categories based on their content. Furthermore, most of these works primarily use English datasets due to their widespread availability. A diverse array of features has been explored to detect abusive language. Initial efforts relied on manually crafted features such as bag-of-words representations, alongside syntactic and semantic features, to train machine learning algorithms including Linear Regression, Support Vector Machine (SVM), Random Forest, and Naive Bayes classifiers (Magu et al., 2017; Robinson et al., 2018; Frenda et al., 2019; Vidgen and Yasseri, 2020; Martins et al., 2018; Madukwe and Gao, 2019; Rai et al., 2020; Pariyani et al., 2021). Research findings suggest that lexical methods have the potential to identify hate speech. However, their decisions are primarily based on single words or small context windows. We want to explore techniques that can account for a significant amount of context for each word.(Koushik et al., 2019; Watanabe et al., 2018; Abro et al., 2020).

Recent research has focused on leveraging deep learning to improve the ability of classifiers to identify abusive language, bypassing the need for manual feature engineering. Convolutional Neural Networks (CNNs) have been a popular approach, as demonstrated by Gambäck and Sikdar (2017); Mozafari et al. (2020) who employed deep contextualized word representations alongside a CNN for supervised fine-tuning. Furthermore, Zhang et al. (2018) incorporated a Gated Recurrent Unit (GRU) layer within their CNN model, benefitting feature extraction and sequential information. Recently, pre-trained language models, such as ELMO, GPT-2, and BERT, have been successfully integrated into abusive language detection systems (Liu et al., 2019; Nikolov and Radivchev, 2019). These models leverage pre-existing knowledge from vast amounts of text data, demonstrably improving detection performance.

As previously mentioned, limited training data presents a significant challenge when training our models, particularly for tasks requiring nuanced understanding. With a restricted pool of examples, models struggle to generalize and perform adequately on novel data. Data augmentation techniques offer a solution by artificially expanding the training set, effectively increasing data size and diversity. Current research on hate speech detection, particularly for non-English languages, lacks exploration of these techniques. This presents a significant opportunity to investigate the effectiveness of data augmentation for hate speech detection.

3 Methodology

Our methodology consists of two parts. The first part trains a diffusion language model to generate synthetic data conditioned to its class (aggressive or not aggressive). The second part augments our original training data using the diffusion model just trained. Then, it trains an aggressiveness classifier on the augmented dataset. Figure 1 presents this whole training and inference process.

3.1 Training a Diffusion Language Model

To train our diffusion model, we create a dataset consisting of sequence pairs (source, target). We want to generate a target sequence that contains specific bad words because we consider those words relevant to the aggressive class. We set a bad word and an example in our source sequence to achieve this. We then follow the next steps to create this new dataset.

1. First, we take our training dataset and determine their most relevant words. As a metric, we use the chi-squared score. We create a list of those words that are offensive, too. We denote it as S. For each word w in S, we create a set T_w that consists of all training tweets that contain w.

2. Given a word w in S, we take a pair of tuples $(x_i, y_i), (x_j, y_j)$ from T_w, where x_i, x_j are tweets and y_i, y_j are their labels. We set the source sequence as: "Generate a tweet with the word w and the y_j polarity. Example with a polarity y_i: x_i". The target sequence consists only of x_j. In Figure 1, we can observe a concrete example.

3.2 Data selection

The diffusion model generates data of different qualities. We aim to understand if a higher or lower data quality leads to a better classifier performance. We fine-tune a pre-trained language model h, RoBERTuito (Pérez et al., 2021), on our training set to measure data quality. Then, we generate a synthetic dataset three times larger than the original. We sort this data regarding its confidence score given by our base model (RoBERTuito). Given a synthesized sentence (x_i', y_i'), we first verify that $\arg\max h(x_i') = y_i'$, and then use h confidence score as a rank for (x_i', y_i'). We define the confidence score as the maximum predicted probability $\max h(x_i')$. We split this sorted set into three pieces that we call low, middle, and high-confidence datasets.

4 Experimental Settings

4.1 Dataset

We consider the MEX-A3T dataset for the aggressiveness detection task (Aragón et al., 2020). This dataset consists of Mexican Spanish tweets and two classes: aggressive and not-aggressive. Table 1 shows the distribution of this dataset.

4.2 Compared Methods

We compare our DA method with two traditional DA techniques. **Contextual augmentation** (Kobayashi, 2018): We use a pre-trained language model, RoBERTuito, for this method. We consider two actions at the word level: insert and substitute. **Easy Data augmentation**(Wei and Zou, 2019a): We consider three main actions at the word level: random swap, random delete, and synonym substitution. We use *nlpaug* library (Ma, 2019) to implement both methods with default hyperparameters.

Class	Train	Test
Not aggressive	5222	2238
aggressive	2110	905

Table 1: Statistic for the MEX-A3T dataset.

4.3 Diffuser training setups

We train a DiffuSeq model from scratch using the following parameters: 2000 diffusion steps, a learning rate of 0.001, a batch size of 100, 100000 learning steps, and a sequence length of 128.

4.4 Classifier

We choose RoBERTuito as our classifier. We fine-tune it in our original dataset and every augmented dataset.

5 Results and Analysis

Table 2 shows the classifier's results trained on several augmented datasets generated by our diffusion model. We also compare our method with standard data augmentation techniques, such as Contextual Augmentation and Easy Data Augmentation. We run each experiment 5 times with a set of 5 random seeds. Table 2 displays their average and standard deviation.

5.1 Complementary analysis

Considering only one method, the best-performing classifier is achieved using the middle-confidence diff augmented data. However, we can observe that individual data augmentation techniques only get a slight improvement concerning our baseline. To determine a more robust model, we look for the most effective way to combine our best-performing models: middle-confidence diff and synonym substitution. We try two ways to accomplish this objective. The first consists of making an ensemble of the two models. We only calculate the average of the two predictions. The second consists of generating different combinations of augmented datasets. We

Method	F1-positive	F1-negative	F1-Macro	Accuracy
W/o augmentation	82.6±0.78	92.72±0.33	87.738±0.54	89.736±0.46
Low-confidence diff	82.09±0.65	92.762±0.13	87.426±0.37	89.692±0.22
Middle-confidence diff	82.772±0.45	93.02±0.15	87.896±0.26	90.068±0.19
High-confidence diff	82.376±0.54	92.898±0.2	87.638±0.34	89.876±0.27
Contextual aug: substitute	82.47±0.73	92.728±0.52	87.6±0.61	89.724±0.63
Contextual aug: insert	82.44±0.45	92.694±0.35	87.568± 0.37	89.684±0.4
EDA	82.168±0.69	92.634±0.46	87.402±0.57	89.578±0.58
Synonym	82.48± 0.39	92.934±0.37	87.706±0.31	89.93±0.4
Combination 1	82.684±2.73	93.028±1.7	87.858±1.87	90.058±1.98
Combination 2	82.644±4.37	92.902±1.87	87.774±2.84	89.928±2.44
Combination 3	81.804±4.89	92.422±2.11	87.114±2.84	89.304±2.46
Ensemble	**83.41**±4.43	**93.166**±5.27	**88.288**±3.69	**90.322**±4.59

Table 2: Classification results for the aggressiveness detection task. We display the average and standard deviation of five runs. Results include an ensemble model and three models trained on different combinations of middle confidence-diff and synonym substitution datasets.

Example	GT	MC diff	Syn
If there's something that really annoys me, it's the pr****tutes who think they're saints, RIDICULOUS	0	1	1
I see this guy as kind of effeminate. It's like he resembles Fabiruchis	1	0	0
For your understanding, Sergio used the terms 'h**ker' and 'sl*t', but he didn't address them to the women with the intention of insulting them.	0	1	0
They have no morals or shame!!!	1	1	0

Table 3: Sample of misclassified examples on the test set for our two best models. GT corresponds to the ground truth labels, MC diff to Middle-confidence diff, and Syn to Synonym substitution method.

consider the following synthetic datasets: $data\_1$ is obtained by applying synonym substitution to the original dataset. $data\_2$ refers to the middle-confidence diff set. $data\_3$ is achieved by applying synonym substitution to $data\_2$. In this way, Combination 1 comprises the original data, $data\_1$ and $data\_3$. Combination 2 is the union of the original data, $data\_1$ and $data\_2$. Finally, Combination 3 consists of the original data, $data\_1$, $data\_2$, and $data\_3$.

We run each experiment five times and calculate the average and standard deviation for every metric. In Table 2, we can observe the most effective approach to combine augmented datasets is through an ensemble of both models. However, it is the most expensive option.

5.2 Error Analysis

According to our results, we conduct an error analysis on our best-performing models, which are those trained on middle-confidence diff and synonym substitution datasets.

Table 3 presents some of the most common error patterns. To maintain data privacy, we paraphrased the original examples in Spanish and translated them into English. In the first example, it was misclassified for both models because it contains some offensive words. However, it is not a harmful message. The third example was misclassified for the same reason, although the synonym substitution model got the correct answer. The second and fourth examples are considered offensive even if they do not contain bad words. That is why at least one of the models was wrong.

5.3 Loss function

Training a diffusion model for the text generation task presents different challenges. For instance, it performs poorly when trained on a small dataset because it has millions of parameters. To address this limitation, we design a framework (detailed in section 3.1) to train our diffusion model effectively. Another limitation we observed is that the model requires enormous training steps to converge. We can notice this behavior in Figure 2, where we can confirm that our model converges successfully.

6 Conclusion and Future work

This work introduces a novel data augmentation technique employing a Diffusion Language Model. We systematically compare our proposed method against conventional data augmentation techniques through a series of experiments through a series of experiments. The outcomes of these experi-

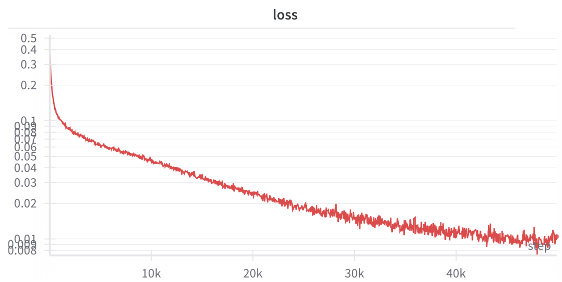

Figure 2: Visualization of the loss function during diffusion model training for the first 50000 steps. Data is displayed on a logarithmic scale.

ments reveal a modest yet discernible enhancement achieved by applying our diffuser data augmentation technique, thereby highlighting the potential for further exploration into related strategies.

We envision our study as a catalyst for delving deeper into DLM's advantages in generating synthetic data. We aim to inspire further investigations into leveraging DLM for similar purposes. Moreover, it's worth noting that there is also a gap in exploring data augmentation techniques for hate speech detection in non-English languages. This opens the opportunity for future research, offering opportunities for innovation and advancement within the field.

In future work, we plan to analyze the potential biases of the MEX-A3T dataset and how the models trained on this corpora could acquire them. We expect to find sexism or gender bias and then conduct an analysis similar to that of (Sap et al., 2019).

Furthermore, we want to employ various metrics to comprehensively assess the diversity of the synthetic data generated by the diffuser. This includes leveraging established metrics like Distinct-N (Li et al., 2015) to quantify the number of unique N-grams and Self-BLUE (Zhu et al., 2018) to measure the intrinsic similarity of the synthetic data. In addition to these quantitative measures, we will also conduct a visual inspection to qualitatively evaluate the data's diversity and richness.

A preliminary analysis has already yielded promising results. It suggests that the diffuser can generate synthetic data significantly different from the original data, indicating a high degree of diversity. We plan to incorporate a more detailed quantitative and qualitative diversity evaluation in our future work.

Limitations and Ethical Concerns

Our work presents the following limitations:

- The dataset was manually labeled, which implies that assignation depends on some factors. The notion of aggressiveness could vary according to gender, education, place of birth, cultural factors, etc. The diversity of annotator backgrounds could introduce a broader range of perspectives and potentially enrich the dataset. However, it is important to consider these biases when analyzing the data.

  Data augmentation techniques are susceptible to propagating biases in a dataset. We note that our method suffers from a particular type of bias. The aggressive class of the data set is closely related to the use of bad words. Our technique propagates this bias by generating text conditional on these words. We plan to reduce this bias by increasing the number of tweets that do not contain these offensive words.

- Our dataset contains 10,475 Spanish tweets. This is a small number of tweets to train efficiently a diffusion model. We address this limitation by pairing tweets to create a more extensive dataset.

Regarding potential ethical concerns, we recognize the intricate nature of analyzing content from social media platforms. Working with such data brings forth concerns regarding privacy and moral conduct. It is imperative to underscore that our research solely relied on existing publicly accessible datasets, and we refrained from direct interaction with users on social media platforms. The dataset used in this study is public and was taken from the MEX-AT3 official site. We meticulously adhered to the terms of service and user agreements governing these datasets. Additionally, it's essential to highlight that measures were taken to anonymize the datasets, safeguarding individual privacy. However, to maintain the confidentiality of our analysis, we paraphrased the examples displayed and translated them into English. Although individuals may share posts publicly, they may not anticipate the widespread dissemination of their content.

Acknowledgements

Antonio D. Reyes-Ramirez, Fernando Sanchez-Vega, and A. Pastor Lopez-Monroy thank CONAHCYT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies and CIMAT Bajio Supercomputing Laboratory (#300832). Antonio D. Reyes-Ramirez (CVU 1227043) thanks CONAHCYT for the support through the master's degree scholarship at CIMAT.

Mario Ezra Aragón, thanks for the support obtained from: I) the financial support supplied by the Consellería de Cultura, Educación, Formación Profesional e Universidades (accreditation 2019-2022 ED431G-2019/04, ED431C 2022/19) and the European Regional Development Fund, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System, and II) the financial support supplied by project PID2022-137061OB-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by the European Regional Development Fund). III) project 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).

Sanchez-Vega acknowledges CONAHCYT's support through the program "Investigadoras e Investigadores por México" (Project ID.11989, No.1311).

References

Sindhu Abro, Sarang Shaikh, Zahid Hussain Khand, Ali Zafar, Sajid Khan, and Ghulam Mujtaba. 2020. Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, 11(8).

Mario Ezra Aragón, Horacio Jesús Jarquín-Vásquez, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villasenor Pineda, Helena Gómez-Adorno, Juan Pablo Posadas-Durán, and Gemma Bel-Enguix. 2020. Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish. In *IberLEF@ SEPLN*, pages 222–235.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Salvatore Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Garima Koushik, K Rajeswari, and Suresh Kannan Muthusamy. 2019. Automated hate speech detection on twitter. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–4. IEEE.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Kosisochukwu Judith Madukwe and Xiaoying Gao. 2019. The thin line between hate and profanity. In *AI 2019: Advances in Artificial Intelligence: 32nd Australasian Joint Conference, Adelaide, SA, Australia, December 2–5, 2019, Proceedings 32*, pages 344–356. Springer.

Rijul Magu, Kshitij Joshi, and Jiebo Luo. 2017. Detecting the hate code on social media. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 608–611.

Ricardo Martins, Marco Gomes, Jose Joao Almeida, Paulo Novais, and Pedro Henriques. 2018. Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 61–66. IEEE.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII*, pages 928–940, Cham. Springer International Publishing.

Alex Nikolov and Victor Radivchev. 2019. Nikolov-radivchev at SemEval-2019 task 6: Offensive tweet classification with BERT and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Bhavesh Pariyani, Krish Shah, Meet Shah, Tarjni Vyas, and Sheshang Degadwala. 2021. Hate speech detection in twitter using natural language processing. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pages 1146–1152. IEEE.

Juan Manuel Pérez, Damián A Furman, Laura Alonso Alemany, and Franco Luque. 2021. Robertuito: a pre-trained language model for social media text in spanish. *arXiv preprint arXiv:2111.09453*.

Neha Rai, Pooja Meena, and Chetan Agrawal. 2020. Improving the hate speech analysis through dimensionality reduction approach. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 321–325. IEEE.

David Robinson, Ziqi Zhang, and Jonathan Tepper. 2018. Hate speech detection on twitter: Feature engineering vs feature selection. In *The Semantic Web: ESWC 2018 Satellite Events: ESWC 2018 Satellite Events, Heraklion, Crete, Greece, June 3-7, 2018, Revised Selected Papers 15*, pages 46–49. Springer.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Bertie Vidgen and Taha Yasseri. 2020. Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6:13825–13835.

Jason Wei and Kai Zou. 2019a. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Jason Wei and Kai Zou. 2019b. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web*, pages 745–760, Cham. Springer International Publishing.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.