Impact of Data Augmentation on Hate Speech Detection in Roman Urdu

Fariha Maqbool[1,*], Blerina Spahiu[1] and Andrea Maurino[1]

[1]*Dipartimento di informatica, sistemistica e comunicazione*
University of Milano-Bicocca
Viale Sarca 336, 20126 Milan, Italy

Abstract

The prevalence of hate speech leads to an increase in hate crimes, online violence, and serious harm to social safety, physical security, and cyberspace. To address this issue, several studies have been conducted on hate speech detection in European languages, whereas little attention has been paid to low-resource South Asian languages, making social media vulnerable for millions of users. Due to the scarcity of the datasets and the samples available, there is a need to apply some strategies to increase the data samples. In this paper, we improved the performance of the already fine-tuned m-Bert model by applying data augmentation techniques to one of the datasets on hate speech on tweets in Roman Urdu language. F1-score and accuracy matrix have been used to compare the results. We also experiment to determine the optimal percentage of augmented data to be included and the percentage of words augmented in each instance of data. The new RUHSOLD++ Dataset containing the augmented data has also been published publicly. The improvement in hate speech detection of the model proved that the performance of the models can be improved by applying data augmentation techniques to the dataset with a limited number of instances.

1. Introduction

The exponential growth of social media platforms like Facebook[1], x (formally Twitter)[2], and YouTube[3] has provided a global stage for individuals from diverse cultures and social backgrounds to communicate and share their opinions on a myriad of topics. While these platforms uphold the principle of freedom of speech, some users also negatively exploit this freedom and abuse other users on the basis of gender, religion or race. This surge in harmful content has underscored the need for increased research in natural language processing (NLP) to effectively detect instances of hate speech. The consequences faced by victims of targeted hate speech are not limited to physical harm; they also experience a profound sense of dread and rejection within their communities. Recognizing the urgency to create online spaces free of racism and hate speech, researchers emphasize the importance of early detection mechanisms to mitigate the pervasive harm caused by such content [1]. This challenge extends beyond the English language, as millions of users worldwide employ diverse languages as vehicles for spreading

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy
*Corresponding author.
f.maqbool@campus.unimib.it (F. Maqbool); blerina.spahiu@unimib.it (B. Spahiu); andrea.maurino@unimib.it (A. Maurino)

[1]https://www.facebook.com/
[2]https://x.com/
[3]https://www.youtube.com/

hate. Despite extensive research in English, there exists a noticeable dearth of datasets and studies on languages like Urdu. Urdu, spoken by over 170 million people globally, faces unique challenges in written communication, with an alphabet that cannot be easily mapped onto an English keyboard. The Urdu language comprises 40 characters, yet English keyboards, accommodating only 26 letters, face limitations in fully supporting Urdu script. Attempting to map the Urdu alphabet directly onto an English keyboard proves impractical due to these constraints. Consequently, the predominant approach among Urdu speakers involves the use of Roman Urdu particularly on social media platforms, Roman Urdu is a transliteration version of Urdu using English letters. The use of Roman Urdu has sharply expanded as a result of social media's rising adoption[2, 3]. Users of these tools regularly utilize these platforms to share their opinions about a variety of products, services, politics, and other items. However, despite its widespread use, Roman Urdu (RU) encounters challenges such as a lack of linguistic resources, annotated datasets, and dedicated language models [4]. To address these limitations and enhance model performance, researchers have explored data augmentation techniques[3]. In a notable application, simple data augmentation techniques were employed on a low-resourced language dataset with a limited number of samples. The results demonstrated a significant improvement in model performance, underscoring the potential of augmentation in mitigating the challenges posed by scarce linguistic resources. Additionally, experiments aimed to identify the optimal percentage of augmented data to be integrated with the original dataset, aiming to boost model performance while minimizing training time. This multifaceted approach contributes to ongoing efforts to combat hate speech across various languages, fostering inclusivity and positive discourse in online spaces. In this paper we make the following contributions: (i) enrich RUHSOLD dataset [3] and create a new RUHSOLD++ dataset; (ii) provide a new custom function in Python to dynamically alter the spelling of selected words within a sentence; and (iii) provide an empirical analysis of the different data augmentation methods. The paper is structured as follows: Section 2 discusses approaches to detect hate speech in the Roman Urdu language. In Section 3 we provide the methodology to augment our initial dataset. Section 4 provides the analysis and findings by applying different methods for data augmentation while conclusions and future work end the paper in Section 5.

2. Related Work

The issue of abusive speech has been a longstanding focus within the research community. In earlier investigations, attempts to identify abusive users primarily relied on lexical and syntactic features extracted from their posts [5]. However, recent advancements in automated hate speech detection have witnessed substantial growth. The availability of large datasets has prompted a shift in academic research towards more data-intensive and sophisticated models, notably leveraging deep learning techniques [6] and graph embedding methods [7]. Notably, transformer-based language models like BERT[8] have gained immense popularity in various downstream tasks, proving to be particularly effective in surpassing traditional deep learning models such as CNN-GRU[9], LSTM[10], etc., for the detection of abusive language [11], [12]. This evolution highlights the dynamic nature of research in addressing the complexities of abusive speech detection. Detecting hate and abusive speech in low-resourced languages

presents a formidable challenge, as exemplified in the context of Roman Urdu. The scarcity of linguistic resources for Roman Urdu spurred the creation of the RUHSOLD dataset by H. Rizwan et al. [3]. Comprising 10,012 tweets, this dataset stands out for its dual approach, offering both coarse-grained and fine-grained labeling of hate speech instances. In their research, Rizwan and colleagues not only curated this valuable dataset but also proposed a deep learning-based architecture specifically tailored for hate speech detection in Roman Urdu. Addressing the broader landscape of multilingual abusive speech, M. Das et al. [13] conducted an in-depth investigation into the performance of multilingual models across eight distinct Indic languages. In a noteworthy application, they employed m-BERT[8] and MuRIL[14] models on the RUHSOLD dataset [3] to gauge their efficacy in detecting abusive speech. Through a series of meticulously designed experiments, encompassing various settings, Das and team explored the nuances of multilingual hate speech detection. Their findings underscored the effectiveness of model transfers, revealing that transferring knowledge from one language to another enhances the overall performance of the models. This body of research not only contributes to the evolving field of hate speech detection but also illuminates the specific challenges associated with low-resourced languages like Roman Urdu. By providing a robust dataset and proposing dedicated architectures, these studies lay essential groundwork for future endeavors aimed at combating hate speech across diverse linguistic landscapes. The insights gained from these investigations, especially regarding the transferability of models, offer valuable guidance for the development of more inclusive and effective hate speech detection systems in multilingual contexts. In a meticulous analysis, M. M. Khan. et al. [15] delved into the complexities of hate speech detection in Roman Urdu, manually examining over 90,000 tweets to curate a substantial corpus of 5,000 Roman Urdu tweets. Their significant contribution extended beyond dataset creation, as they systematically employed five supervised learning approaches, including a sophisticated deep learning technique, to rigorously evaluate and compare their effectiveness in hate speech detection. The results of their comprehensive study revealed that, across two levels of categorization, logistic regression outperformed all other techniques, opening up a viable path for robust of hate speech detection in Roman Urdu. Recognizing the challenges posed by the low resources of Roman Urdu, Azam et al. [16] undertook a proactive exploration of data augmentation strategies. Leveraging both easy data augmentation and transformer-based augmentation approaches, they aimed to enhance hate speech detection capabilities in Roman Urdu. The researchers conducted experiments using existing datasets in Roman Urdu and baseline models to meticulously assess the impact of augmentation techniques. Their findings unequivocally demonstrated that the performance of hate speech detection models could indeed be significantly improved by the strategic application of augmentation techniques to the dataset. This research not only contributes to the optimization of hate speech detection in low-resourced languages like Roman Urdu but also highlights the potential of augmentation strategies as valuable tools in mitigating the impact of resource constraints, providing valuable insights for the ongoing evolution of hate speech detection methodologies.

3. Methodology

3.1. Dataset

We employed the RUHSOLD dataset, a comprehensive collection of tweets in Roman Urdu created by H. Rizwan et al. [3]. The authors meticulously established a gold standard for two distinct sub-tasks. Our focus centered on the first sub-task, which involves binary labels categorizing content as either Hate-Offensive (labeled as 0) or Normal (labeled as 1), representing inoffensive language. The dataset comprises a total of 10,000 tweets, thoughtfully partitioned into training, testing, and validation sets in a ratio of 70%, 20%, and 10%, respectively.

3.2. Data Augmentation

In our quest to boost our dataset's size and improve the overall performance, we strategically employed Noise-based Data Augmentation techniques on training data. We dove into a detailed exploration, trying out different percentages for augmenting the dataset to strike the right balance. Moreover, we played around with the augmentation process by adjusting the percentage of words in each tweet that underwent these augmentation techniques. This nuanced approach was not just about expanding the dataset; it was about delicately adjusting the augmentation impact and finding the sweet spot between quantity and quality to strengthen our model's resilience. Through methodical experimentation, we aimed to uncover the most effective configurations that could genuinely enhance the overall performance of our model.

3.2.1. Random Swapping

Random swapping is an effective technique in the realm of noise-based data augmentation. This technique is based on randomly swapping the words or tokens from a tweet. For example: *"hum kisi se km nhi"* becomes *"km kisi se hum nhi"*.

This operation adds variances to the dataset without changing the overall sentiment or context of the text. In order to help the model generalize and function well on a variety of linguistic patterns, it is intended to be exposed to various word configurations.The percentage of word augmentation in random swapping directly controls the degree of variability injected into the dataset. Therefore, it is essential to find the optimal percentage of words which should be swapped during augmentation.

3.2.2. Random Deletion

Another noise-based data augmentation is Random Deletion. This strategy involves the deliberate and random removal of words or tokens from a given sentence, introducing an element of unpredictability and variability. We designate a specific percentage of words within the sentence for potential removal, aiming to strike a careful balance between introducing noise for robustness and preserving the coherence of the text. By implementing this intentional randomness, we infuse the dataset with a dynamic quality, fortifying the model's adaptability to diverse linguistic nuances. This method serves as a potent instrument, enriching our model's adaptability and efficacy across a wide range of textual inputs.

3.2.3. Spelling Augmentation

Spelling Augmentation adds a layer of complexity by altering the spelling of words within a sentence. This process entails replacing one or more characters in a word with randomly chosen alternatives and deliberately introducing a controlled amount of noise into the data. The purpose here is to diversify the linguistic patterns in the dataset, enhancing the model's ability to handle variations in spelling and promoting resilience against potential inconsistencies in user-generated content. This meticulous introduction of noise through character substitution serves as a strategic maneuver, refining our model's capacity to adapt to a wide array of spelling idiosyncrasies. For example: *"chal ja tujhy maaf kia"* becomes *"chal aa tujha maaf kia"*.

3.3. The new RUHSOLD++ Dataset

After employing the data augmentation techniques, we created a new RUHSOLD++ Dataset[4] which can be accessed publicly to promote future work. The dataset consists of 3 types of data with swap, delete and spelling augmentation applied. The dataset contains the augmented data in train and validation sets distributed uniformly and unaltered test data.

3.4. Model

The m-BERT model has garnered significant attention in the realm of abusive speech research. Its efficacy is attributed to being pre-trained on a comprehensive dataset, comprising the extensive content of Wikipedia[5], employing a masked language modeling (MLM)[17] objective across 104 languages. This pre-training involves 12 fully connected transformer encoder layers, incorporating a self-attention mechanism to efficiently process contextual information. It is worth noting that m-BERT, while powerful, has a token limit of 512, necessitating the use of a fine-tuned variant introduced by Das, M. et al. [7]. Das and colleagues enhanced the original m-BERT by incorporating a fully connected layer, aligning its output with the CLS (classification) token in the input. This added layer introduces a level of specificity, with the output reflecting the model's interpretation of the input sentence, often represented by the CLS token output. This nuanced modification allows the model to capture and interpret complex contextual nuances within the given token limit, contributing to its efficacy in understanding and classifying abusive speech patterns.

4. Experimentation and Results

In this section, we describe the experimentation conducted on the RUHSOLD [3] dataset using fined tuned m-Bert model, a refinement proposed by M. Das et al. [7]. In our implementation, we use the PyTorch library[6] in Python, configuring each model to run for 10 epochs with an Adam optimizer and a batch size of 16.

[4] https://github.com/fariha231/impact-of-augmentation-ruhsoldplusplus
[5] https://www.wikipedia.org/
[6] https://pytorch.org/

The experimentation extended to the exploration of data augmentation techniques, aiming to strike an optimal balance between computational efficiency and accuracy. The primary focus was on determining the most suitable percentage of the dataset for applying augmentation. To achieve this, we selected a portion of the original training dataset that underwent Random Swap Augmentation. Leveraging the NLPAug[7] library in Python, a random percentage of 30% for words was chosen, indicating that 30% of the total words in a tweet would undergo swapping with each other. Following the generation of augmented data, it was seamlessly integrated with the original dataset and subsequently shuffled to mitigate overfitting concerns. The outcomes of this experiment are succinctly presented in Table 1, showcasing the impact of Random Swap at varying percentages of the dataset. This experimentation highlights the strategic choices made in augmenting the data to achieve an optimal trade-off between efficiency and accuracy.

Table 1
Random Swap with Varying Overall Augmented Data

% of original data augmented	Precision	Recall	mF1-score	Accuracy
0	0.873	0.876	0.863	0.875
20	0.910	0.907	0.902	0.903
30	0.913	0.917	0.909	0.909
50	**0.914**	**0.925**	**0.913**	**0.913**
60	0.927	0.891	0.904	0.905
80	0.922	0.893	0.902	0.903
100	0.917	0.883	0.898	0.898

We employed the Macro F1 score (mF1-score) as a performance metric, along with other evaluation measures. The Macro F1 score allows to assess the performance of each class individually while giving equal weight to all classes. Examining the results, we observe a consistent improvement in both accuracy and mF1-score as the model is trained on augmented data. However, a notable finding emerges: the highest accuracy is attained when augmentation is applied to 50% of the data. Beyond this threshold, further increasing the size of augmented data leads to diminishing returns, resulting in a decline in accuracy and mF1-score. This suggests that the model tends to overfit the training data when subjected to an excessive amount of augmented information. While the accuracy of the validation data may show promising signs, the model's performance on unseen data, specifically the test data, begins to decrease.

With the optimal augmented data percentage identified, our exploration extends to varying the percentage of words swapped in each iteration. The results, as depicted in Table 2, indicate that the overall accuracy and mF1-score exhibit minimal fluctuations with changes in the word augmentation percentage. However, precision and recall values do showcase variations corresponding to alterations in the word augmentation percentage. This nuanced observation

[7]https://pypi.org/project/nlpaug/0.0.5/

underscores the importance of fine-tuning not only the quantity of augmented data but also the specific aspects of augmentation.

Table 2
Random Swap with Varying Word Swap Rate

Words swapped per instance (%)	Precision	Recall	mF1-score	Accuracy
20	0.909	0.916	0.906	0.907
30	0.914	0.925	0.913	0.913
40	0.929	0.907	0.913	0.913
50	0.926	0.910	0.913	0.913

Subsequently, we implemented the Delete Data Augmentation on 50% of the original training dataset. Leveraging the NLPAug library in Python, we conducted experiments to generate new data by selectively removing certain words in each row. This augmented dataset was seamlessly integrated with the original data, effectively amplifying the training set by 50%. Our exploration further extended to varying the percentage of words designated for deletion in each iteration. The outcomes of this experiment are presented in Table 3. The results demonstrate that this approach not only diversifies the training data but also involves fine-tuning the augmentation to achieve improvements in model performance.

Table 3
Random Delete with Varying Word Deletion Rate

Words deleted per instance (%)	Precision	Recall	mF1-score	Accuracy
20	0.92	0.88	0.89	0.90
30	0.92	0.88	0.895	0.895
40	0.916	0.894	0.899	0.90

In implementing spelling augmentation, we crafted a custom function in Python to dynamically alter the spelling of selected words within each sentence. This function provides the flexibility to adjust the percentage of words in each row subject to augmentation. The outcomes of this experiment are presented in Table 4. Notably, the results reveal an improvement in accuracy as we incrementally raise the percentage of augmented words in each row. However, a cautious approach was adopted, refraining from further increasing the percentage to prevent potential distortion of the sentence's meaning. Beyond a certain threshold, excessive alterations could compromise the contextual integrity of the sentence, potentially undermining the model's overall performance.

After conducting a comprehensive array of experiments, our findings show that the most

Table 4
Spelling Augmentation with Varying Word Augmentation Rate

Words augmented per instance (%)	Precision	Recall	mF1-score	Accuracy
30	0.921	0.885	0.898	0.898
50	0.928	0.898	0.908	0.908

favorable accuracy was achieved through swap data augmentation. The optimal model, exhibiting the highest accuracy, emerged from training with an additional 50% of data, where 30% of words in each row were subject to swapping. This configuration demonstrated the finest balance between data enrichment and model performance enhancement.

To provide a visual representation of the model's performance, Figure 1 presents the confusion matrix for this optimal model, showcasing the details of how well the model navigated and classified instances with the applied swap data augmentation. This synthesis of experimentation outcomes reinforces not only the efficacy of swap data augmentation but also the significance of precise configurations in achieving the model's peak performance.

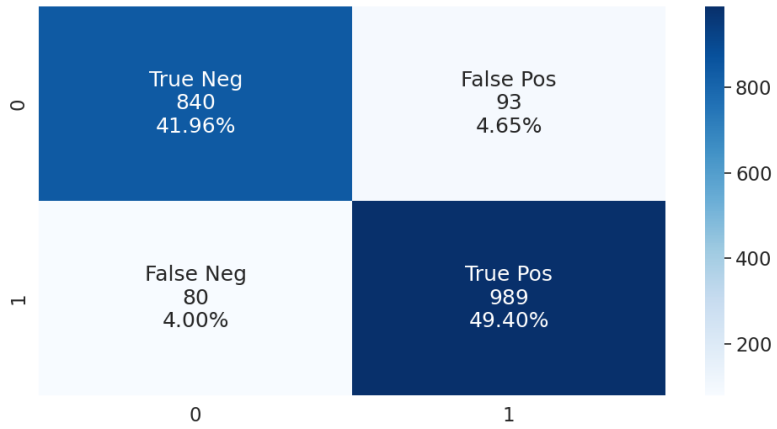

Figure 1: Confusion Matrix for best model with swap data augmentation

5. Conclusion and Future Work

Recognizing the pressing need to combat hate speech within the constraints of limited resources, our paper lies in the strategic application of data augmentation techniques to linguistic datasets on social media. We assert that applying data augmentation techniques to the dataset helps to increase the dataset size and improves the overall model performance. We experimented with determining the ideal percentage of augmented data to seamlessly integrate with the original dataset. This exploration aimed not only to enhance model training efficiency but also

to circumvent the pitfalls of model overfitting. Our experimentation involved the application of three distinct data augmentation techniques: random swap, random deletion, and spelling augmentation. Notably, the results underscore the prowess of swap data augmentation, exhibiting the highest accuracy at 91.3%. This achievement was realized with a 30% word augmentation rate and a 50% augmented data incorporation. We also published the new RUHSOLD++ dataset containing the augmented data. For future work we envision the exploration of additional augmentation techniques, setting the stage for a comprehensive comparison of model performances. This will improve and add more tools to the ongoing fight against hate speech on social media especially for under-resourced languages such as Roman Urdu.

References

[1] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Comput. Surv. 51 (2018) 85:1–85:30. URL: https://doi.org/10.1145/3232676. doi:10.1145/3232676.

[2] M. M. Khan, K. Shahzad, M. K. Malik, Hate speech detection in roman urdu, ACM Trans. Asian Low Resour. Lang. Inf. Process. 20 (2021) 9:1–9:19. URL: https://doi.org/10.1145/3414524. doi:10.1145/3414524.

[3] H. Rizwan, M. H. Shakeel, A. Karim, Hate-speech and offensive language detection in roman urdu, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 2512–2522. URL: https://doi.org/10.18653/v1/2020.emnlp-main.197. doi:10.18653/V1/2020.EMNLP-MAIN.197.

[4] A. Dewani, M. Memon, S. Bhatti, Development of computational linguistic resources for automated detection of textual cyberbullying threats in roman urdu language, 3C TIC: Cuadernos de desarrollo aplicados a las TIC 10 (2021) 101–121. doi:10.17993/3ctic.2021.102.101-121.

[5] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting offensive language in social media to protect adolescent online safety, in: 2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Confernece on Social Computing, SocialCom 2012, Amsterdam, Netherlands, September 3-5, 2012, IEEE Computer Society, 2012, pp. 71–80. URL: https://doi.org/10.1109/SocialCom-PASSAT.2012.55. doi:10.1109/SOCIALCOM-PASSAT.2012.55.

[6] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: R. Barrett, R. Cummings, E. Agichtein, E. Gabrilovich (Eds.), Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017, ACM, 2017, pp. 759–760. URL: https://doi.org/10.1145/3041021.3054223. doi:10.1145/3041021.3054223.

[7] M. Das, P. Saha, R. Dutt, P. Goyal, A. Mukherjee, B. Mathew, You too brutus! trapping hateful users in social media: Challenges, solutions & insights, in: O. Conlan, E. Herder (Eds.), HT '21: 32nd ACM Conference on Hypertext and Social Media, Virtual Event, Ireland, 30 August 2021 - 2 September 2021, ACM, 2021, pp. 79–89. URL: https://doi.org/10.1145/3465336.3475106. doi:10.1145/3465336.3475106.

[8] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional

transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1, Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/V1/N19-1423.

[9] Z. Zhang, D. Robinson, J. A. Tepper, Detecting hate speech on twitter using a convolution-gru based deep neural network, in: A. Gangemi, R. Navigli, M. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, M. Alam (Eds.), The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings, volume 10843 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 745–760. URL: https://doi.org/10.1007/978-3-319-93417-4_48. doi:10.1007/978-3-319-93417-4_48.

[10] G. V. Houdt, C. Mosquera, G. Nápoles, A review on the long short-term memory model, Artif. Intell. Rev. 53 (2020) 5929–5955. URL: https://doi.org/10.1007/s10462-020-09838-1. doi:10.1007/S10462-020-09838-1.

[11] H. S. Alatawi, A. Alhothali, K. Moria, Detection of hate speech using BERT and hate speech word embedding with deep model, CoRR abs/2111.01515 (2021). URL: https://arxiv.org/abs/2111.01515. arXiv:2111.01515.

[12] M. Bilal, A. Khan, S. Jan, S. Musa, S. Ali, Roman urdu hate speech detection using transformer-based model for cyber security applications, Sensors 23 (2023) 3909. URL: https://doi.org/10.3390/s23083909. doi:10.3390/S23083909.

[13] M. Das, S. Banerjee, A. Mukherjee, Data bootstrapping approaches to improve low resource abusive language detection for indic languages, in: A. Bellogín, L. Boratto, F. Cena (Eds.), HT '22: 33rd ACM Conference on Hypertext and Social Media, Barcelona, Spain, 28 June 2022- 1 July 2022, ACM, 2022, pp. 32–42. URL: https://doi.org/10.1145/3511095.3531277. doi:10.1145/3511095.3531277.

[14] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, P. P. Talukdar, Muril: Multilingual representations for indian languages, CoRR abs/2103.10730 (2021). URL: https://arxiv.org/abs/2103.10730. arXiv:2103.10730.

[15] M. M. Khan, K. Shahzad, M. K. Malik, Hate speech detection in roman urdu, ACM Trans. Asian Low Resour. Lang. Inf. Process. 20 (2021) 9:1–9:19. URL: https://doi.org/10.1145/3414524. doi:10.1145/3414524.

[16] U. Azam, H. Rizwan, A. Karim, Exploring data augmentation strategies for hate speech detection in roman urdu, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022, European Language Resources Association, 2022, pp. 4523–4531. URL: https://aclanthology.org/2022.lrec-1.481.

[17] J. Salazar, D. Liang, T. Q. Nguyen, K. Kirchhoff, Masked language model scoring, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 2699–2712. URL: https://doi.org/10.18653/v1/2020.acl-main.240. doi:10.18653/V1/2020.ACL-MAIN.240.