Classification and Analysis of Indonesian Hate Speech on Twitter Using XGBoost

Timoteus Fabian Halim  
*School of Computer Science*  
*Bina Nusantara University*Jakarta, Indonesia  
timoteus.halim@binus.ac.id

Belinda Mettadevi  
*School of Computer Science*  
*Bina Nusantara University*Jakarta, Indonesia  
belinda.mettadevi@binus.ac.id

Felix Indra Kurniadi  
*Computer Science Department*  
*School of Computer Science*  
*Bina Nusantara University*Jakarta, Indonesia 11480  
Benny Strata Wijaya  
*School of Computer Science*  
*Bina Nusantara University*Jakarta, Indonesia  
benny.wijaya002@binus.ac.id

*Abstract*— In social media platforms such as Twitter, particular negative content classified as hate speech is often insufficiently regulated and moderated. The use of Hate Speech Detection systems to handle such content is an often-reached conclusion to this problem. However, most detection systems focus on English-specific hate speech whereas each language typically has a unique and specific set of speech that can be explicit or implicitly hateful. This study aims to analyze and classify Indonesian Twitter text using a machine learning algorithm known as Extreme Gradient Boosting (XGBoost). The algorithm has the capability to handle large datasets such as the dataset used in this study with 13,169 tweets. Although the dataset is multi-classed—with each tweet labelled for different types of hate speech—this study focuses on the detection of hate speech in general. The pre-processing techniques applied to the dataset include lowercasing, removal of non-alphanumeric characters, normalization of slang words, and removal of stop words. XGBoost is then trained to classify the tweets into hate or non-hate speech. The trained model is then evaluated using the accuracy, precision, recall, and F1-score as the assessment metric of the model’s performance. The results achieved an accuracy score of 0.82 on the testing set, indicating that the model makes the correct prediction at most instances.

Keywords—hate speech, machine learning, Indonesian, Twitter, XGBoost

# Introduction

In the world wide web, content is often able to be posted and published for the global audience instantaneously. Unfortunately, the moderation of online content is often slow and limited, causing the nature of such content to be difficult to monitor effectively or at all [1]. Since the current environment of the internet is effectively spread across people from all walks of life, all kinds of content are exposed to both normal and more vulnerable people including young people whose cognitive development is highly unstable and easy to influence [2].

Although hate speech has a clear-cut definition, speech itself is difficult to define into hate or non-hate. Hate speech is an umbrella term often used to define a manner of speech which is detrimental or meant to degrade or abuse certain peoples or objects by targeting certain undesirable traits (discrimination), using slurs and offensive language (hate), and various other negatively impacting content and speech [3]. The prevalence of freedom of speech allows for the uncontrollable spread of hate speech, yet the speech itself is difficult to constrain without restricting freedom of speech [1]. In order to moderate and monitor the online content for hate speech quicker and more effectively, a system to detect such speech is crucial [4]. Even if not to completely overtake the task of monitoring, a detection and recognition system will be able to assist human moderation by automating the filtering of massive amounts of content at speeds nearing that of its generation.

The field of machine learning has produced many methods in an attempt to detect hate speech, including various classifiers and even deep-learning related methods [4]. Natural Language Processing (NLP) is a field of computer science that studies the analysis and processing of text and other speech mediums to detect, comprehend, and extrapolate meaning [5]. Hate speech detection falls under this category and related methods in this field have been implemented for this purpose. A more specific sub-field of study is known as Sentiment Analysis which is defined by Oxford dictionary [6] as “computationally identifying and categorizing opinions… especially in order to determine whether the author’s attitude… is positive, negative, or neutral.” whose methods may be implemented partially in order to detect hate speech that is typically negative in sentiment.

However, most studies [7] focuses on implementations of analysis methods for text-speech using the most commonly acknowledged language on social media platforms, i.e. English. Hate speech as a form of communication online is not limited to the English language and each language has a unique national- or even region-specific set of speech that is negative or offensive in nature and intent. Local researchers and developers has done sparse research on each of their own local languages for the detection of hate speech. Several studies [3] attempted to do the same in the Indonesian language with varying results. Despite so, the existence of machine learning models dedicated to the detection of Indonesian hate speech is difficult to find within the scope of this study.

This study aims to analyze a dataset of Indonesian twitter text in order to train a machine learning model capable of classifying Indonesian text as hate or non-hate speech. In order to detect the hate speech patterns, the machine learning algorithm used in this study is Extreme Gradient Boosting (XGBoost) [8] due to its robust capabilities in processing large datasets and flexibility in use. The aforementioned Indonesian Twitter dataset is taken from a GitHub repository that in itself is created based on a study from 2019 by Ibrohim & Budi [27]. The dataset is accompanied by related dictionaries for specific words labeled as hate and fixes for all typos, misspellings, and slang or slurs used in the dataset.

Between 2019 to 2021, an approximate number of 3640 content on the internet has been taken down by the Indonesian Ministry of Communication and Information (KOMINFO) [31] not including other content that has been blocked from access within Indonesia. The content that was taken down is mostly online content, with some being certain mediums of communication often used to propagate or express slurs targeting specific races, religions, and other discriminating features or hate speech in general. While some words are explicitly used as hateful and slurs, many Indonesian hate speech uses implicit negative connotations. And to further complicate the task of detecting hate speech, offensive speech is also used with non-harmful intentions as a type of dark humor—typically directed to oneself or one’s own discriminating feature. Due to this complex nature, a specific and particular model is necessary to be trained in detecting purposefully offensive content.

# Related Works

A study from Durban University of Technology in South Africa investigates hate speeches in social media, namely Twitter [Using Transformer Method]. By utilizing a dataset containing 24783 labeled tweets, they used the DistilBERT method and compared the result to BERT, XLNet, RoBERTa, and LSTM. DistilBERT is shown to perform the best on the dataset, with a result of 92% on Accuracy and 75% on Precision. XLnet scored the same on Precision with 75% result and 91% Accuracy. While LSTM scored the lowest at 66% on both Accuracy and Precision. [30]

Another study created two versions of the same software to detect hate speech in Twitter for CLEF 2021. This research created two versions in order to satisfy the need for both English and Spanish detection. The study used LR, SVM, RF, XGB, Statistical XGB, and Dictionary-based n-gram XGB. The result of both software showed that the models used are more suitable for the English corpus with the Spanish corpus achieving lesser accuracy. [9]

Researchers at Keio University conducted a similar research on Twitter as well, in order to combat the rise of cyber conflicts in social media. The research uses a dataset consisting of 2010 tweets. By using RF, SVM, and J48graft, they evaluated the result to show that their study sees an increased 87.4% accuracy when checking whether a tweet is offensive and 78.4% accuracy when checking whether the tweet is hateful. [7]

People from the Norwegian University of Science and Technology conducted a research they claimed to be effective by using Recurrent Neural Networks (RNN) model. The LSTM model is used to process the Waseem dataset that contains 1943 tweets labeled as racism, 3166 tweets labeled as sexism, and 10889 tweets labeled as neutral. By using the Area Under Curve (AUC), the study achieved a score of 80% on NS, while the results did not achieve the same success on the other 5 classes. [11]

# Research Methodology

The dataset used in this study is taken from the GitHub repository [3]. It provides a multi-label collection of text from Twitter (tweets) to detect hate speech and abusive language in the Indonesian language. Said labels are binary which indicates whether a Tweet is considered a hate speech or not. In addition the dataset also provides binary labels on whether the hate speech is targeted to an individual, group, religion, race, physical appearance and gender. While the dataset has binary labels that specify the degrees and categories of hate speech, the research only focuses on the hate speech in general. Table I is an example of tweets that are labeled as HS (hate speech) and non-hate speech.

1. Random Example of Classified Tweets

| No | Value | Hate Speech? |
| --- | --- | --- |
| 1 | USER USER USER USER USER USER Lawan bicara gw gak intelek kyk loe, yg otak gak punya. Ttg kencing onta, gw mengakui hadis nabi dan itu sahih. Kafir kyk loe pasti menolak, makanya loe ahlunnar. | Yes |
| 2 | Belakangan ini kok fikiran ampas banget ya | No |

A picture containing text, screenshot, receipt, font

Description automatically generatedFig. 1. describes the flow of the experiment run in this study, from data preprocessing until testing through model prediction. The preprocessing techniques used are the lowercasing of the text to ensure uniformity, removal of unnecessary characters such as links and Twitter-specific functions, normalization of slang words, and removal of Indonesian stop words.

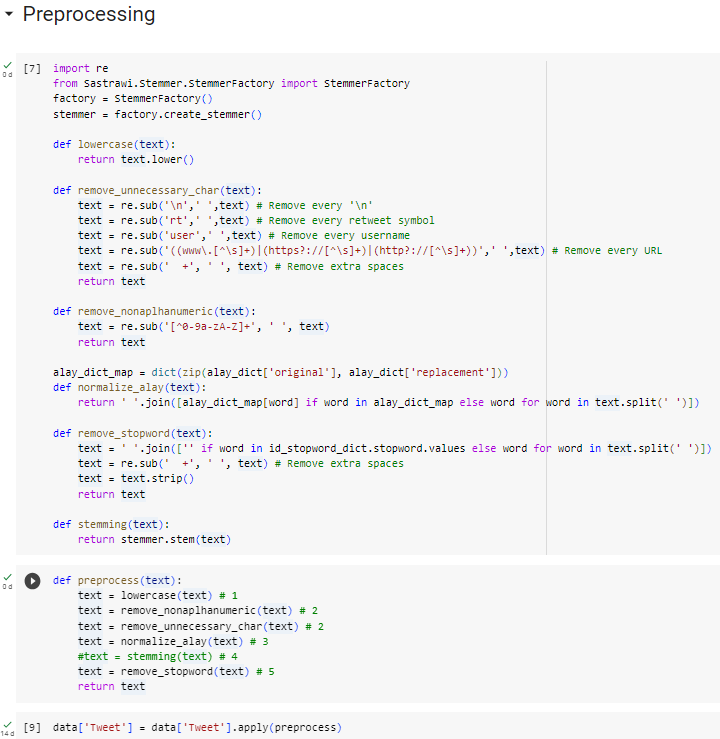
1. Workflow chart of the experiment.
2. Code to import and use the dataset and dictionaries.

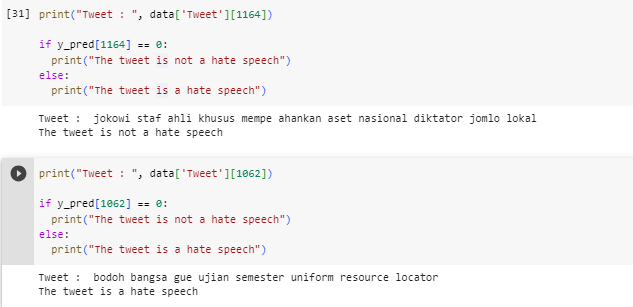
For the normalization of slang words, a dictionary consisting of slang words and their formal definitions are used [27], whereas the Indonesian stop words uses a different dictionary [14]. These dictionaries are also imported and read by the experiment code with the Indonesian Twitter dataset as shown in Fig. 2.

The combined codes necessary to complete the described preprocessing techniques are merged into one function and applied to the dataset as seen in Fig. 3.

The following step is the feature extraction. This study uses the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer. The vectorizer works by combining Term Frequency and Inverse Document Frequency. It calculates a numerical value for each term in each document to represent the importance of that term in that particular document relative to others in the dataset. When applied to the processed dataset, it is able to learn the vocabulary used in the dataset and transform the Twitter text into a feature matrix. This feature matrices represents each document as a set of numerical features.

The dataset is then split with a ratio of 8:2 for training and testing. The machine learning model is trained using the training set data and tested using a separate testing set data to evaluate its performance when predicting using previously unseen data. The algorithm XGBoost is used in this study because of its gradient boosting capabilities, ability to handle imbalanced classes, feature importance analysis, flexibility, and compatibility with the scikit-learn library in Python. Patterns and relationships between the input features (*X*) and target variable (*y*) are learned and analyzed according to the algorithm during the training.

The trained model is then tested using the testing dataset to assess its performance and evaluated using evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics are able to provide a quantification of the model’s performance by scoring the predictions made based on the testing dataset.

1. Preprocessing Techniques implemented in code
2. Classification and Evaluation process in code
3. Model prediction for testing in code with results

Accuracy is the percentage of correctly predicted labels in the test dataset. Precision is the ability of the model to correctly identify the positive class among the predicted positive instances. Recall is the ability of the model to identify the true positive instances from the actual positive instance. Lastly, the F1-score is the mean of precision and recall which provides a balanced measure of the model’s performance. This whole process is called Classification and Evaluation, as seen in Fig. 4.

Fig. 5. shows the prediction made by the model where two random indexes from the Tweet dataset are taken. The results are also visible in Fig. 5.

# Results and Discussions

Table II and III shows two tweets that are both raw and preprocessed. Starting from the conversion of the text to lowercase until removal of stopwords. As the name suggests, the text is converted to lowercase to avoid inconsistencies caused by capitalization. The second text which is full of capital letters is decapitalized in the clean text.

Removal of unnecessary characters such as retweets “RT”, usernames “USER”, URLS, and extra spaces is done to eliminate irrelevant information from the text.

1. Raw Hate Speech Tweets

| No | Raw Text |
| --- | --- |
| 1 | USER USER USER USER USER USER Lawan bicara gw gak intelek kyk loe, yg otak gak punya. Ttg kencing onta, gw mengakui hadis nabi dan itu sahih. Kafir kyk loe pasti menolak, makanya loe ahlunnar. |
| 2 | BODO AMAT BNGS GUE LAGI UAS URL |

1. Cleaned Hate Speech Tweets

| No | Cleaned Text |
| --- | --- |
| 1 | jokowi staf ahli khusus mempe ahankan aset nasional diktator jomlo lokal |
| 2 | bodoh bangsa gue ujian semester uniform resource locator |

Slang words that are found in the text are replaced by their normalized forms which makes the text more consistent and easier to analyze. For example the word “BODO” from the second text is changed into “bodoh” (stupid). Even though “bodo” could have a meaning of not caring, the word is converted into its normalized form based on the “new\_kamusalay” dataset [3].

Stopwords such as pronouns and prepositions like “Apa” and “pak” are removed to reduce noise and allow the model to focus on the more meaningful words in the text, also based on “stopwordbahasa” dataset [29]. The researchers are still uncertain why some words are missing a letter like “mempertahankan” and “jomblo”. In the clean text those words become “mempe tahankan” and “jomlo”. These steps help to clean the text data and standardize its format and remove unnecessary information which improves the quality of the data thus preparing it for further analysis.

The classification report on Table IV summarizes the performance of the XGBoost model for hate speech classification. For the non-hate speech, the model performs well in correctly identifying instances of non-hate speech as proven by the high precision and recall values. The second class which is hate speech the values provided suggests that the model is somewhat less accurate in identifying hate speech instances compared to non-hate speech instances, as the precision and recall are slightly lower. Both the macro and weighted average F1-score indicates that the model is quite effective in identifying hate speech. Table V shows the overall performance of the XGBoost model. These results suggest that the XGBoost model performed reasonably well in classifying hate speech instances because the accuracy, precision, recall and F1-score are all above 0.80. It means that the model correctly classifies approximately 82% of the instances and achieves a good balance between correctly classifying hate speech while avoiding misclassification.

1. Classification Report from the Experiment

|  | Evaluation Metrics | | | |
| --- | --- | --- | --- | --- |
| Precision | Recall | F1-Score | Support |
| Non-Hate Speech (0) | 0.81 | 0.91 | 0.86 | 1516 |
| Hate Speech (1) | 0.86 | 0.70 | 0.77 | 1118 |
| Accuracy |  |  | 0.82 | 2634 |
| Macro Average | 0.83 | 0.81 | 0.81 | 2634 |
| Weighted Average | 0.83 | 0.82 | 0.82 | 2634 |

1. XGBoost Model Overall Performance

| Evaluation Metric | Score |
| --- | --- |
| Accuracy | 0.8230827638572513 |
| Precision | 0.8268166871498865 |
| Recall | 0.8230827638572513 |
| F1-Score | 0.8198672365659038 |

# CONCLUSION

In this study, we developed a hate speech detection system using XGBoost, a gradient enhancement algorithm, and applied it to a dataset containing raw and clean text from social media. Through experiments, we achieved an accuracy of 82.31%, demonstrating the effectiveness of his XGBoost in detecting hate speech.

The classification report and performance metrics results showed that the model performed well in detecting hate speech, with a weighted average accuracy of 82.68%, a recall of 82.31%, and an F1-score of 81.99%. This demonstrates that our model can accurately classify cases of hate speech which can be used to identify and mitigate harmful online content.

However, our study has room for improvement. Fine-tuning the model by optimizing hyperparameters and exploring different feature engineering techniques could potentially enhance the performance of our hate speech detection system. Additionally, incorporating explainable AI techniques like LIME (Local Interpretable Model-Agnostic Explanations) could provide insights into the model's decision-making process and improve transparency, allowing for better understanding and interpretation of the results.[3]

The future work could focus on expanding the dataset to include a more diverse range of hate speech instances, as well as non-hate speech examples, to further enhance the generalizability of the model. Additionally, exploring the use of other machine learning algorithms, such as deep learning models like Recurrent Neural Networks (RNN) or Transformer-based architectures like BERT, could potentially improve the performance of hate speech detection. [14]

In conclusion, our study demonstrates the effectiveness of XGBoost in hate speech detection and highlights the importance of developing robust and accurate models to combat hate speech online. By further refining our approach, incorporating explainable AI techniques, and expanding our datasets, we can further advance the field of hate speech detection and help build a safer, more inclusive online space.

##### References

1. C. O’Regan, Hate Speech Online: an (Intractable) Contemporary Challenge?, Current Legal Problems, vol. 71, issue 1, 2018, pp. 403–429, doi: 10.1093/clp/cuy012.
2. M. Estellés and J. Castellví, The Educational Implications of Populism, Emotions and Digital Hate Speech: A Dialogue with Scholars from Canada, Chile, Spain, the UK, and the US, Sustainability, vol. 12, no. 15, pp. 6034, Jul. 2020, doi: 10.3390/su12156034.
3. M. O. Ibrohim and I. Budi, “Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter,” in Proceedings of the Third Workshop on Abusive Language Online, Aug. 2019, pp. 46–57. doi: 10.18653/v1/W19-3506.
4. J. S. Malik, G. Pang, and A. van den Hengel, Deep Learning for Hate Speech Detection: A Comparative Study. 2022.
5. A. Stevenson and C. A. Lindberg, ‘sentiment analysis’. Oxford University Press, 2011.
6. E. Katona, J. Buda, and F. Bolonyai, “Using N-grams and Statistical Features to Identify Hate Speech Spreaders on Twitter,” 2021.
7. “What is natural language processing?,” IBM, https://www.ibm.com/topics/natural-language-processing (accessed Jun. 15, 2023).
8. R. T. Mutanga, N. Naicker, and O. O. Olugbara, “Hate Speech Detection in Twitter using Transformer Methods,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 9, 2020, doi: 10.14569/IJACSA.2020.0110972.
9. T. Chen and C. Guestrin, “XGBoost,” Aug. 2016. doi: 10.1145/2939672.2939785.
10. H. Watanabe, M. Bouazizi, and T. Ohtsuki, “Hate Speech on Twitter A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection,” IEEE Access, vol. PP, pp. 1–1, Feb. 2018, doi: 10.1109/ACCESS.2018.2806394.
11. G. Pitsilis, H. Ramampiaro, and H. Langseth, “Effective hate-speech detection in Twitter data using recurrent neural networks,” Applied Intelligence, vol. 48, p. in press., Dec. 2018, doi: 10.1007/s10489-018-1242-y.
12. M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding, “Peer to Peer Hate: Hate Speech Instigators and Their Targets”, *ICWSM*, vol. 12, no. 1, Jun. 2018.
13. A. Al-Hassan and H. Al-Dossari, “Detection of Hate Speech in Social Networks: A Survey on Multilingual Corpus,” Feb. 2019, pp. 83–100. doi: 10.5121/csit.2019.90208.
14. H. Saleh, A. Alhothali, and K. Moria, “Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model,” Applied Artificial Intelligence, vol. 37, Feb. 2023, doi: 10.1080/08839514.2023.2166719.
15. K. Halevy, A Group-Specific Approach to NLP for Hate Speech Detection, 2023.
16. P. P. Jemima, B. R. Majumder, B. K. Ghosh and F. Hoda, "Hate Speech Detection using Machine Learning," *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 2022, pp. 1274-1277, doi: 10.1109/ICCES54183.2022.9835776.
17. F. Del Vigna, A. Cimino, F. Dell’Orletta, M. Petrocchi, and M. Tesconi, “Hate me, hate me not: Hate speech detection on Facebook,” Jan. 2017.
18. C. S. Wu and U. Bhandary, "Detection of Hate Speech in Videos Using Machine Learning," *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, 2020, pp. 585-590, doi: 10.1109/CSCI51800.2020.00104.
19. A. Haque and M. N.-U.-R. Chowdhury, “Hate Speech Detection in Social Media Using the Ensemble Learning Technique”. TechRxiv, Apr. 2023, doi: 10.36227/techrxiv.22583857.v1
20. X. Zhou et al., “Hate Speech Detection Based on Sentiment Knowledge Sharing,” Jan. 2021, pp. 7158–7166. doi: 10.18653/v1/2021.acl-long.556.
21. P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep Learning for Hate Speech Detection in Tweets,” 2017. doi: 10.1145/3041021.3054223.
22. K. Makram et al., “CHILLAX - at Arabic Hate Speech 2022: A Hybrid Machine Learning and Transformers based Model to Detect Arabic Offensive and Hate Speech,” in Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur’an QA and Fine-Grained Hate Speech Detection, Jun. 2022, pp. 194–199. [Online]. Available: https://aclanthology.org/2022.osact-1.25
23. S. Goswami, M. Hudnurkar and S. Ambekar, “Fake News and Hate Speech Detection with Machine Learning and NLP”, *J Arch.Egyptol*, vol. 17, no. 6, pp. 4309 - 4322, Dec. 2020.
24. T. T. A. Putri, S. Sriadhi, R. D. Sari, R. Rahmadani, and H. D. Hutahaean, “A comparison of classification algorithms for hate speech detection,” IOP Conference Series: Materials Science and Engineering, vol. 830, no. 3, p. 032006, Apr. 2020, doi: 10.1088/1757-899X/830/3/032006.
25. A. Khan, J. Yousaf, T. Muhammad, and M. Ismail, “Hate speech detection using machine learning and N-gram techniques,” May 2023.
26. A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, and S. H. Malik, “Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques,” International Journal of Information Management Data Insights, vol. 2, no. 2, p. 100120, 2022, doi: https://doi.org/10.1016/j.jjimei.2022.100120.
27. T. Davidson, D. Warmsley, M. Macy, and I. Weber, Automated Hate Speech Detection and the Problem of Offensive Language, 2017.
28. O. R. Hartono, “Indonesian stoplist,” Kaggle, https://www.kaggle.com/datasets/oswinrh/indonesian-stoplist (accessed Jun. 15, 2023).
29. Z. Waseem and D. Hovy, “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter,” in Proceedings of the NAACL Student Research Workshop, Jun. 2016, pp. 88–93. doi: 10.18653/v1/N16-2013.
30. P. Kominfo, “Siaran Pers no. 143/HM/KOMINFO/04/2021 tentang sejak 2018, Kominfo Tangani 3.640 ujaran Kebencian Berbasis Sara di Ruang Digital,” Website Resmi Kementerian Komunikasi dan Informatika RI, https://www.kominfo.go.id/content/detail/34136/siaran-pers-no-143hmkominfo042021-tentang-sejak-2018-kominfo-tangani-3640-ujaran-kebencian-berbasis-sara-di-ruang-digital/0/siaran\_pers (accessed Jun. 15, 2023)